СЕМАНТИЧЕСКИЙ АНАЛИЗ И ОТВЕТЫ НА ВОПРОСЫ: СИСТЕМА В СТАДИИ РАЗРАБОТКИ

Богуславский И. М., Диконов В. Г., Иомдин Л. Л., Лазурский А. В., Сизов В. Г., Тимошенко С. П.

Институт проблем передачи информации РАН им. А. А. Харкевича, Москва, Россия

В статье представлена система семантического анализа и вопросно-ответная система, реализованная на ее основе. Предметной областью являются новости про футбол. На входе система получает вопрос на естественном языке, в качестве ответа выдаёт элемент из базы данных. Модуль семантического анализа лингвистического процессора ЭТАП-3 строит для каждого предложения семантическую структуру, представляющую собой набор троек вида **`семантическое_отношение(индивид,индивид)`**. Семантические отношения и индивиды, из которых состоит семантическая структура, соответствуют элементам онтологии, которая таким образом становится функциональным аналогом словаря для «семантического языка». Семантические структуры предложений одного текста объединяются благодаря установлению кореференции между объектами и конвертируются в OWL-документ, использующийся далее в качестве базы данных. В базу данных также помещаются фоновые сведения из базы индивидов о конкретных командах, футболистах, матчах. Благодаря этому становится возможным находить ответ на вопрос, используя информацию, содержащуюся не только в разных предложениях текста, но и в базе индивидов. Так, если пользователь задал вопрос *Какая команда нанесла поражение чемпиону Испании?*, а мы располагаем текстом, в котором сообщается, что *Подопечные Слуцкого обыграли мадридский «Атлетико»*, то система установит соответствие между вопросом и этим текстом и даст правильный ответ: *ЦСКА*. Семантическая структура, полученная из вопроса, конвертируется в SPARQL-запрос к базе данных. На данный момент все части системы функционируют, работа находится в стадии отладки.

Ключевые слова: глубокий семантический анализ, семантический словарь, онтология, вопросно-ответная система, кореферентность

SEMANTIC ANALYSIS AND QUESTION ANSWERING: A SYSTEM UNDER DEVELOPMENT[1]

Igor Boguslavsky, Vyacheslav Dikonov, Leonid Iomdin, Alexander Lazursky, Victor Sizov, Svetlana Timoshenko

A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

The paper presents a system of semantic analysis and a question answering system implemented on its basis for a specific subject domain: (European) football match news. As input, the system obtains a natural language question (in Russian), which it answers with an element (or elements) from the repository of individuals. The core part of the system is the semantic analyzer of natural language texts. For each sentence of the text processed, the special semantic analysis component of ETAP-3 linguistic processor constructs a semantic structure, which consists of a set of triples of the type **`semantic_relation (individual,individual)`**. Semantic relations and individuals constituting this structure correspond to the elements of the ontology, which can thus be viewed as a functional analogue of a dictionary for the semantic language. Semantic structures of sentences belonging to a particular text are integrated thanks to coreference and anaphora resolution and converted into an OWL-document, which is later used as a database. This database is supplemented by background knowledge from the repository of individuals concerning specific teams, football players, and games. Thanks to this resource, we are able to find an answer to the question using not only the data contained in different sentences of the text but also in the repository of individuals. If the user asks "*What team defeated the champion of Spain?"* while we have a text reporting that "*Slutsky's players outplayed Atletico Madrid*" then the system will establish the correspondence with the question, the text, and the records in the depository of individuals, and will come with the correct answer "CSKA". The semantic structure obtained from the natural language question is converted into a SPARQL query addressed to the database. Currently, all parts of the system are operating in the test mode.

Key words: deep semantic analysis, semantic dictionary, ontology, question answering, coreference

1. Introductory remarks

Semantics is the most essential and probably the most complex component of the full model of natural language. So far, the task of semantic analysis is far from being

[1] The paper is partially funded by the Russian Humanitarian Scientific Foundation, grants No. 13-04-0043 and 15-04-00562, and the Russian Foundation of Basic Research, grant No. 15-06-09208. The authors express their gratitude to both Foundations.

fulfilled, even though numerous attempts to achieve this goal using diverse approaches have been made. Many researchers view the task of semantic analysis in tagging the text by semantic elements, such as WordNet synsets, ontology classes or individuals, semantic roles, or FrameNet frames (Shi and Mihalcea, 2004; Coppola, Moschitti, 2010; Azmeh et al., 2011). In a different approach, many authors attempt to translate natural language sentences into a logical language (see e.g. Bos, 2008, 2011; Copestake et al., 2006, Allen et al., 2008). On the other hand, many papers focused on semantic analysis tend to use, in addition to linguistic data, also background information contained in the ontology. This approach, primarily pursued in the OntoSem project (Nirenburg, Raskin, 2004; Akshay Java et al., 2007; Raskin, Taylor, 2010; Raskin et al., 2010; Nirenburg, McShane, 2012) is helpful in tasks of lexical and/or syntactic disambiguation and can be used in deducing all sorts of inferences which contribute to deeper and more comprehensive understanding. A series of articles written by the Spanish FuncGram group who works in the framework of Lexical Constructional Model develop a semantic database, used for inferences (see e.g. Mairal Usón, Periñán-Pascual, 2009; Periñán-Pascual, Arcas-Túnez, 2010a, b; Periñán-Pascual, Mairal Usón, 2010). Automatic semantic analyzers are actively developed within the machine learning paradigm, especially under supervised learning (cf. Ge and Mooney, 2005; Poon and Domingos, 2009; Clarke et al., 2010; Titov and Klementiev, 2011; Liang et al., 2011). An obvious obstacle here is the lack of sufficiently large semantically tagged corpora. It should be added that some semantic parsers combine mixed technique: machine learning and linguistic rules (Moldovan et al., 2010).

Unlike these last ones, our analyzer of Russian texts is strictly rule-based, which seemingly contradicts the current trend. Our choice of strategy is based on two considerations. First, there exist no corpora annotated with the kind of structure we are interested in. Once we construct our analyzer, it will open the possibility to develop such a corpus, which could then be used for refining and evaluating the analyzer, as well as for developing other semantic parsers. The second, and more important, reason for our no statistics approach is our firm belief that the modelling of real understanding of texts requires knowledge-intensive methods (for details, see Boguslavsky 2011).

Most researchers agree that the goal of constructing a broad-coverage system of deep semantic analysis is currently unachievable. There are two possible ways to produce such a system: (1) start with an extensive shallow understanding system and gradually deepen it, as suggested in (Riloff, 1999), or else begin with a small-scale deep understanding system and gradually broaden the coverage (Mueller, 2006). We advocate the latter approach. Our semantic analyzer prototype is aimed at deep understanding of texts belonging to a restricted subject domain: football match news. To achieve this, we need to use not just linguistic knowledge but subject domain knowledge which is provided by the ontology.

The football subject domain has repeatedly attracted the attention of research groups working with ontologies. Several ontologies have been built, which contain a rather comprehensive nomenclature of football terms. These include the SWAN Soccer Ontology by DERI (http://sw.deri.org/ knud/swan/ontologies/soccer), Ranwez's ontology (http://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.daml), the sports fragment of the OpenCyc Ontology (http://sw.opencyc.org/2009/04/07/concept/en/Soccer), the sports fragments in the DAML repository (http://www.cse.sc.edu/~dukle/

ontologies/football-ont.daml) (Dukle, 2003), the soccer ontology of the i3media project (Bouayad-Agha et al. 2011), and the Kiktionary Ontology (kiktionary.de). Yet, neither of these ontologies meets the needs of our project so that we had to develop our own ontology. The reason is that the basic application of the ontologies listed above is annotation of texts or multimedia objects aimed at facilitating database search. Typical examples are Ranwez ontology intended for video file annotation, and the paper (Tsinaraki et al., 2005), who discuss methods of using the ontology for semantic indexation of audiovisual material. The Kiktionary ontology is a trilingual dictionary of football terms arranged into several hierarchies, frames, and scenes. Somewhat different is the i3media ontology intended for the generation of summaries of football matches from statistical reports of these matches. The goals determine the contents of these ontologies. Information on each concept supplied therein basically confines to the concept's referral to a higher class of hierarchy and the list of equivalents in one or more languages. To give an example, the Goal concept is presented in Ranwez in the following way (using OWL):

```
<rdfs:Class rdf:ID="Goal">
<rdfs:subClassOf rdf:resource="#Stoppage"/>
<rdfs:label xml:lang="fr"> But
</rdfs:label>
<rdfs:label xml:lang="en"> Goal
</rdfs:label>
</rdfs:Class>
```

This description reads as follows: «Concept Goal is an element of the Stoppage class. In English, the corresponding term is *goal*, and in French it is *but*». Other ontologies provide similar information. This level of explication is quite sufficient for tagging football news reports with ontological concepts. However, it offers no understanding on what this event is about and does not allow drawing any inferences that require such understanding. A more detailed description of this concept given in our OntoEtap ontology is shown below, in section 3.

An important component of the ontology is the set of class instances, which may be many times greater than the set of classes and properties. For example, the i3media project ontology contains 47 classes, 50 properties, and ca. 70 thousand instances. Often, ontologies are automatically populated with data scraped from web pages (Bouayad-Agha et al., 2011; Dukle, 2005) by means of specially designed programs.

In our project, the construction of a complete ontology of football and its population with instances is not our main focus of attention. These tasks were solved previously, and technologies of populating the ontology by different methods, including automatic ones, are known. Our main route of advancement is to learn how to extract implicit knowledge from texts and make inferences based on this knowledge. The extent of ontological coverage plays no essential role in the development of methods that allow for such reasoning. Much more important is to understand exactly which knowledge the system needs to solve this task, to learn how to produce this knowledge and operate it. This approach entails the requirements that should be met by our resources to be, and the order of action. In accordance with this approach, we decided to start

by constructing a fragment of an ontology which should not be big but which should contain knowledge that allows the system to fetch the implicit information extractable from texts. Should the experiments performed in this restricted area prove successful, we could proceed with the expansion of the area and move towards a fuller coverage of the subject domain. This is why we venture to report on this project without waiting for its maturity and high recall that would allow for its full-scale evaluation.

More specifically, our semantic analyzer strives to advance in the following directions:

(i) The use of extralinguistic knowledge in addition to linguistic data presented in the dictionary and the grammar. Extralinguistic knowledge is stored in two repositories: the Ontology and the Repository of Individuals. While the Ontology stores hierarchically arranged information on concepts and their properties, the Repository of Individuals accumulates data on individual objects (like Moscow) or situations (like 2014 FIFA World Cup).

(ii) Explicit presentation of word meanings for inference purposes. We proceed from the assumption that the depth of understanding is growing with the number of inferences we can draw from the text. In many cases, a detailed description of word meanings helps produce additional conclusions and in this achieve a deeper understanding.

(iii) Going beyond the sentence boundary. Normally, syntactic and semantic analysis of text is limited to one sentence, so that it is impossible to look from the sentence processed to a neighbouring one. It is however a serious obstacle for many tasks. Importantly, going beyond the sentence is essential for finding antecedents of pronouns which are very often located in one of the preceding sentences. We will also show below that in order to answer relevant questions to the text it may be essential to resort to the material of several sentences.

Our analyzer is built on the basis of ETAP-3 linguistic processor as its new module. For more details see (Iomdin et al., 2012), (Boguslavsky et al., 2013).

The paper is written according to the following plan. Section 2 will be focused on OntoEtap ontology, covering the three issues just listed. Section 3 will show how word/phrase meanings are represented. Section 4 will discuss how the context is treated. Section 5 will give a few detailed examples demonstrating the system implementation in ETAP-3. Although full-fledged system evaluation in terms of precision and coverage is not possible at this stage, it is nevertheless desirable to make sure that the information produced by the system can work. Therefore, we decided to carry out some restricted experiments in the question-answering scenario in which this information would come in handy. These experiments are described in Section 6. In section 7 we will outline directions of future work.

2. OntoEtap Ontology

The ontology is a natural intermediary link that connects natural language with extralingusitc knowledge and the processes with which this knowledge is manipulated. There are two distinctly different sorts of such knowledge.

One sort is Ontology proper. It consists of a hierarchically arranged set of concepts with formal properties assigned to them. For instance, the class concept **City** denotes the class of cities, which is a subclass of the class **GeopoliticalArea**; in its turn, **City** has a subclass **CapitalCity**. The concept **City** is assigned such properties as the country where the city is situated, its population, area, geographic coordinates etc. All properties of a class are inherited by its subclasses. Many concepts refer to more than one class, thus inheriting properties coming from different sources.

The other sort of extralinguistic knowledge is the Repository of individuals, which contains information on individual objects or situations that are concrete instances of Ontology concepts. Individual objects are e.g. Moscow, France, the Thames, or Cervantes, and individual situations include World War II, Sochi 2014 Winter Olympics, or Yesterday's match between Arsenal and Manchester United.

We have chosen OWL (Web Ontology Language) to present both the Ontology and the Repository of Individuals, because this language is common for ontology developers and ontological semantics community and because a number of useful tools are available for OWL manipulation (such as editors, reasoners, or workplaces).

Our OntoEtap ontology has two sources. One source is the well-known upper/middle ontology SUMO, or Suggested Upper Merged Ontology (www.ontologyportal.org), rewritten in OWL. SUMO is an open formal ontology formulated in the first order predicate logic language. It contains references to WordNet and involves several restricted subject domain ontologies, covering jointly about 20,000 concepts and 60,000 axioms. The other source is our own small ontology of the football news domain, written in OWL using SWRL rules. The resulting OntoEtap ontology is maintained in Protegé enviroment. Today (April 2015), OntoEtap contains 10,963 classes and 5,587 individuals.

3. Representation of Word Meanings

The role of Ontology in our project is twofold. On the one hand, it is the source of structured world knowledge. On the other hand, it serves as a metalanguage of semantic representations. Concepts, individuals and their formal properties are viewed as semantic elements with which semantic structures are built. All (non-functional) Russian words are interpreted in terms of these semantic elements.

To give an example, all the diverse designations of a victory/defeat in a football match, such as *победить* 'defeat', *выиграть* 'win', *переиграть* 'outplay, beat', *разгромить* 'rout', *заработать 3 очка* 'win 3 points'; *проиграть* 'lose', *уступить* 'concede', *потерпеть поражение* 'suffer a defeat', *оказаться сильнее/слабее* 'be stronger/weaker against' etc. are eventually represented by one general concept **WinEvent**. The difference between winning and losing is not reflected at the level of concepts but manifested in the way in which the **WinEvent** concept's roles **hasWinner** and **hasLoser** are implemented.

In addition to words that correspond to a single concept (*победить* 'win' to **WinEvent**, *футболист* 'footballer' to **FootballPlayer** etc.) there are words whose meaning is represented by a structure formed with several concepts. The word

генерал '(army) general' may refer to a military rank or to a human who has this rank. The first sense of the word is tantamount to the concept **GeneralRank**, which belongs to the class **MilitaryRank**, while the second sense corresponds to the structure **hasRole(Human,GeneralRank)**. A different example: the Ontology includes concepts corresponding to animal species, like **Lion**, but it would be redundant to introduce special concepts for words like *львица* 'lioness' or *львенок* 'lion cub'. Such words are best represented with structures that explain their senses with the concept **Lion**: **hasGender(Lion,female)** and **developmentalForm(Lion,NonFullyFormed)**.

Rather often, the link between the natural language and the ontology is understood as mapping NL expressions to ontology nodes. Yet, natural language expressions do not always have a fixed ontological correlate. For instance, expressions like *local team* correspond to different ontological individuals depending on the context. For such cases, special rules of contextual interpretation should be written.

Complex events are described by means of multi-propositional explications (scripts). For example, the central event that happens in a football match—scoring a goal—is supplied with a description that interprets this event as a script formed with three sub-events that are causally related with each other[2]: a **GoalEvent** situation takes place if:

(1) Player who belongs to Team-1 hits the ball in the direction of the GoalArea of Team-2.
(2) As a result, the ball is located in the GoalArea of Team-2.
(3) As a result, Team-1's score increases by one.

Due to the decomposition of the goal concept into several components we can generate an adequate representation of expressions in which these components come into play. For instance, phrases like *забить гол головой (с 20 метров)* 'score a goal with one's head (from 20 meters)' may only be understood if we take into account that a goal event includes hitting the ball (cf. *hitting the ball with the head (from 20 meters)*).

Besides, the description of an event with the help of a script opens a possibility of treating certain kinds of metonymy. It allows one to identify the event even in cases when the text only mentions a part of the relevant components or even a single crucial component. The following sentences do not contain even the word *гол* 'goal' but obviously refer to this sort of event: (1) *Уже на второй минуте вратарь достал мяч из ворот* 'As early as on the second minute, the goalkeeper took the ball out of the goal'; (2) *Месси отправил мяч в верхний левый угол ворот*. ‹Messi sent the ball to the upper left corner of the goal›.

Let us see how these sentences are processed. First, we produce the Basic Semantic Structure (BSemS), which conveys the basic meaning of the sentence in terms of ontological units. BSemSs of sentences (1) and (2) are shown in Fig. 1 and Fig. 2, respectively.

---

[2] To save space, we are not reproducing the formal definition. Many examples of using this language for word definitions can be found in Boguslavsky et al. 2013.

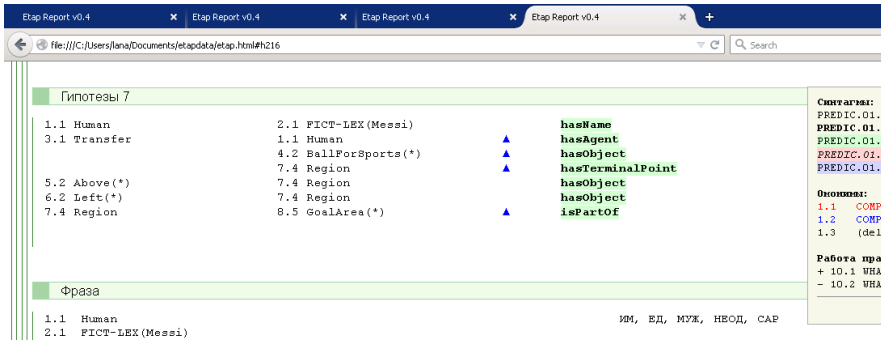


Fig. 1. Basic Semantic Structure (BSemS) for sentence (1)

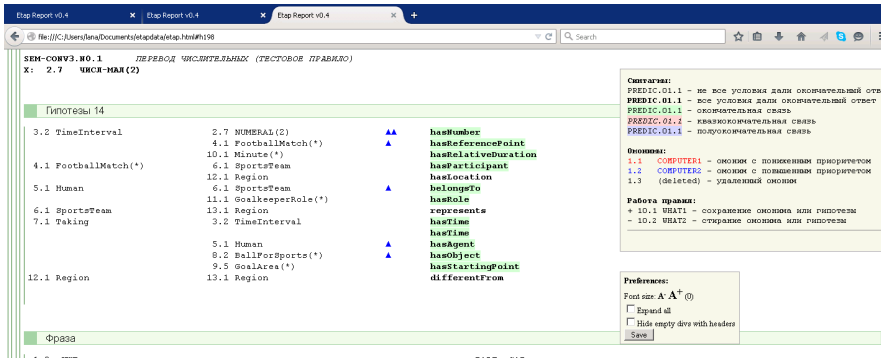


Fig. 2. Basic Semantic Structure (BSemS) for sentence (2)

After that, inference rules are applied, if need be. These rules make various kinds of inferences and have different degree of generality. Among them, there are (a) concept-centered rules, such as effect and precondition rules, which are specific to particular concepts, and (b) more general common sense rules. Both kinds of rules are needed for processing sentences (1) and (2). They are written in one of two formalisms—SWRL (Semantic Web Rule Language) or FORET (ETAP rules language). Below, we formulate these rules in plain words to facilitate understanding by the readers.

(i) if a physical object is taken out of some place, it was previously located in this place (the precondition of the TakeOut event);

(ii) if a ball is located in the goal area, this is the result of the goal (the effect of the GoalEvent; this rule is a kind of abductive inference).

(iii) if a physical object is moved to a place, it becomes located in this place (the effect of the Transfer event)

Rules (i)–(ii) are applied to sentence (1) and generate four new triples boxed in Fig. 1. Sentence (2) requires rules (iii) and (ii), but before (ii) could be applied, a common sense rule (iv) should have been applied:

(iv) if an object is located in place A which is part of place B, it is located in place B (transitivity of the hasLocation relation).

After these rules are applied to sentence (2), its BSemS is supplemented with new triples (boldfaced in Fig. 2), which show that the GoalEvent has taken place.

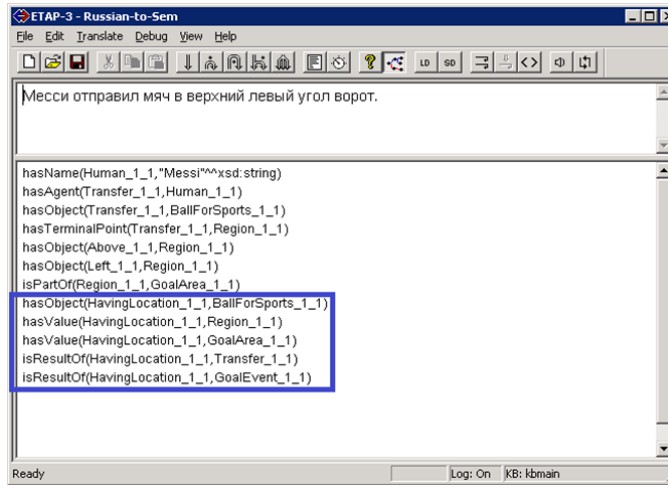


Fig. 3. Inferred semantic structure of sentence (1)

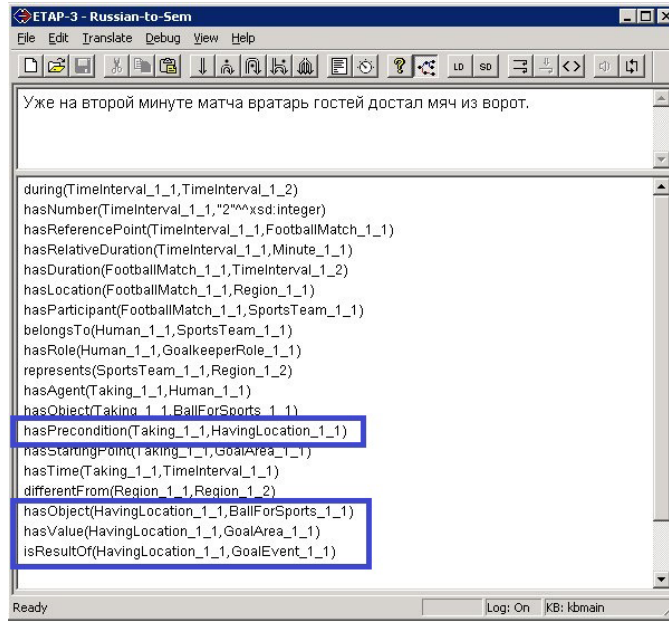


Fig. 4. Inferred semantic structure of sentence (2)

Thus, our analyzer not only interprets complex concepts in terms of simpler ones, but also makes certain kinds of inference.

4. From Sentence to Text

We have mentioned above that the semantic analyzer must operate across sentence boundaries. For a question answering system based on such an analyzer (see Section 6 below for details), this requirement can be reformulated as follows: ideally, the system should be able to find an answer to any question even if it involves collecting the material scattered in different sentences.

The material for the experimental question answering system has been a collection of football news published on the lenta.ru portal. Every message consists of 5 to 10 sentences and normally represents one event. When semantic structures of different sentences are merged into a single description, one of the most difficult problems to be solved is establishing the coreference (or absence thereof) between the objects mentioned in one (or, especially, more) sentences. In our texts, this problem often arises with different names of matches, teams, and players, because the authors tend to avoid repeating the same designations in adjacent sentences. A typical example could be seen in the message *ЦСКА и «Торпедо» встретились в Испании в матче турнира Pinatar Cup. Победу в столичном дерби одержали черно-белые* 'CSKA and Torpedo met in Spain in the Pinatar Cup tournament. In the capital derby, the black-and-whites won'.

In many cases, different designations could be identified as referring to the same entities with the help of the Repository of Individuals. For instance, the entry for the Torpedo team contains the information **`hasNickname "Cherno-belye"`** ('black and whites'). Yet we do not have a general solution of the problem, so we need to look for partial solutions. Considering the fact that news messages normally (though not always) cover only one sports event, we assume that different mentions of an event within a message are coreferent if no evidence to the contrary are available. Pointing to different times, locations, or lists of participants may form such evidence.

Suppose for example that we need to find an answer to the following question: *Как сыграли «Бавария» и «Реал Мадрид»?* 'How did Bavaria and Real Madrid play?'

In our news collection, we have the following message fragment:

(3) *Мюнхенская «Бавария» обыграла мадридский «Реал» в первом полуфинальном матче футбольной Лиги чемпионов. Встреча, проходившая на стадионе «Альянц-Арена» в Мюнхене, завершилась со счетом 2:1.*
'Munich Bavaria beat Real Madrid in the first semifinal match of the football Champions League. The meeting, held at the stadium "Allianz Arena" in Munich, ended with the score 2:1'.

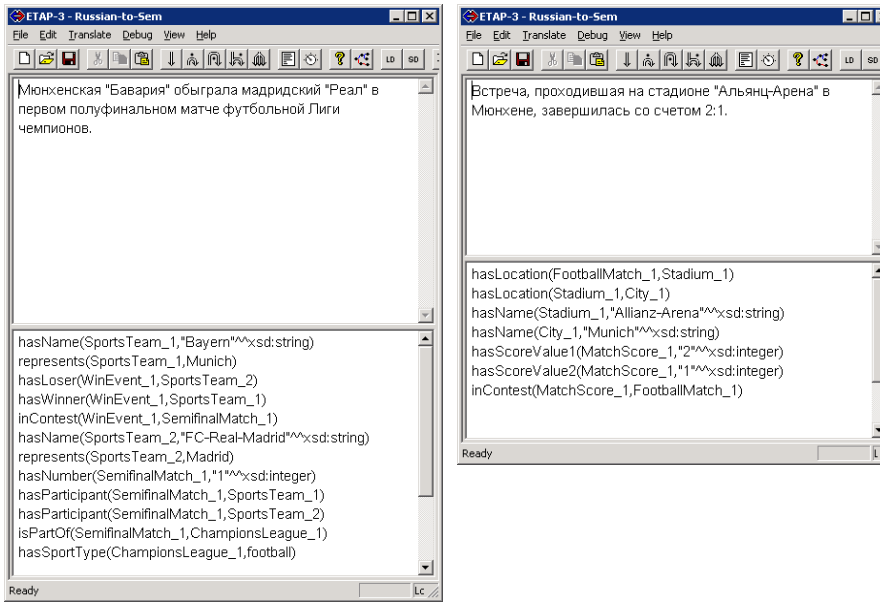
In order to extract the full answer to this question from the given text (namely, that *«Бавария» победила «Реал» со счетом 2:1* 'Bavaria beat Real with the score 2:1'), we need to understand that both sentences report on the same match. This could be done because there is no evidence that the matches are different.

5. Case studies

Below, we will demonstrate two techniques of semantic analyzer operation.

5.1. Building a semantic structure of the message with semantic structures of individual sentences

Let us return to text fragment (3) considered in Section 4 above. The semantic structures of the two sentences are presented in Fig. 5a-b. Both figures are screenshots of ETAP-3 linguistic processor operation.



a b

Fig. 5. Semantic Structures of Text Fragment (3): a—first sentence, b—second sentence

The structure in Fig. 5a can be "read" in the following way: "The team SportsTeam 1 is called Bayern (which is German for Bavaria) and represents Munich (lines 1–2). It lost to team SportsTeam 2 in a semifinal match SemifinalMatch_1 (lines 3–5). Team SportsTeam 2 is called FC-Real-Madrid and represents Madrid (line 6–7). The semifinal match in which these teams participated is part of the Champions League and has sequence number 1 (lines 8–12).

The structure in Fig. 5b says that some football match FootballMatch_1 took place in some stadium Stadium_1 located in some city City_1 (lines 1–2), this stadium is called Allianz-Arena and the city is called Munich (lines 3–4), the score for one team ScoreValue1 in MatchScore_1 was 2 (line 5), the score for the other team

ScoreValue2 in the same MatchScore_1 was 1 (line 6), and this MatchScore_1 was reached in football match FootballMatch_1 (line 7).

The two semantic structures are merged into one integral structure of the text, in which the coreference is established between the event SemifinalMatch_1 from the first sentence and the event FootballMatch_1 from the second sentence. Now the structure is ready to be used in answering the question.

5.2. Addressing the information on individuals stored in the Repository of Individuals

As was already mentioned, the semantic analyzer can access the information on individuals that is absent from the text and is only represented in world knowledge. This access allows us to state the referential identity of expressions which look extremely different. In its turn, the question answering system enhances its potential and becomes able to answer a much broader range of questions. If, for example, the user asks (4) *Какая команда нанесла поражение чемпиону Испании?* 'What team defeated the champion of Spain?' while we have a text reporting that (5) *Подопечные Слуцкого обыграли мадридский «Атлетико»* 'Slutsky's players outplayed Atletico Madrid' then the system will establish the correspondence between the question, the text, and the records in the depository of individuals, and will come with the correct answer "CSKA". This inference relies upon the background knowledge that Slutsky is the trainer of CSKA and that Atletico Madrid won La Liga 2013–14 tounrament, which is the Championship of Spain. All these data can be seen in Fig. 6–7 below that represent the ETAP-3 semantic analysis operation screenshots, and in Fig. 8a–b that represents information from the depository of individuals.

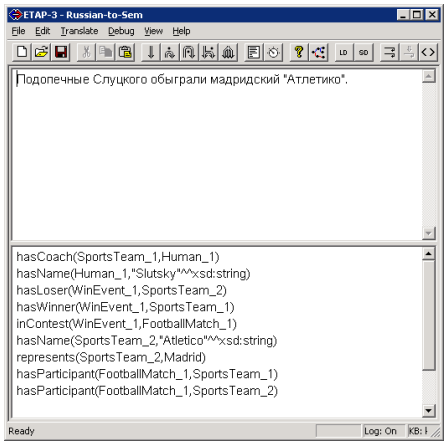


Fig. 6. Semantic structure of the question

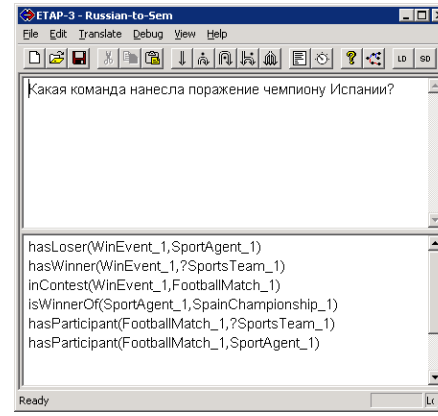
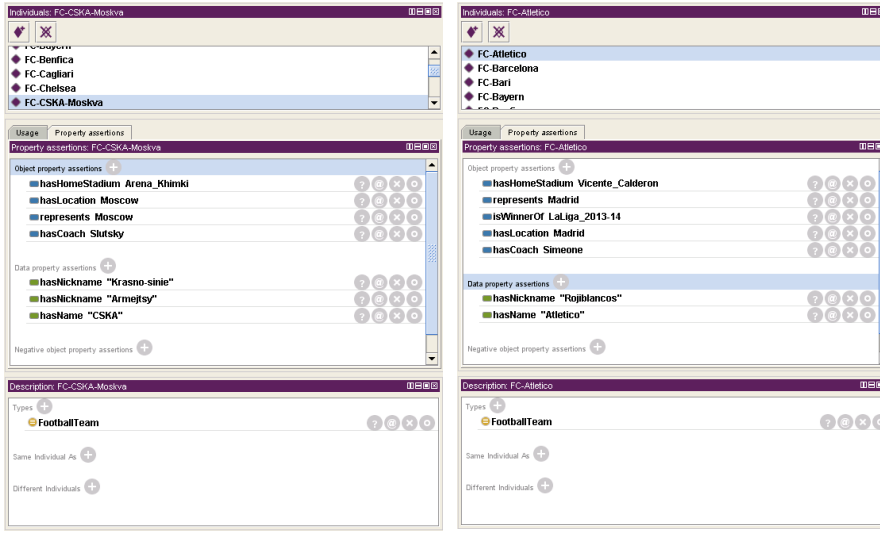


Fig. 7. Semantic structure of the sentence—answer



a b

Fig. 8. Background knowledge on the teams: a) CSKA; b) Atletico Madrid

6. Question Answering Based on Semantic Analysis

We have already mentioned that our semantic analyzer is applied to the task of answering natural language questions. To solve this task,

(1) natural language texts that presumably contain answers to the questions are processed by the semantic analyzer that builds a semantic structure for every sentence;
(2) the set of semantic structures generated by the semantic analyzer is converted into an OWL-document, which is uploaded into Protégé 4.3, a knowledge base management system;
(3) the semantic structure obtained for the question is converted into a SPARQL-query, which is then implemented at the SPARQL access point incorporated into Protégé 4.3.

Since the technique of semantic structure generation has been discussed in detail in Sections 3–5 above, we will focus here on the remaining two issues.

6.1. Generating an OWL-document for semantic structures

The language of semantic structure is based on the notions borrowed from the Ontology, which is represented in OWL. Respectively, the elements of semantic structures (individuals and their properties expressed with semantic relations) remain virtually unchanged when transformed into OWL elements. The individuals are assigned a special property **`belongToSent`**, which points to the sentence of the processed text that has served as basis for semantic structure generation. In this way, we arrange the semantic elements according to their first occurrence in the text.

Every OWL-document generated by the converter receives its own unique space of names (in IRI[3] format) which lists all individuals that are created in this document. The full name of such an individual includes a unique identifier which unambiguously determines the text where the individual was mentioned and even the sentence in which it first appeared. Due to this property we are able to match the facts with the texts from which they were extracted even if the respective OWL-document in which they were introduced is merged into some specialized repository of facts.

The following screenshot is a graphic Protégé 4.3 representation of an OWL-document fragment which was built by the semantic structure converter for the first sentence of text (1) from Section 4.

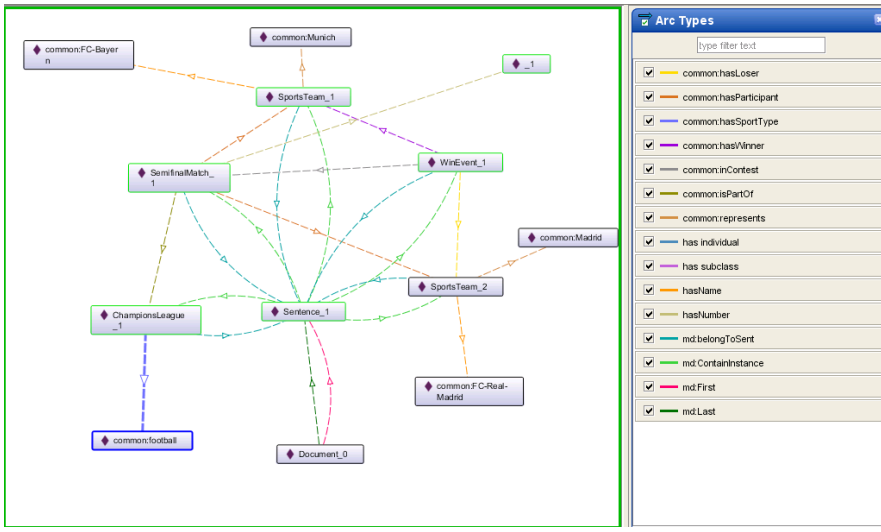


Fig. 9. An OWL-document representing a semantic structure

6.2. Generation of answers to natural language questions

The mechanism of answer generation used in our question answering system is based on the comparison of the semantic structure of the user question with structures present in the repository of individuals. The comparison consists in processing the repository with a SPARQL-query, generated from the semantic structure of the question Technically, SPARQL access point incorporated into Protege 4.3 is used for this purpose.

In the course of system development, we found that semantic structures for NL questions can be basically generated in the same way as the structures for declarative sentences from texts used to extract facts. One minor difference is the fact that we need to identify interrogative words, which is done by assigning these words with a special feature QUEST and forming a special name containing a question mark in the prefix. Such names instruct the SPARQL generator to determine whether

[3] Abbreviation of Internationalized Resource Identifier.

we have to do with a general or a special question. For general questions, SPARQL generates a so-called ASK-query which can be answered with a “yes” or “no”; whilst special questions generate a SELECT-query in which interrogative words are viewed as variables for which all possible values are selected and presented to the user.

Semantic structure elements that are not interrogative words are processed in ASK- and SELECT-queries in the same way. The processing consists in 1) substituting variables for individuals introduced in the question sentence; 2) using individuals specified in the Ontology in their original way: 3) transforming the semantic structure into a graph of triples, and 4) in referring individual variables to the respective classes.

The latter task may prove to be non-trivial if we need to check whether an individual indirectly belongs to a particular class. In this case, instead of the regular record of the form **`<variable> rdf:type <class>`**, where the variable and the class are connected with the relation rdf:type (to be an individual of class) we use the record like **`<variable> rdf:type/rdfs:subClassOf* <class>`**, which means that the variable and the class are linked by a string of relations, in which the first one is **`rdf:type`** followed by a zero (direct belonging to the class) or more relations **`rdfs:subClassOf (indirect belonging to the class)`**.

The following two figures exemplify queries generated for a special question (6) *Кого обыграла “Бавария”?* ’Who did Bavaria outplay?’ (Fig. 10) and a general question (7) *Мадридский «Реал» победил «Баварию»?* ‘Did Real Madrid defeat Bavaria’? (Fig. 11).

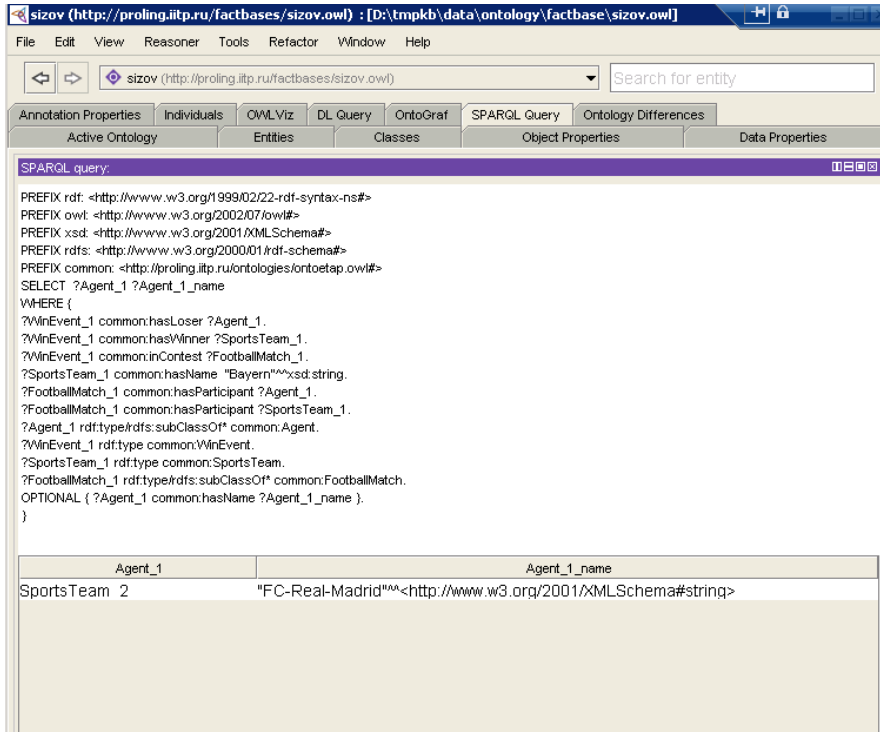


Fig. 10. A SPARQL query for a wh-question. The lower part shows the answer

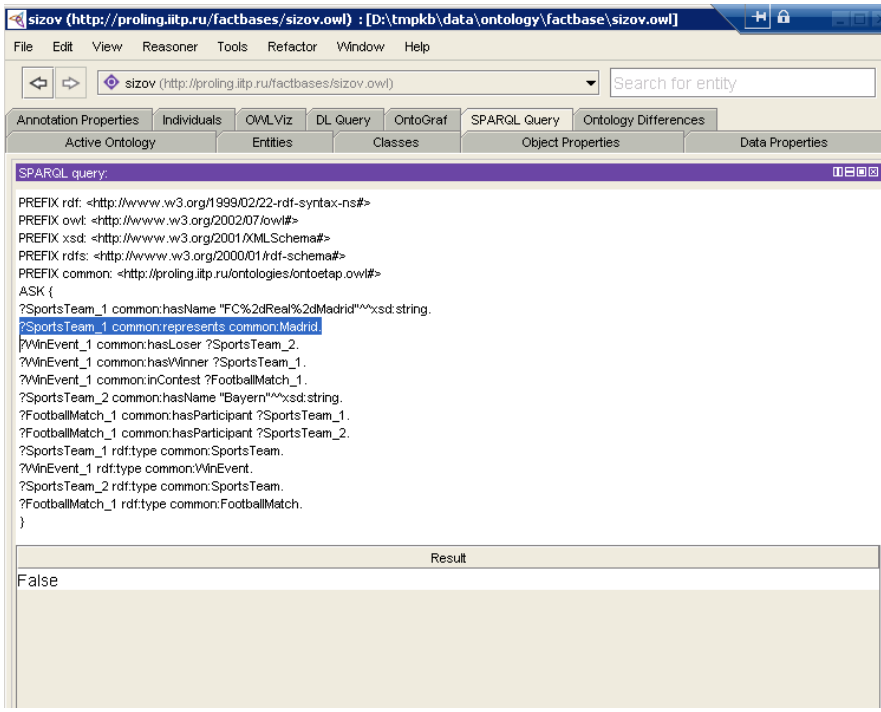


Fig. 11. A SPARQL query for a general question. The answer is negative

7. Future work and conclusion

We have presented a system of deep semantic analysis of Russian texts and shown how it can be used in a question answering system. The database used in our experiments is constructed on the basis of news texts, which are subject to a number of transformations. First, for every sentence a semantic representation is built. Semantic representations of sentences belonging to the same news message are merged into an integrated semantic structure using the coreference relation. The integrated semantic structure is converted into an OWL document. Natural language questions to the system are processed in a similar way: for any such question, a semantic structure is built, to be later converted into a SPARQL query to the database.

There are several important directions of work to be done. The specialized football ontology should be extended. On the one hand, many new instances (such as teams, players, stadiums) should be introduced and their properties described. On the other hand, many more concepts should be supplied with descriptions. In particular, we plan to develop a number of scripts to account for major complex events that occur during the match. One can wonder if it is at all feasible to compile and implement a reasonably complete list of scripts to be applied to match descriptions. In our opinion, this task is realizable, given that the total number of event types

in football is quite restricted, and each of them consists of a small number of elementary events (cf. our description of the scoring event above). Therefore, the task of identifying a complex event by its crucial components seems practicable.

A number of linguistic tasks need to be solved. In particular, the coreference resolution, which is the most sensitive component of the current system, should be made more reliable. As of today, we use a rather simplified rule: within a message, differently presented individuals belonging to one class are considered coreferential if there is no evidence to the contrary, like references to different participants of an event or different names of an object. This rule needs to be made more precise in order for system performance to improve.

Upon the extension of both the ontology and the linguistic data, a full-fledged evaluation of the system will be carried out.

References

1. *Akshay Java, Nirenburg S., McShane M., Finin T., English J., Anupam Joshi* (2007), Using a Natural Language Understanding System to Generate Semantic Web Content, International Journal on Semantic Web and Information Systems, 3(4), pp. 50–74.
2. *Allen J. F., Swift M., Beaumont W.* (2008), Deep Semantic Analysis of Text, Symposium on Semantics in Systems for Text Processing (STEP), volume 2008.
3. *Azmeh Z., Falleri J.-R., Huchard M., Tibermacine C.* (2011), Automatic Web Service Tagging Using Machine Learning and WordNet Synsets, Web Information Systems and Technologies. Lecture Notes in Business Information Processing. Vol. 75, pp. 46–59.
4. *Boguslavsky I. M.* (2011), Semantic Analysis based on linguistic and ontological resources, Proceedings of the 5th International Conference on Meaning-Text Theory (MTT'2011). Barcelona, September 8–9. pp. 25–36.
5. *Boguslavsky I. M., Dikonov V. G., Iomdin L. L., Timoshenko S. P.* (2013), Semantic representation for NL understanding, Computational Linguistics and Intellectual Technologies. International Conference Dialogue-2013 Proceedings, Issue 12 (19) in two volumes, Bekasovo, May, 29—June, 2. Moscow, RGGU Publishers, pp. 132–144.
6. *Bos J.* (2008) Wide-Coverage Semantic Analysis with Boxer, Semantics in Text Processing, STEP 2008 Conference Proceedings. W08–2222.
7. *Bos J.* (2011) A Survey of Computational Semantics: Representation, Inference and Knowledge in Wide-Coverage Text Understanding, Language and Linguistics Compass, Vol. 5, Issue 6, pp. 336–366.
8. *Bouayad-Agha N., Casamayor G., Mille S., Rospocher M., Serafini L., Wanner L.* (2012), From Ontology to NL Generation of Multilingual User- Oriented Environmental Reports, Proceedings of NLDB 2012: 17th International Conference on Applications of Natural Language Processing to Information Systems. Groningen.
9. *Clarke, J., Goldwasser D., Chang M. and Roth D.* (2010), Driving Semantic Parsing from the World's Response, Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010).

10. *Copestake A., Flickinger D., Pollard C. and Sag I.* (2006), Minimal recursion semantics: An introduction, Research on Language and Computation 3 (4), pp. 281–332.
11. *Coppola B. and Moschitti A.* (2010), A General Purpose FrameNet-based Shallow Semantic Parser, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Eds. Nicoletta Calzolari et al. Valletta, Malta: European Language Resources Association (ELRA).
12. *Dornescu I.* (2009), EQUAL: Encylopaedic Question Answering for Lists, Working notes for the CLEF 2009 Workshop. Corfu, Greece.
13. *Dukle K.* (2003), A Prototype Query-Answering Engine Using Semantic Reasoning, Master of Science Thesis, University of South Carolina. Manuscript.
14. *Ferrandez O., Spurk C., Kouylekov M., Dornescu I., Ferrandez S., Negri M., Izquierdo R., Tomas D., Orasan C., Neumann G., Magnini B., Vicedo J. L.* (2011), The QALL-ME Framework: A specifiable-domain multilingual Question Answering architecture, Web Semantics: Science, Services and Agents on the World Wide Web. Vol. 9, Issue 2, pp. 137–145.
15. *Ge R., Mooney R. J.* (2005), A Statistical Semantic Parser that Integrates Syntax and Semantics, Proceedings of the Ninth Conference on Computational Natural Language Learning. Ann Arbor, MI, pp. 9–16, June 2005.
16. *Iomdin L. L., Petrochenkov V. V., Sizov V. G., Leonid Tsinman L. L.* (2012), ETAP parser: state of the art, Computational Linguistics and Intellectual Technologies. International Conference (Dialog'2012). Moscow: RGGU Publishers, Issue 11(18). pp. 830–843. ISSN 2221-7932.
17. *Liang P., Jordan M., Klein D.* (2011), Learning Dependency-Based Compositional Semantics. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — v. 1, p. 590–599.
18. *Mairal Usón R., Periñán-Pascual J. C.* (2009), The anatomy of the lexicon component within the framework of a conceptual knowledge base. Revista Espanola de Linguistica Aplicada 22, pp. 217–244.
19. *Moldovan D., Tatu M., Clark Ch.* (2010), Role of Semantics in Question Answering, Phillip C.-Y. Sheu, Heather Yu, C. V. Ramamoorthy, Arvind K. Joshi, Lotfi A. Zadeh (Eds.) Semantic Computing, pp. 373–420.
20. *Mueller E.* (2006), Common sense reasoning. Elsevier, Morgan Kaufmann Publishers.
21. *Nirenburg S., Raskin V.* (2004), Ontological Semantics. The MIT Press. Cambridge, Mass., London, England.
22. *Nirenburg S., McShane M.* (2012), Agents modeling agents: Incorporating ethics-related reasoning. Proceedings of the symposium Moral Cognition and Theory of Mind at the AISB/IACAP World Congress 2012, Birmingham, UK.
23. *Periñán-Pascual J. C., Arcas-Tunez F.* (2010a), Ontological Commitments in FungramKB. Procesamiento del Lenguaje Natural, p. 44.
24. *Periñán-Pascual J. C., Arcas-Tunez F.* (2010b), The architecture of unGramKB. Proceedings of ELRA Conference. Malta.
25. *Periñán-Pascual J. C., Mairal Usón R.* (2010), La Gramatica de COREL: un lenguaje de representation conceptual. Onomazein 21. Universidad de Chile.

26. *Poon H., Domingos P.* (2009), Unsupervised semantic parsing. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 EMNLP 09 (p. 1).
27. *Raskin V., Taylor J.* (2010), Fuzzy Ontology for Natural Language. 29th International Conference of the North American Fuzzy Information Processing Society, Toronto, Canada, July 2010.
28. *Raskin V., Hempelmann C. F., Taylor J.* (2010) Application-guided Ontological Engineering. International Conference on Artificial Intelligence, Las Vegas, NE, July 2010.
29. *Riloff E.* (1999), Information extraction as a stepping stone toward story understanding, Ashwin Ram and Kenneth Moorman, editors, Understanding Language Understanding: Computational Models of Reading. MIT Press.
30. *Shi L. and Mihalcea R.* (2004), Open Text Semantic Parsing Using FrameNet and WordNet. Proceedings HLT-NAACL—Demonstrations '04 Demonstration Papers at HLT-NAACL 2004. pp. 19–22.
31. *Titov I., Klementiev A.* (2011), A Bayesian Model for Unsupervised Semantic Parsing. Learning Dependency-Based Compositional Semantics. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — v. 1,USA, Oregon,.Portland. pp. 1445–1455.
32. *Tsinaraki C., Polydoros P., Kazasis F., and Christodoulakis S.* (2005), Ontology-based semantic indexing for mpeg-7 and tv-anytime audiovisual content. Multimedia Tools and Applications, Vol. 26, No. 3, pp. 299–325.