

Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

SEMANTIC ANALYSIS WITH INFERENCE: HIGH SPOTS OF THE FOOTBALL MATCH

Boguslavsky I. M. (bogus@iitp.ru)^{1,2},
Frolova T. I. (tfrolova@gmail.com)¹,
Iomdin L. L. (iomdin@gmail.com)¹,
Lazursky A. V. (lazursky@mail.ru)¹,
Rygaev I. P. (irygaev@gmail.com)¹,
Timoshenko S. P. (nyrestein@gmail.com)¹

¹ A. A. Kharkevich Institute for Information Transmission
Problems, Russian Academy of Sciences, Moscow, Russia

² Universidad Politécnica de Madrid, Madrid, Spain

The paper describes a new version of the semantic analyzer SemETAP. Our approach is based on the assumption that the depth of understanding is growing with the number of inferences we can draw from the text. The salient features of SemETAP include: 1) intensive use of both linguistic and background knowledge. The former is incorporated in the Combinatorial Dictionary and the Grammar, and the latter is stored in the Ontology and Repository of Individuals. 2) Words and concepts of the ontology may be supplied with explicit decompositions for inference purposes. 3) Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological elements. Enhanced semantic structure (EnSemS) extends BSemS by means of a series of inferences. 4) A new logical formalism Etalog is developed in which all inference rules are written. Semantic analysis with inference allows us to extract implicit information. The analyzer is tested on the task of interpreting high spots of the football match.

Keywords: semantic parser, ontology, inference, co-reference, question answering

СЕМАНТИЧЕСКИЙ АНАЛИЗ С ЛОГИЧЕСКИМ ВЫВОДОМ: ОСТРЫЕ МОМЕНТЫ ФУТБОЛЬНОГО МАТЧА

Богуславский И. М. (bogus@iitp.ru)^{1,2},
Иомдин Л. Л. (iomdin@gmail.com)¹,
Лазурский А. В. (lazursky@mail.ru)¹,
Рыгаев И. П. (irygaev@gmail.com)¹,
Тимошенко С. П. (nyrestein@gmail.com)¹,
Фролова Т. И. (tfrolova@gmail.com)¹

¹Институт проблем передачи информации РАН
им. А. А. Харкевича, Москва, Россия

²Мадридский политехнический университет, Мадрид, Испания

1. Introduction

In this paper, we describe the current state of the semantic analyzer SemETAP, different aspects of which we presented in our previous publications [Boguslavsky et al. 2015], [Boguslavsky 2017], [Rygaev 2017]. The justification of our approach and the review of the state-of-the-art were given in these publications and we will not come back to that again (except for the analysis of some recent publications in section 3).

The salient features of SemETAP are as follows.

- SemETAP is an option of the ETAP-3 linguistic processor and reuses its non-semantic modules (morphological analysis, syntactic dependency parsing, and normalization).
- Semantic analysis makes use of linguistic data and extralinguistic information (background knowledge). The linguistic data are provided by the Combinatorial Dictionary and the Grammar, and the background knowledge is stored in the Ontology and Repository of Individuals (RI). Whereas 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).
- Both words and concepts of the ontology may be supplied with explicit decompositions 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 inferences and thus achieve a deeper understanding.
- Semantic decomposition is carried out in terms of ontological elements. Thus, Ontology is not only a structured repository of background knowledge, but also a metalanguage for semantic description.
- Semantic analysis goes beyond the sentence boundaries. Usually, syntactic and semantic analysis of text is limited to one sentence, so that it is impossible to look

from the sentence under analysis to a neighboring one. It is however a serious obstacle for many tasks. Importantly, going beyond the sentence boundaries is essential for finding antecedents of pronouns which are very often located in one of the preceding sentences.

- Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological elements. Enhanced semantic structure (EnSemS) extends BSems by means of a series of inferences.
- From the formal point of view, semantic structures of both types are represented in the RDF format, i.e. as sets of triples of the type **relation(Ontoelement-1, Ontoelement-2)**, where **relation** is an object or data property of the ontology, and **Ontoelement-i** is a variable or a constant denoting a concept or an instance. The RDF formalism was chosen because, on the one hand, it is very flexible and expressive, and on the other hand, it is supported by a wide range of tools and is easily integrated with many Semantic Web applications.

In previous publications, we put forward the principles which underlie the system and showed its relevant features by means of some examples. In this paper, we will give a more systematic view of the system, emphasizing the new features that were introduced lately (section 4). This section will be preceded by the problem statement (section 2) and the analysis of related work (section 3). In section 5 we will present a case study. Then we will evaluate the system (section 6) and give a brief error analysis (section 7).

2. Problem statement

Given the current state of computational semantics, semantic parsing so detailed and deep as we are aiming at is impossible to achieve for the texts of unrestricted semantics. So far, the only feasible option seems to be working with more or less narrow domains.

We believe that successive ontological, semantic and logical coverage of different domains will in the final analysis enable us to work with increasingly larger-domain texts. This approach can be illustrated by a series of studies carried out by the commonsense reasoning community, which are dedicated to the logicosemantic modeling of different domains ranging from very narrow ones (such as breaking an egg and pouring it into a bowl—[Morgenstern 2001]) to larger ones, such as emotions, interpersonal relations, commonsense psychology, causality, change of state, etc.—[Gordon, Hobbs 2004]; [Gordon, Hobbs 2011]; [Gordon, Hobbs et al. 2011]; [Hobbs, Gordon 2008]; [Hobbs, Gordon 2010]; [Hobbs, Sagae et al. 2012]; [Montazeri, Hobbs, 2011], [Montazeri, Hobbs 2012]).

In this paper, the subject domain selected is that of football reports. The texts of sports reports are an interesting object for ontosemantic studies due to a number of features. The world of a football match is relatively small and universally understandable. This alone makes it a convenient object of modeling. On the other hand, maybe just due to the restricted size of this world, commentators and journalists do their best to make the reports less dull. This may explain high variability, a large number of individual, figurative and metaphorical expressions.

For example, the result of a match between the Russian teams Zenit and Luch-Energiya is described in one of the sites as follows: *V matče 24 tura Čempionata Ros-sii po futbolu «Zenit» doma razgromil «Luč-Energiju» iz Vladivostoka so ščetom 8:1* ‘≈ in the match of the 24th round of the football championship of Russia Zenit defeated Luch Energiya at home with the score of 8 to 1’ (Soccer.ru portal), while another site had a much more picturesque account: *Za 90 s lišnim minut igrovogo vremeni mjač pobyval v setke vorot “Luča” 8 raz, v to vremja kak vratarju “Zenita” Vjačeslavu Malafeevu prišlo’ vytaskivat’ sportivnyj snarjad iz setki svoix vorot liš odnaždy* ‘≈in slightly more than 90 minutes of playing time, the ball visited the Luch’s goal net 8 times, while Zenit’s goalkeeper Vyacheslav Malafeev had to pick the sporting implement out of his goal net only once’ (Lenta.ru portal). The diversity of nominations used to denote the same object or situation may be striking; cf., for example, *mjač* ‘ball’—*igrovoj snarjad* ‘playing implement’—*sportivnyj snarjad* ‘sports implement’—*sfera* ‘sphere’—*kruglyj* ‘the round one’. Here is how the Benfica team was referred to during one and the same report: *«Benfica»—portugal’tsy* ‘the Portuguese’—*lissabontsy* ‘the Lisboners’—*gosti* ‘the away team’—*sopernik* ‘the adversary’—*komanda* ‘the team’—*portugal’skij klub* ‘the Portuguese club’—*podopečnye Rui Vitoria* ‘the charges of Rui Vitoria’. The goal-scoring situation is denoted in an even more diverse way. Let us give some typical examples which are far from exhausting the whole set of nominations used in the reports: *zabil* ‘scored’—*zabil gol* ‘scored a goal’—*otygral odin mjač* ‘won one goal back’—*otkryl ščet* ‘opened the scoring’—*sravnjal ščet* ‘evened the score’—*vyšel vpered* ‘took the lead’—*uveličil (sokratil) razryv* ‘increased (reduced) the gap’—*oformil dubl’* ‘made the double (scored the second goal in the match)’—*otličilsja* ‘excelled’—*otpravil (poslal, zakinul, perepravil, votknul, zakatil, zapulil) mjač v setku vorot* ‘sent the ball to the net’—*porazil (rasstreljal) vorota* ‘hit (shot) the goal’—*realizoval penal’ti* ‘realized the penalty kick’—*dobil mjač v pravij ugol* ‘dealt the final blow in the right corner’—*zamknul pas (naves)* ‘closed the pass (high cross)’—*razvel mjač i golkipera po raznym uglam* ‘separated the ball and the goalkeeper putting them in opposing corners’—*nakazal vratarja* ‘punished the goalkeeper’—*probil mimo golkipera* ‘kicked beside the goalkeeper’—*ščet stanovitsja 1:0* ‘the score becomes 1:0’—*mjač okazalsja (pobyval) v setke vorot* ‘the ball was in the net, visited the net’—*mjač (gol) vletel v vorota* ‘the ball (the goal) flew in the goal’—*vzjatje vorot* ‘seizure of the goal’—*zabil pobednyj gol* ‘scored the victory goal’—*postavil pobednuju točku v matče* ‘make a victory full stop in the match’—*snjal vse voprosy o pobeditele v etom matče* ‘dispelled all the doubts about the winner of this match’—*emu ostavalos’ tol’ko ne promaxmut’sja* ‘it remained only not to miss’—*peredaća (kombinatsija) okazalas’ golevoj* ‘the pass (combination) turned to be goal-scoring’—*zastal golkipera vrasplox* ‘took the goalkeeper by surprise’—*neotrazimo probil* ‘kicked irresistibly’—*ne ostavil golkiperu ni edinogo šansa* ‘did not leave a single chance to the goalkeeper’—*vyjti vpered (v ščete)* ‘take the lead’—*ataka zaveršilas’ rezul’tativnym udarom Xondy* ‘the attack ended in a successful kick by Honda’.

Many of these expressions are not synonymous. E.g. besides scoring a goal they may contain other important components of meaning. For example, *otkryt’ ščet* ‘open the scoring’ means ‘score a goal, which results in the score 1:0’. *Sravnjat’ ščet* ‘even the score’ means ‘score a goal, which results in the tie score’. *Otygrat’ odin mjač* ‘win

one goal back’ means ‘score a goal when the scoring team scored fewer goals than its adversary; as a result, this difference becomes smaller but not equal to zero’. As these examples clearly show, to adequately represent the content of many expressions semantic decomposition is an absolute must.

Besides that, it is characteristic of sports reports to recur to the indirect mode of expression. Many meaning components are expressed implicitly, and the text interpretation system should be able to restore them. Let us give a typical example.

(1) *Korner u vorot xozjaev polja zaveršaetsja udarom Netsida v upor, no Dikan okazyvaetsja na vysote* ‘the corner kick at the goal of the home team ended in the kick point blank by Necid, but Dikan was up to the mark’.

If the Hearer is aware of the background information, he will easily understand that Necid failed to score a goal, although this was not said directly. We will come back to this example below (in 4.3) and show how SemETAP manages to cope with it.

All of the aforesaid makes sports reports understanding a linguistically non-trivial and exciting task. In processing football reports, we lay emphasis on the understanding of “high spots” of the match, similar to (1). We call high spots the moments fraught with scoring a goal, for example when the goal of one of the teams is being attacked. Our aim is to identify major details of the situation making use of all the information available.

3. Related work

Although football is a popular topic of computational linguistics experiments, most of the relevant efforts have been focused on a football ontology construction (cf. Tsinaraki et al. 2005, Schmidt 2006, Abreu et al. 2010, Ranwez Soccer Ontology, SWAN Soccer Ontology) or on generating football match summaries using an ontology (cf. Nadjet Bouayad-Agha et al. 2011). A notable exception is a recently published book Cimiano et al. 2014, which is also using football as its subject domain. At the level of foundational principles, the approach defended in this book is very similar to ours. It proclaims that in order to interpret natural language texts with respect to the domain knowledge, a machine needs (a) a formalization of the domain knowledge by means of an ontology, (b) a process for building meaning representations that are aligned to that domain knowledge, and (c) a way to draw inferences and use the resulting information in the interpretation process. We cannot agree more with these theses. However, their implementation in Cimiano et al. 2014 and in our project is quite different. Besides, it remained unclear to us up to what extent these principles have been implemented in a real system. In particular, it was difficult to make an idea of syntactic and semantic complexity of sentences the system copes with.

One of the differences between our approach and the one of Cimiano and his co-authors is that their ontology does not support the representation of events and their modification (Cimiano et al. 2014 48–49). As we will show below, our language of meaning representation is much more expressive.

Another important difference between our approaches concerns the role played by the ontology and the status of the NL dictionary. The Cimiano approach is radically

ontology-centric. According to it, each text belongs to some specific domain, and one should first of all create an ontology of this domain and then compile a NL dictionary whose role is to specify NL equivalents for the ontological elements. A domain-independent dictionary also exists but its scope is limited to representing closed-class words such as determiners, pronouns, auxiliary verbs, etc. Dictionaries induced by different domain ontologies will be different even in the number and granularity of meanings of particular words.

We cannot accept this approach. It implies that there is no such a thing as a dictionary of a particular language. There are as many dictionaries as specific domains (such as the football domain), which may contain the same words that have different meaning sets and different granularity. We think that such an approach will be very difficult to implement, since there is no clear-cut border between the vocabularies of different domains, as well as between domain-specific and general vocabulary. Besides, it is often very difficult to assign a text to a specific domain. Then which dictionary should be used for its processing? In our opinion, the domain-independent dictionary should not be restricted to closed-class words, since even domain-specific texts contain a large number of general vocabulary words.

We adopted a different approach. We have an integrated dictionary for Russian, in which all domain-specific information is marked in a special way. Such a marking is needed not only for the words that do not occur beyond domain-specific texts. Very often, it is only some senses of a word (or some phrases containing this word) that are domain-specific, other senses being quite neutral. For example, the phrase *red card* can be easily encountered in a free text where it merely means a card whose color is red. Yet in football it denotes a specific punishment and corresponds to a concept (**RedCard**) of the ontology. The connection between the phrase *red card* and this concept is marked as relevant for the sports domain (DOMAIN:SPORT-DOMAIN). As an illustration, below is a fragment of the dictionary entry for КАРТОЧКА 'card'.

```
ENTRY:КАРТОЧКА
...
  ZONE:EN
    TRANS: CARD
  ZONE:SEM
    DOMAIN:SPORT-DOMAIN
  <a rule stating that red card corresponds to RedCard>
  ...
```

If the text we are processing belongs to this domain, the **RedCard** interpretation will be preferred. Otherwise, it will have the status of only one of possible alternatives. This strategy allows us to have a single dictionary matched with one or more domain-specific ontologies.

4. Semantic analyzer SemETAP

In its present state, the SemETAP analyzer is a follow-up of the system described in Boguslavsky et al. 2015, Boguslavsky 2017 and Rygaev 2017. Below, we will show what it looks like today with a particular focus on the components developed recently.

4.1. Analysis of football reports

At the input, SemETAP receives the Normalized Syntactic Structure (Norm-SynS), constructed by the regular ETAP-3 parser. By this moment, all strongly governed prepositions and conjunctions, as well as auxiliary verbs have been deleted, zero copulas have been substituted by the verb *byt'* 'to be', lexical functions (such as Oper, Func and others) have been identified, antecedents of anaphoric pronouns have been found and some other normalization operations have been performed. Further, NormSyntS is subjected to three stages of processing: 1) preparation of Norm-SyntS for semantization, 2) construction of BSemS, 3) construction of EnSemS.

4.1.1. Preparation of the Normalized Syntactic Structure for semantization

At this stage, the following operations are carried out, among others:

- Substitution of antecedents for anaphoric pronouns and making explicit zero actants.

Pust' vratar' sygral i ne očen' uverenno, no ugrozu ot svoix vorot on [⇒ vratar'] otvel. '≈ Even though the goalkeeper did not play very strongly but he [⇒ the goalkeeper] fended of the threat to his goal'

Traore skinul mjač pod udar Ionovu, kotorogo [⇒ Ionova] v poslednij moment operedil Samba. '≈ Traore kicked the ball to Ionov who [⇒ Ionov] was outrun by Samba at the last moment'

- Resolving non-anaphoric coreference based on the background knowledge extracted from the Repository of Individuals.

Dumbia, obygrav neskol'kix sopernikov, vyvel Tošiča odin na odin s Fil'tsovym, posle čego serbu [⇒Tošiču] ostavalos' tol'ko ne promaxnut'sja. . '≈ Dumbia who outplayed several adversaries brought Tošič head to head with Filtsov, after which the Serb [⇒Tošič] only needed not to miss'

- Processing of support verbs aiming at obtaining identical BSemSs for sentences like:

Spartak pobedil Dinamo 'Spartak defeated Dinamo' = *Spartak oderžal pobeđu nad Dinamo* 'Spartak gained a victory over Dinamo' = *Spartak nanjos poraženie Dinamo* 'Spartak inflicted a defeat to Dinamo' = *Dinamo poterpelo poraženie ot Spartaka* 'Dinamo suffered a defeat from Spartak'.

- Splitting the sentence into predications (subordinate, participial, infinitival clauses, predicative NPs).

Sentence V seredine pervogo tajma Netsid posle pasa Dumbija bjet po vorotam, i Malafeev s trudom perevodit mjač na uglovoj, posle podači kotorogo ivuariets

popadaet v perekladinu ≈ In the middle of the first period, Netsid, after a pass by Doumbia kicks the ball towards the goal, Malafeev, with difficulty, moves the ball over the goal line to enable a corner kick, so that, after the corner was kicked, the Ivorian hits the crossbar' is represented by means of 5 temporally ordered predications: 1) Doumbia gives a pass, 2) Necid kicks the ball towards the goal, 3) Malafeev moves the ball over the goal line, which results in a corner kick, 4) somebody kicks the corner, 5) the Ivorian hits the crossbar.

- Transformation of the passive voice into the active one.

Rossijskij futbol byl predstavlen v plej-off srazu dvumja komandami 'Russian football was represented at the play-off by two teams at once' ⇒ *Srazu dve komandy predstavljali v plej-off rossijskij futbol* 'two teams at once represented Russian football at the play-off'.

4.1.2. Constructing Basic Semantic Structure

Basically, this stage contains semantic interpretation of words, syntactic constructions and morphological features by means of ontological elements. If a word has an exact equivalent among the ontology concepts, it is replaced with this concept. If needed, this concept will be semantically interpreted at the next stage. For example, *gol* 'goal'

⇒ **GoalEvent**.

If the ontology does not have such an equivalent, and it is inexpedient to create it, then a rule is composed which constructs a fragment of BSemS. For example, *vratar* 'goalkeeper' is translated as **Human hasRole GoalkeeperRole** ("person that fulfills the goalkeeper role").

A more complicated rule is responsible for interpreting relational adjectives such as *frantsuzskij* 'French'. For readers' convenience, we will not reproduce it here in the formal language, but give its simplified NL gloss:

- 1) If *frantsuzskij* 'French' modifies a noun which corresponds to the ontological class **SportAgent**, then *frantsuzskij* translates as 'representing France' (e.g. *frantsuzskaja sbornaja* 'French national team').
- 2) If *frantsuzskij* 'French' modifies a noun which corresponds to the ontological class **Human**, then *frantsuzskij* translates as 'living in France' (e.g. *frantsuzskie bolet'sčki* 'French fans').
- 3) If *frantsuzskij* 'French' modifies a noun which corresponds to the ontological class **Organization**, then *frantsuzskij* translates as 'acting in France' (e.g. *frantsuzskij muzej* 'French museum').
- 4) If *frantsuzskij* 'French' modifies a noun which corresponds to one of the ontological classes **OrganizedEvent**, **Building**, **StationaryArtifact**, **GeographicArea**, then *frantsuzskij* translates as 'situated in France' (e.g. *frantsuzskie bul'vary (reki)* 'French boulevards (rivers)').
- 5) If *frantsuzskij* 'French' modifies a noun which corresponds to one of the ontological classes **Document**, **Food**, **Artifact**, then *frantsuzskij* translates as 'made in France' (e.g. *frantsuzskoe vino* 'French wine').

4.1.3. Constructing Enhanced Semantic Structure

The rules that operate at this stage mostly explicate the semantics of concepts. To give an example, it is not sufficient to state that all numerous ways of denoting goal scoring correspond to the concept **GoalEvent** (this task is already solved in BSemS). It is no less important to show what exactly goal scoring is. Briefly, a goal is scored if a player of team A kicks the ball with the aim of moving it in the goal of team B; as a result, the ball gets into the goal of team B and the score of team A increases by 1; this event is beneficial for team A and unbeneficial for team B.

Obviously, such a decomposition enables us to obtain a much deeper comprehension of the text and to make more inferences, than if we restricted ourselves to merely establishing the fact that **GoalEvent** takes place. For instance, we can infer what the player of team A has done, where the ball is located just after the event, what happened to the score, who benefitted from the event, etc. By the same token, we can better understand texts like *Udar pjatkoj, i mjač v setke vorot* ‘a kick with the heel, and the ball is inside the goal’. We are informed that a goal has been scored although *goal scoring* has not been mentioned.

In describing events, special attention is paid to such aspects as preconditions of the event and its results, both obligatory and possible, objectives of the participants, the actions they perform, the assessment of the event from the point of view of different participants, etc.

It is to be stressed that predications may have different degrees of epistemic modality. In particular, the maximal degree (**EpistModality hasDegree MaximalDegree**) is assigned to an event that definitely took place. The medium degree (**EpistModality hasDegree MediumDegree**) corresponds to a possible event. Due to this, we can differentiate between the 100%-reliable logical entailments and the inferences that are no more than plausible expectations. The importance of the latter for the interpretation of discourse and, in particular, dialogues, is exemplified in [Boguslavsky et al. 2016].

At the time of writing (February, 2018), we dispose of 261 rules for transforming BSemS into EnSemS. Some of these rules are related to the general vocabulary, and others describe domain (football) concepts. Even the general vocabulary is far from being completely covered by the rules, to say nothing of the concepts of other domains, so that the inventory of rules should be significantly augmented in the future.

These rules are written in a special language, which will be discussed in the next section.

4.2. Language for inference and inference rules

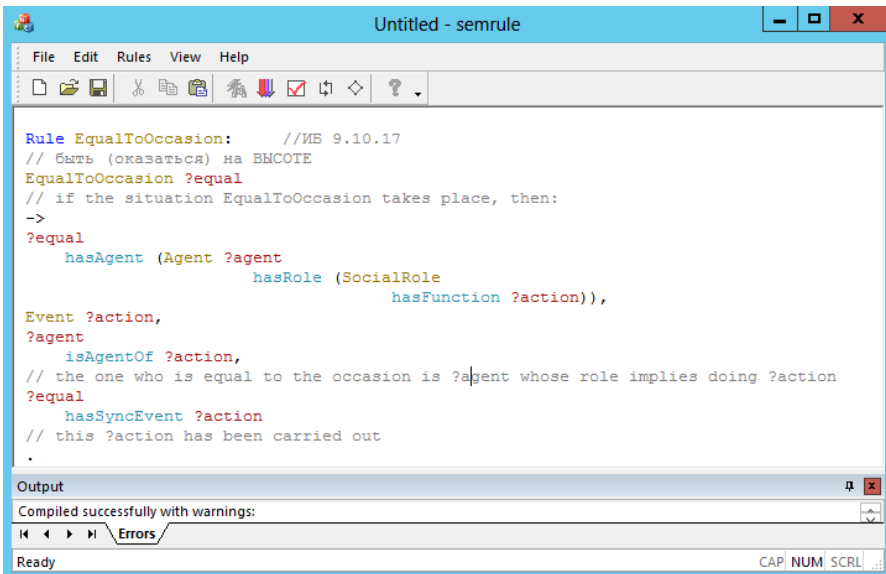
When we were selecting a formalism for writing inference rules, we came to the conclusion that none of the existent formalisms we were aware of could be directly used for our task. We decided that it would be better to develop a formalism of our own, which would be sufficiently expressive for defining the meaning of the concepts and at the same time allow for efficient implementation of the algorithms for logical inference. In developing such a formalism, the following requirements were taken into account:

1. The formalism is destined for formulating logical inference rules. An important particular case of such rules are rules for decomposing the meaning of concepts.
2. The rules should apply to the semantic structure represented by an RDF graph.
3. The rule application should result in adding new triples to SemS. They embody new knowledge obtained by logical inference.
4. The rule application should be efficient from the point of view of the system productivity.
5. The formalism should be easily understandable for the linguists.

As a result of taking these requirements into account, the Etalog language was born, which combines some elements of Datalog (requirements 1 and 4), RDF/SPARQL (requirements 2 and 3) and natural language (requirement 5).

A rule in Etalog consists of a title (which contains the keyword `Rule` and the name of the rule), a logical premise, the implication sign (`->`) and a conclusion. The premise and the conclusion are one or more predicates separated by a comma (stands for conjunction). The rule terminates with a full stop.

Etalog is used to write $B_{SemS} \Rightarrow E_{nSemS}$ rules. Here is an example of one of these rules. It interprets the concept **EqualToOccasion** and is used for processing sentence (1), which we already mentioned in section 2 and will discuss in more detail below (symbol `//` introduces a NL comment):



```

Rule EqualToOccasion: //ИБ 9.10.17
// БЫТЬ (оказаться) на ВЫСОТЕ
EqualToOccasion ?equal
// if the situation EqualToOccasion takes place, then:
->
?equal
  hasAgent (Agent ?agent
            hasRole (SocialRole
                    hasFunction ?action)),
Event ?action,
?agent
  isAgentOf ?action,
// the one who is equal to the occasion is ?agent whose role implies doing ?action
?equal
  hasSyncEvent ?action
// this ?action has been carried out
.
  
```

Output
Compiled successfully with warnings:

Ready

Fig. 1. A semantic rule written in Etalog

4.3. Reasoner

Rules written in Etalog are applied by the RDFox reasoner developed at Oxford University (Motik 2014, Nenov 2015). We chose this reasoner because it suits our needs very well in a number of ways:

1. Very efficient and scalable reasoner. Shows the top results in benchmarking even for large datasets (Benedikt 2017).
2. Optimized for RDF model. Provides an RDF triple store and supports efficient SPARQL query execution.
3. Supports new variables in the consequent of the rules not present in the antecedent. Such variables are known as existentials or anonymous individuals and are essentially required in concept definition rules to represent various parts of the definition.
4. Supports controlled materialization of the existentials (adding new individuals and relations to the semantic structure) allowing for custom filters to prevent infinite loops and guarantee termination. Such procedure is known in the literature as the ‘restricted chase’ (Benedikt 2017, p. 40).
5. Has a built-in support for equality relation (Motik 2015) and a special query mode where different but equal individuals are treated as one individual. It is very helpful in coreference processing.
6. Originally written in C++ and has a solution for Windows. So it integrates very well with ETAP which is also written in C++ for Windows.

Each Etalog rule is translated into several RDFox inference rules. The rule is split into several chunks which are applied independently. First of all functional relations are extracted from the Etalog rule consequent and a separate RDFox rule is created for each of them. The rest of the consequent is split into independent chunks and an RDFox rule is created for each chunk.

This is done to maintain integrity and avoid creation of duplicated entities. Each RDFox rule includes a filter—it does nothing if the corresponding subgraph already exists. The filter works at the level of RDFox rules, so for an Etalog rule it is possible that some individuals will be accommodated from the existing data while others will be added (see more details in Rygaev 2017). This happens invisibly for the linguists who create rules in Etalog, so they can concentrate on the concept definition and can ignore technical aspects of the rule application.

The filters do not always prevent creation of duplicated objects. Because of that we also use equality rules to join duplicated objects together. First of all such rules are created automatically for each functional relation. For more complex cases, additional equality rules can be written manually in Etalog.

We also have an additional filter to guarantee termination of the reasoning. If all the variables in the antecedent of the RDFox rule are anonymous (i.e. do not come from the original data but are created by other rules) the rule does nothing. This is required to avoid infinite chains such as the following: if a player controls the ball he can pass it to another player who then will be controlling the ball and will be able to pass it to another player and so on ad infinitum. This empirical filter works surprisingly well, preventing unnecessary inferences and very rarely blocking good inferences.

4.4. Repository of Individuals

We have built a large Repository of Individuals, which contains data on more than 200K individuals automatically extracted from DBpedia. These individuals belong to the following ontology classes: **Human**, **FootballTeam**, **TimeInterval**, **IndependentState**, **City**, **SportsLeague**. The data on the football players include the name, family name, place and date of birth, country of residence, team (or teams) he played for during his carrier, playing position in the team, etc.

Let us show how all the three resources—Combinatorial Dictionary, Ontology and the Repository of Individuals—contribute to the interpretation process. Let us go back to sentence (1) referred to at section 2.

- (1) *Korner u vorot hozjaev polja zaveršaetsja udarom Netsida v upor, no Dikan' okazyvaetsja na vysote* ‘the corner kick at the goal of the home team resulted in the kick point blank by Necid, but Dikan was up to the mark’.

We want to know if a goal has been scored. To answer this question, we will have to recur to three sources of information:

- Combinatorial Dictionary tells us that the expression *byt' na vysote* ‘be up to the mark’ corresponds to the concept **EqualToOccasion**, interpreted as ‘do well what one is expected to do’ (cf. the rule in the previous section);
- Repository of Individuals contains the information that Andrei Dikan is a goalkeeper of Spartak Football Club;
- Ontology describes the goalkeeper role as preventing the ball from penetrating the goal of his team.

These three pieces of information allow the reasoner to infer that Dikan, being a goalkeeper, performed well his function of preventing a goal and, consequently, a goal has not been scored. Obviously, if the Repository of Individuals had told us that Dikan had the position of a forward, then, given that the Ontology specifies the function of a forward as scoring goals, the overall conclusion would have been opposite.

Again, a conclusion concerning scoring a goal has been made in the context which does not mention the word *goal* nor any of its synonyms.

5. Case study

Let us give another example to illustrate the interpretation of a sentence by means of a series of inferences. We will analyze sentence (2) and show how the analyzer comes to the conclusion that the team for which Aršavin was playing has suffered a defeat.

- (2) *Aršavin tak i ne smog spasti matč* ‘Aršavin could not save the match’.

Among the data at the disposal of the analyzer there are the following three facts which we will for the readers' convenience formulate in NL and not in Etalog, in which they are stored in the system:

1. The verb *smoč* ‘be able’, in the perfective aspect, is implicative (for more details on the implicative verbs in Russian and the impact the verbal aspect

has on the implicativity cf. Rygaev 2015). Therefore, $X \text{ smog } P$ 'X could do P' implies that P took place, while $X \text{ ne smog } P$ 'X could not do P' implies that P did not take place.

2. The phrase *spasti matč* 'save the match' is interpreted as 'prevent the defeat of one's team'.
3. 'Prevent' is also an implicative predicate, but of a different type than 'be able'. $X \text{ prevented } P$ implies that P did not take place.

These facts underlie the following inference chain: Aršavin could not save the match \Rightarrow does not take place: Aršavin saved the match \Rightarrow does not take place: Aršavin prevented the defeat of his team \Rightarrow does not take place: the team for which Aršavin played was not defeated \Rightarrow the team for which Aršavin played was defeated.

Let us take a look how the system formally makes these inferences.

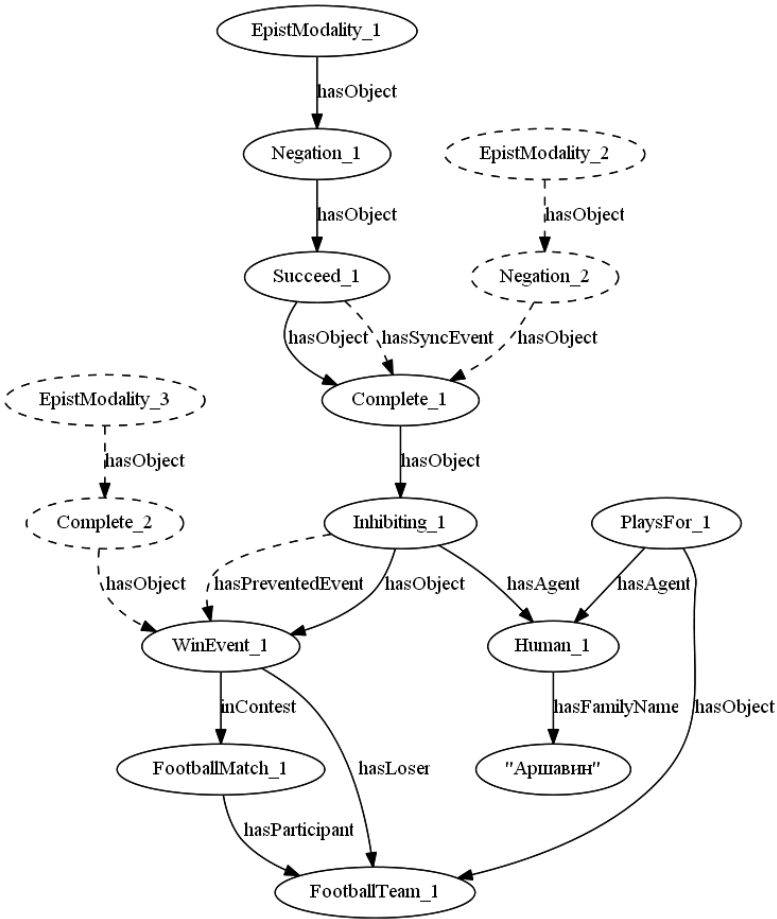


Fig. 2. Elements of BSemS (solid) and EnSemS (dashed) of the sentence *Aršavin tak i ne smog spasti matč* (Aršavin could not save the match)

In Fig. 2 we can see in solid lines the relevant elements of the BSemS which are created based on lexical and grammatical properties of the sentence. Nodes **Negation_1** **hasObject** **Succeed_1** correspond to the lexical items *ne smog* ‘could not’. Nodes **Human_1** **hasFamilyName** “**Аршавин**” correspond to *Aršavin*. **Complete_1** comes from the perfective aspect of *spasti* ‘save’. **EpistModality_1** marks the top predicate of the sentence indicating its facticity status. The rest (**Inhibiting_1** and down) comes from the lexical meaning of *spasti matč* ‘save the match’ and is created by a dictionary rule of the verb SPASAT ‘save’ shown below:

```

ZONE:SEM
DOMAIN:SPORT-DOMAIN
REG:SEM-CONV2.D0
TAKE:*
CHECK
  1.1 DOM-EQUN(X,Z,ПРЕДИК,Human)
N:1 // Иванов спас матч
CHECK
  1.1 DOM-LEXR(X,W,1-КОМПЛ,МАТЧ)
DO
  1 REPLACE-SEM(X,Inhibiting)
  2 REPLACE-SEM(W,FootballMatch)
  3 ADD-NODE-SEM(Z1,WinEvent)
  4 REPLACE-LINK([X,Z,*],[X,Z,hasAgent])
  5 LINK-NODES(X,Z1,hasObject)
  6 ADD-NODE-SEM(U,FootballTeam)
  7 REATTACH-NODE([X,W,*],[Z1,W,inContest])
  8 LINK-NODES(Z1,U,hasLoser)
  9 ADD-NODE-SEM(U3,PlaysFor)
  10 LINK-NODES(U3,Z,hasAgent)
  11 LINK-NODES(U3,U,hasObject)
  12 LINK-NODES(W,U,hasParticipant)

```

Once BSemS is created, inference rules are applied. First the definitions of the concepts **Succeed** and **Inhibiting** are processed, adding two relations **hasSyncEvent** and **hasPreventedEvent**. **hasSyncEvent** means that two events have the same facticity status, while **hasPreventedEvent** means that they have the opposite facticity status. The corresponding parts of the definitions are presented below:

Succeed ?x ->

```
?x    hasObject (Event ?event)
      hasSyncEvent ?event.
```

Inhibiting ?x ->

```
?x    hasObject (Event ?event)
h     asPreventedEvent ?event.
```

Then the inference rule for **hasSyncEvent** relation creates **EpistModality_2** and **Negation_2** nodes thus marking the fact that **Inhibiting_1** did not take place based on the fact that **Succeed_1** did not take place.

```
?event1 hasSyncEvent ?event2,
    EpistModality hasObject (Negation hasObject ?event1) ->
    EpistModality hasObject (Negation hasObject ?event2).
```

And finally the inference rule for **hasPreventedEvent** relation creates **EpistModality _ 3** and **Complete _ 2** nodes thus marking the fact that **WinEvent _ 1** took place based on the fact that **Inhibiting _ 1** did not take place.

```
?event1 hasPreventedEvent ?event2,
    EpistModality hasObject (Negation hasObject(Complete
hasObject ?event1)) ->
    EpistModality hasObject (Complete hasObject ?event2).
```

Since EnSemS contains a very large number of predications (up to several hundred) and is difficult to survey, the most convenient way to make sure that the analyzer obtained the expected inference is the question-answering option. In this option, the analyzer constructs the EnSemS of both the initial sentence, and the question, transforms the EnSemS of the question into SPARQL and infers the answer with the help of the RDFox reasoner. In Fig. 3 one can see the result of processing sentence (2) in the question-answering mode. In the upper window is the text (*Aršavin could not save the match*) and the diagnostic question (*Did Aršavin's team lose the match?*). The lower window contains EnSemSs of both sentences (only the last lines of the EnSemS of the question are seen) and the answer returned.

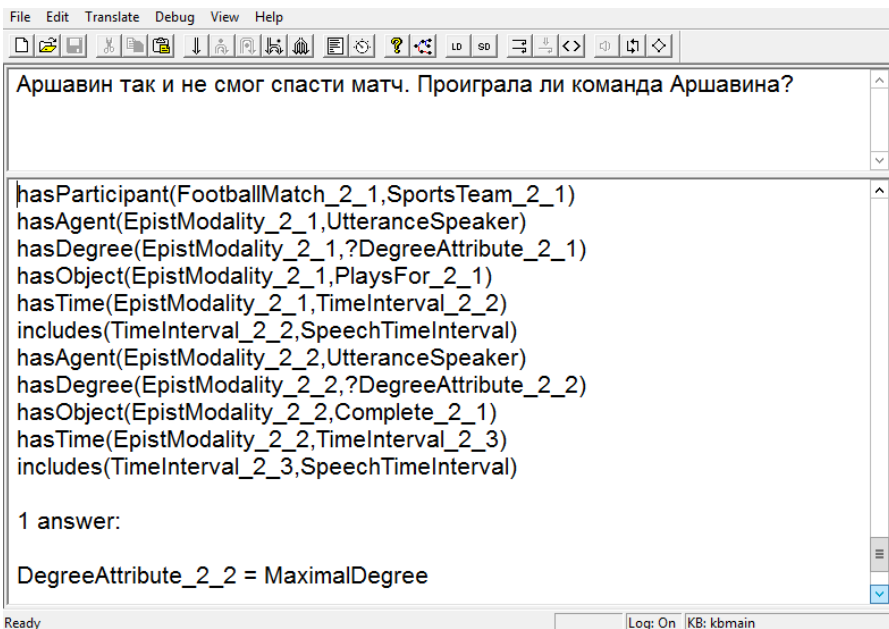


Fig. 3. EnSemS of sentence (2) and of the question and the answer obtained

EnSemS of the question can be glossed as follows: what is the value of the epistemic modality of the statement “the team for which Aršavin was playing was defeated”? In plain words, it means: is it true that Aršavin’s team was defeated? The answer, which can be seen in the lower window, reads that the value of this modality is maximal. This means that the question was answered in the affirmative.

6. Evaluation

For the evaluation, we selected 50 sentences the analyzer did not see before which are having to do with high spots of the match, i.e. the moments fraught with scoring a goal, such as attacks on the goal, direct free kicks, corners, goalkeeper interventions, etc. The sentences were extracted from the online running commentaries on the Football World Cup qualifying games. The analyzer constructed full EnSemS of all the sentences. The aim of the evaluation was to determine up to what extent the analyzer was able to extract explicit or infer implicit information on the main features of the high point. Concretely, we found that:

- 11 out of 50 test file sentences describe the situations that ended in scoring a goal. In 10 cases (91 %) the semantic analyzer successfully identified the goal event, in 5 cases (45.5 %) it also was able to identify the author of the goal event.
- 7 sentences out of 50 specify the distance from which a goal scoring shot was performed, and in 6 cases (85.7%) the semantic analyzer showed the correct distance.
- 12 sentences of the test file contain the information about the so-called “starting point” of the shot (which corresponds to the location of the person performing the shot). In 11 cases (91.6 %) the EnSemS indicate the starting point correctly.
- The terminal point of the shot (the part of the goal the shot is aimed at) is mentioned in 17 sentences. In 16 cases (94%) EnSemS indicate it correctly.
- Whenever the goalkeeper managed to prevent the goal event (6 sentences out of 50), the semantic analyzer showed the correct result (100 %).

7. Error analysis

The quality of semantic structures strongly depends on the accuracy of syntactic structures they are built on. 9 syntactic structures obtained for the test file contained some syntactic errors. Since our aim was to evaluate the semantic component of the system, we performed some interventions to the syntactic component in order to correct these errors.

The defects encountered in EnSemS are of the following types.

1. Wrong generation of an explicit subject in a noun phrase with a zero subject.
2. Wrong interpretation of the NP *štrafnaja X-a* ‘X’s penalty box’. The rule assumes that such a NP is only appropriate if X is the goalkeeper of the team to whom the penalty box belongs. However, one of the sentences of the test file refutes this supposition: in the sentence “*Ferreira Carrasco made a fault near his penalty box*” Ferreira Carrasco is not a goalkeeper.

3. The resolution of the non-anaphoric co-reference (*US President—Donald Trump*) requires that the individual in question be represented in the Repository of Individuals. Some of the players referred to in the test file are absent from RI, and therefore the co-reference between their mentions was not established.
4. Verbal tense is not always interpreted correctly. In Russian, the present tense may have a so called “commentary interpretation” (*nastojasščee reportažnoe*). The sentence *Ečše odin mjač zabivajut bel’gijtsy* ‘Belgium scores one more point’ denotes a terminated event although its verb is in the present imperfective. This use of present is typical of the genre of commentaries. Our rules fail to distinguish between the regular present and the present of commentaries.

Conclusions

The SemETAP semantic analyzer is an option of the ETAP-3 Linguistic Processor aiming at producing in-depth semantic interpretation of the Russian text. SemETAP makes use of both linguistic and extra-linguistic (background) knowledge, the former being stored in the Combinatorial Dictionary and the Grammar, and the latter—in the Ontology and the Repository of Individuals. Semantic analysis represents the text on two levels: Basic semantic structure (BSemS) interprets the text in terms of ontological elements. Enhanced semantic structure (EnSemS) extends BSemS by means of a series of inferences. An important feature of the analyzer is its capacity to infer implicit information, which is very useful for a variety of applications including question answering, story understanding, and dialogue processing.

Acknowledgements

This work was supported by the RSF grant 16-18-10422, which is gratefully acknowledged.

References

1. *Abreu, P. Faria, M., Reis L., Garganta, J.* (2010), “Knowledge Representation in Soccer Domain: An Ontology Development”. 5th Iberian Conference on Information Systems and Technologies (CISTI), 2010.
2. *Benedikt, M., Konstantinidis, G., Mecca, G., Motik, B., Papotti, P., Santoro, D., and Tsamoura, E.* (2017). Benchmarking the chase. In Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (pp. 37–52). ACM.
3. *Boguslavsky I.* (2017), Semantic Descriptions for a Text Understanding System. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2017), p. 14–28.
4. *Boguslavsky I., V. Dikonov, Frolova T., L. Iomdin, A. Lazursky, I. Rygaev, V. Sizov, S. Timoshenko.* (2016), Plausible Expectations-Based Inference for Semantic Analysis // Proceedings of the 2016 International Conference on Artificial Intelligence (ICAI’2016). USA: CSREA Press, 2016. pp. 477–483.

5. *Boguslavsky I., V. Dikonov, L. Iomdin, A. Lazursky, V. Sizov, S. Timoshenko.* (2015), *Semantic Analysis and Question Answering: a System Under Development*. In: *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2015)*, p.62.
6. *Bouayad-Agha, N., Casamayor, G., Wanner, L., Díez, F., López Hernández, L.* (2011), "FootbOWL: Using a generic ontology of football competition for planning match summaries". *Extended Semantic Web Conference ESWC 2011: The Semantic Web: Research and Applications*, pp 230–244.
7. *Cimiano Ph., Unger Ch., McCrae J.* (2014), *Ontology-based Interpretation of Natural Language. Synthesis Lectures on Human Language Technologies*. Morgan and Claypool Publishers.
8. *Gordon, Andrew S., and Jerry R. Hobbs.* (2004), "Formalizations of Commonsense Psychology", *AI Magazine*, Winter 2004, pp. 49–62.
9. *Gordon, Andrew S., and Jerry R. Hobbs.* (2011), "A Commonsense Theory of Mind-Body Interaction", in *Proceedings of the 10th Symposium on Logical Formalizations of Commonsense Reasoning, AAAI Spring Symposium Series*.
10. *Gordon, Andrew S., Jerry R. Hobbs, and Michael T. Cox.* (2011), "Anthropomorphic Self-Models for Metareasoning Agents", in *Michael T. Cox and Anita Raja (eds.), Metareasoning: Thinking about Thinking*, The MIT Press, Cambridge, Massachusetts, pp. 295–305.
11. *Hobbs, Jerry R., and Andrew Gordon.* (2008), "The Deep Lexical Semantics of Emotions", *Proceedings, LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology, and Terminology, Marrakech, Morocco, May 2008*.
12. *Hobbs, Jerry R., and Andrew Gordon.* (2010), "Goals in a Formal Theory of Commonsense Psychology", in *A. Galton and R. Mizoguchi (eds.), Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010)*, IOS Press, Amsterdam, pp. 59–72.
13. *Hobbs, Jerry R., Alicia Sagae, and Suzanne Wertheim.* (2012), "Toward a Commonsense Theory of Microsociology: Interpersonal Relationships", in *M. Donnelly and G. Guizzardi (eds.), Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012)*, IOS Press, Amsterdam, Netherlands, pp. 249–262.
14. *Montazeri, Niloofar, and Jerry R. Hobbs.* (2011), "Elaborating a Knowledge Base for Deep Lexical Semantics", in *J. Bos and S. Pulman (eds.), Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, January 2011, pp. 195–204.
15. *Montazeri, Niloofar, and Jerry R. Hobbs.* (2012), "Axiomatizing Change-of-State Words", in *M. Donnelly and G. Guizzardi (eds.), Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012)*, IOS Press, Amsterdam, Netherlands, pp. 221–234.
16. *Motik, B., Nenov, Y., Piro, R., Horrocks, I., and Olteanu, D.* (2014). *Parallel Materialisation of Datalog Programs in Centralised, Main-Memory RDF Systems*. In *AAAI* (pp. 129–137).
17. *Motik, B., Nenov, Y., Piro, R. E. F., and Horrocks, I.* (2015). *Handling Owl: sameAs via Rewriting*. In *AAAI* (pp. 231–237).

18. *Morgenstern, Leora.* (2001), MidSized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking. *Studia Logica*, April 2001, Volume 67, Issue 3, pp 333–384
19. *Nenov, Y., Piro, R., Motik, B., Horrocks, I., Wu, Z., and Banerjee, J.* (2015). RDFox: A highly-scalable RDF store. In *International Semantic Web Conference* (pp. 3–20). Springer, Cham.
20. *Rygaev I.* (2015), Implicative verbs in Russian and their semantic analysis in ETAP-3 linguistic processor. [Implikativnye glagoly v rusском jazyke i ix semantičeskij analiz v ramkax lingvističeskogo protsessora ETAP-3]. Master thesis, RGGU, 2015.
21. *Rygaev I.* (2017), Rule-based Reasoning in Semantic Text Analysis. Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR 2017 hosted by International Joint Conference on Rules and Reasoning 2017 (RuleML+RR 2017).
22. *Schmidt, Thomas.* (2006), “Interfacing Lexical and Ontological Information in a Multilingual Soccer FrameNet”. Proceedings of OntoLex 2006—Interfacing Ontologies and Lexical Resources for Semantic Web Technologies.
23. *Tsinaraki, C., Polydoros, Kazasis, F., and Christodoulakis, S.* (2005), Ontology-based semantic indexing for mpeg-7 and tv-anytime audiovisual content. *Multimedia Tools and Applications*, Vol. 26, Num .3, 2005, pp. 299–325.