Semantic Analysis Based on Linguistic and Ontological Resources

Igor Boguslavsky

Artificial Intelligence Department – Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain /
Computationl Linguistics Lab – IITP RAS
19, Bolshoj Karetnyj pereulok, GSP-4, 129447 Moscow

igor.m.boguslavsky@gmail.com

Abstract

We describe a recently launched project whose objective is to develop an advanced converter of natural language text to semantic structures. The project basically consists in enriching the ETAP-3 linguistic processor, developed by the Institute for Information Transmission Problems, Russian Academy of Sciences, with a new module – that of semantic analysis. An important feature of this module is that it will use not only linguistic knowledge incorporated in the grammar and the combinatorial dictionary, but also extralinguistic knowledge stored in the ontology and contextual information accumulated in the fact repository. We developed a small ontology that serves as the semantic metalanguage for Semantic Structures. Several examples are given that show how the ontology is used and how the meaning is represented.

Keywords

Semantic analysis, ontology, Semantic web, semantic rules, semantic structure

1 Introduction

Modern search engines such as Google, Yahoo, or Russian Yandex, have long come into our everyday life and we hardly imagine how we could do without them. Nevertheless, however useful these applications may be, they are rightfully reproached for “not understanding” the texts they are dealing with. They find far too many texts, while the overwhelmingly most part of them has nothing to do with what the user is asking about. On the other hand, if a text conveys the relevant meaning but it is expressed by words different from the ones used in the user’s query, this text will hardly be found at all. For many NL applications, first of all, for Information Retrieval and Extraction as well as for Question Answering, it is essential that they should be able to discover semantic similarity between the texts if they express the meaning in different ways.

As is known, most of the web content is created to be read by humans and not for meaningful operations on this content by machines. This is what brought to life the Semantic Web initiative which aims at making the web content more understandable for computers (Berners Lee et al., 2001). This requires, in particular, the elaboration of methods of representing semantic information contained in the texts and creating efficient technologies of handling it. However, there is a serious problem here. Annotating texts with semantic markup and developing resources is costly, and hardly anybody would be willing to work on that on a large scale unless there are valuable applications that will use these resources. But at the same time, applications are not likely to be developed before there exist semantic markup and resources. This is one of the reasons why Semantic Web is slow to realize its potential.

A natural way out of this impasse is automatic or semi-automatic semantic processing of NL texts. Advanced NLP systems will play a twofold role here. On the one hand, they could produce semantic annotation of texts on the massive scale, and on the other hand, they will benefit from the created resources themselves (Pall, 2006). To some extent, the situation is similar to the creation of treebanks. To build a large treebank, one needs an effective parser, but once a treebank exists, it can be used for creation of new parsers as well as for testing and evaluation of the existing ones.

Many NL applications need a much deeper semantic analysis of the text than is used today. In section 2 we will show with a simple example what perspectives deep semantic and encyclopaedic information opens for Information Retrieval and Question Answering. To make use of this information one has to rely on both linguistic and extralinguistic resources. In subsequent sections we will describe a system under construction at the Institute for Information Transmission Problems of the Russian Academy of Sciences and at the Technical University of Madrid which aims at combining both types of resources in order to produce semantic structures to be used in different applications. In section 3 we will briefly present the ETAP linguistic processor, and in section 4 – the SemAnOn project, which is responsible for developing the semantic analysis module for ETAP. Section 5 will highlight the domain ontology developed for the project. In section 6 we show that even simple class/subclass and part/whole relations can be effectively used for syntactic disambiguation and co-reference resolution. Section 7 will illustrate with a series of examples the kind of semantic structures we are striving for. In section 8 we list some directions of future research.

2 Information Extraction and Question Answering in Need of Semantic Analysis

Suppose we wish to find information about

(1) *the losses of warships during World War II.*

If we want to solve the task with Google, we have two main options. First, we could try "this exact wording" option. In this case, no match will be found. Then, we could look for the texts that contain all the words of the query, although not necessarily in the same order and maybe in different sentences. In this case, we obtain more than 16.800.000 links. Most of the texts found are noise, irrelevant for our quest. On the other hand, many of the documents which are relevant will be buried among hundreds of thousands of irrelevant texts, which is tantamount to not being found at all.

One of the ways for increasing the accuracy of information retrieval is taking into account the syntactic structure of both the query and the candidate text. If the query is *Working plan*, the relevant response should not be *We are working on the plan*, because syntactic relations between *working* and *plan* are different in the query and in the response. In our warship example, the syntactic structure of the query manifests the temporal relation between the *losses* and *World War II*. Therefore, we should skip all the texts where these words are not connected in this way.

However, the syntactic structure is by far not sufficient for real understanding of the query. We should also be aware that *the loss (of the ship)* means that the ship ceased to exist (FinFunc$_0$), and for this reason belongs to the same semantic class as the nouns *death, wreck, crash,* etc. These nouns are (quasi)synonyms and could be found in WordNet. But we should also understand that there exist a number of verbs that denote the same situation, such as *to die, to sink, to perish, to wreck,* etc. (Note that the noun-verb type of synonymy is not reflected in WordNet). Moreover, we need not only the words that denote the same or a similar situation. An important semantic relation is entailment. In our case, it is essential to know that the end of existence of a thing can be a result of different events of the LiquFunc0 class, such as *to destroy, to kill, to explode, to sink, to eliminate, to exterminate, to liquidate,* etc. A semantic structure that conveys all this information could look like this.

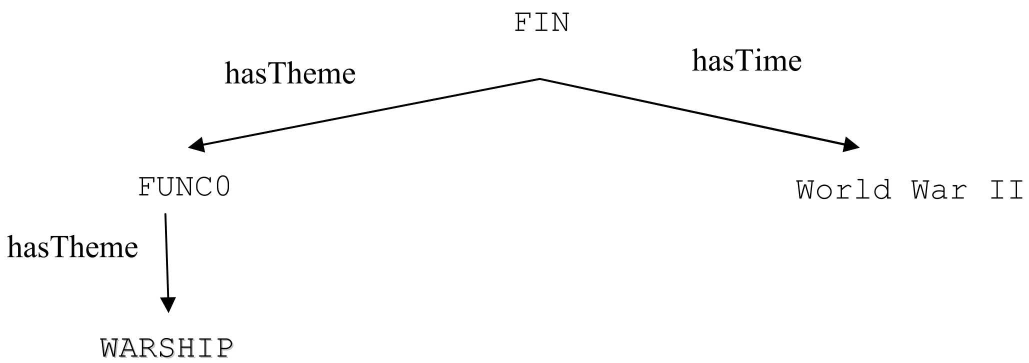

Fig. 1 Semantic structure of query (1): *the losses of warships during World War II.*

Besides semantic information that pertains to linguistic competence and should be incorporated in the linguistic description, there are other important sources of knowledge necessary for understanding the text. This is first of all the encyclopaedic knowledge (contained in an ontology) and an inference mechanism responsible for drawing conclusions from all the information available.

Let us show how the combination of these resources helps ensure a more accurate processing of our example. First, we should enrich the semantic structure in Fig. 1 with encyclopaedic information. Given that World War II began in 1939 and ended in 1945, the representation of the phrase *during World War II* can be made more precise; cf. Fig. 2.

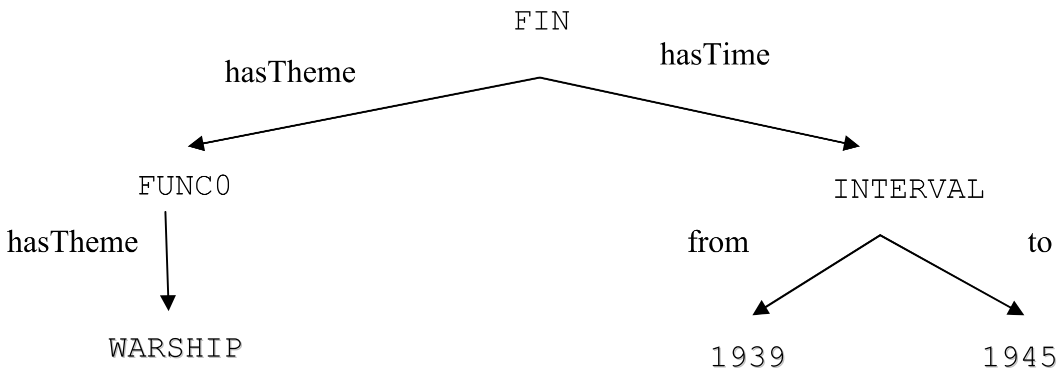

Fig. 2. Semantic structure of query (1) enriched with encyclopaedic information.

Now it is not difficult to see that texts (2) and (3) correspond to query (1), although they are composed of totally different words and are therefore completely beyond the possibilities of word-based retrieval methods.

(2) *On May, 27 of 1941 the Royal Navy destroyed the German battleship "Bismarck".*

To make sure that (2) corresponds to the semantic structure in Fig. 2, one should have access to the following information:

(a) *battleship* is a kind of a warship;

(b) *X destroyed Y* entails that 'Y ceased to exist';

(c) the date *May 27, 1941* belongs to the interval between 1939 and 1945.

This information by far surpasses the data available in NLP systems. Although the data on class-subclass relations, as manifested in (a), can be found in WordNet, (b) requires semantic decomposition rules, and (c) involves reasoning about time.

To come to the conclusion that phrase (3) also corresponds to query (1), one needs to consult an encyclopaedia and find that HMS Hood is a battle cruiser, which is also a member of the warship class.

(3) *the unexpected wreck of Hood at the 8th minute of the battle on May, 21, 1941.*

At the same time, comparing semantic structure of the query with the semantic structures of the texts permits to disqualify the texts which do not match the query. Cf. (4) which, as well as (2) and (3), refers to a wreck of a warship, but in a wrong time span.

(4) *The second film of the series is dedicated to the wreck of the 57-canon motor sailing frigate "Oleg", perished in August 1869.*

Concluding, to implement semantic search, the following resources are necessary:

- a sophisticated linguistic model. It should not only cover domain terms, but also be able, in particular, discover semantic identity in different syntactic contexts (cf. *the wreck of the frigate vs. the frigate whose wreck became known...*). It should be able to handle collocations, modality, and negation; cf. *suffered* [$Oper_1$] *wreck* (relevant for query (1)) vs. *avoided a wreck* (irrelevant)*; didn't suffer wreck* [neg + $Oper_1$] (irrelevant) vs. *didn't avoid a wreck* (relevant);

- various external ontological and encyclopaedic resources and the capacity to integrate them with NLP modules (WordNet, SUMO, FreeBase, DBpedia);

- logical inference engine;

- a large semantic index, which covers the corpus where the search is supposed to be carried out.

3 ETAP-3 Linguistic Processor

The multifunctional ETAP-3 linguistic processor, developed by the Computational Linguistics Laboratory of the Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, (see e.g. Apresjan *et al*. 2003), is the product of decades of research and development in the field of language modelling. At the moment, ETAP-3 consists of a number of options, including

1) a rule-based machine translation system working both ways between Russian and English (plus several prototypes for other languages – French, German, Spanish, Korean and Arabic);

2) a system of synonymous and quasi-synonymous paraphrasing of sentences (Apresjan, Tsinman, 2002);

3) an environment for deep annotation of text corpora, in which SynTagRus, the only corpus of Russian texts tagged morphologically, syntactically (in the dependency tree formalism), and lexically was created (Boguslavsky et al., 2000a), and

4) a Universal Networking Language (UNL) module, responsible for automatic translation of natural language text into a semantic interlingua, UNL, and the other way around (Boguslavsky et al., 2000b).

The ETAP-3 processor is largely based on the general linguistic framework of the Meaning ⇔ Text theory by Mel'čuk. An important complement to this theory was furnished by the theory of systematic lexicography and integrated description of language proposed by Jurij Apresjan (Apresjan, 2000).

For each sentence the processor successively builds several representations: Morphological Structure (MorphS), Syntactic Structure (SyntS), and Normalized Syntactic Structure (NormSyntS). Some options of ETAP-3 make use of a deeper representation. It is a UNL Graph in the UNL option, and Semantic Structure in the semantic analysis option, described below.

One of the major resources used in ETAP-3 is the combinatorial dictionary. It offers ample and diverse data for each lexical entry. In particular, the entry may list the word's syntactic and semantic features, its subcategorization frames, as well as rules (or reference to rules) of a dozen types, which make it possible to describe peculiar behaviour of individual words and exceptions to general rules in a complete and consistent way. Many dictionary entries contain information on lexical functions.

The entry of the combinatorial dictionary has a number of zones, one of which provides the properties of the word that are manifested in the given language, while all the other zones contain information on the match between this word and its equivalent in a particular language. For example, the EN zone in the Russian combinatorial dictionary entry contains information on the translational equivalents of the respective Russian word into English. One field (TRANS) gives the default single-word translation (or several such translations) of this word in English. Other fields contain less trivial translation rules, or references to such rules.

A newly introduced ONTO zone offers information underlying the match between the Russian word and its counterparts in the ontology.

4 SemAnOn project

A new project underway at IITP is aiming at moving ETAP further towards semantics. We add a new module – that of SEMantic ANalysis based on ONtology (SemAnOn). Its task is to transform NormSyntSs into Semantic Structures (SemS).

Our analyser is rule-based, which may seem bizarre nowadays when most of the analysers, both shallow and deep, are statistical. Cf., for example, interesting results obtained in supervised and unsupervised semantic parsing in Ge and Mooney, 2005; Poon and Domingos, 2009; Clark et al. 2010. A combination of machine learning and rule–based approaches is used for semantic processing in Moldovan et al., 2010. Our choice of the 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 analyser, it will open the possibility to develop such a corpus, which could then be used for refining and evaluating the analyser, as well as for developing other semantic parsers. The second, even more important, reason for our non-statistical approach is our firm belief that the modelling of real understanding of texts requires knowledge-intensive methods.

Our approach to semantic analysis is closely related to the OntoSem approach (Nirenburg, Raskin, 2004), although linguistic frameworks adopted in these projects are largely different. Semantic analysis will be done in two steps. First, Basic Semantic Structures are produced, which present literal meaning of the sentence to the extent it can be extracted from the sentence itself. Then, they are transformed into Extended Semantic Structures, that are enriched with ontological and contextual information available.

We are going to use two new types of resources: an ontology and a fact repository The ontology is a collection of concepts connected with relations and provided with attributes and rules. The fact repository accumulates semantic structures which store data about concrete situations.

These resources are to be used for two purposes. On the one hand, they are the basis of semantic analysis. As the matter of fact, ontology provides the metalanguage of semantic structures, so that semantic structures interpret natural language sentences in terms of the ontology. On the other hand, they can be used for disambiguation.

In this paper, we will not give a systematic description of the SemAnOn project. Instead, we will give several examples that illustrate some of its aspects. More details on the project can be found in Boguslavsky et al. 2010.

In the next section, we will demonstrate the small domain ontology we built for the project.

5 Football ontology

The ontology we are working with focuses in the first place on football. It contains information on teams, players, football field, sport events, and their properties. However, we want it to be extendable to other sports as well. That is why some classes are more general than would be needed for football alone. For example, instead of having one class `FootballPlayer`, the ontology has a more general class `Sportsman`, of which `FootballPlayer` is a subclass. An equivalence restriction states that `FootballPlayer` is a `Sportsman` whose `SportType` is `football`. In this way, sportsmen doing different types of sports can be treated by the ontology in a uniform way.

The football ontology is written in SWRL (Semantic Web Rule Language), which is OWL (Ontology Web Language) augmented with rules (Horrocks et al. 2004). In compiling it, we used some existing ontologies dealing with football (e.g. http://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.daml). As usual, properties of classes are inherited by the subclasses. For example, the `Match` class is a subclass of `SportEvent`, which in its turn is a subclass of `Event`. `Match` inherits from `Event` the properties of having definite `Time` and `Place`. From `SportEvent` it inherits the fact that its participants should be `SportAgents`. Its own properties are: the number of participants is 2 (as opposed to championships, which have more) and it has a definite sport type (as opposed to Olympics, which involve many sport types). A subclass of `Match` is `Derby`, in which both participants should be from the same city or region. This property is implemented by means of a SWRL rule. Another rule assigned to `Match` states that if its `SportType` is football (or any other team sport), then its participants should be teams and not individual sportsmen, as is the case in tennis or chess. `Sportsman` is a subclass of two classes: `Person`, from which it inherits the property of having a name and a birth date, and `SportAgent`, which includes also `Team` and from which it inherits the property of having a definite sport type and a coach.

6 Ontological information in parsing

As is well-known, an important source for the increase in the parsing quality is restrictions on the semantic class of arguments of predicates. This information exists in the ETAP-3 combinatorial dictionary for a long time. The restrictions are based on the mini-thesaurus that contains several dozen concepts. In the future, we hope to replace it with a full-fledged upper and middle ontology. However, besides checking semantic agreement between the predicate and its arguments, there are many other tasks that are solved at the stage of parsing and that can be enhanced by using information available in the ontology. In the first place, it is the resolution of syntactic and lexical ambiguity. Let us illustrate it with one example, which manifests syntactic ambiguity:

(5) *Zaly dlja priemov inostrannyx delegatsij, steny kotoryx byli ukrašeny lepninoj* 'halls for the reception of foreign delegations whose walls were covered with stucco moulding'

The relative clause 'whose walls were covered with stucco moulding' can be syntactically linked to any of the three preceding nouns – 'halls', 'reception' и 'delegation'. In order to select among them the one which is most plausible semantically, we have to check which of the phrases – *walls of the halls*, *walls of the reception* or *walls of the delegations* – makes more sense. Obviously, it is the first of them that can be more easily comprehended, although on terms of word order *halls* is further than other candidates from the relative clause. The semantic agreement between 'halls' and 'walls' can be easily discovered by means of the ontology. Indeed, concepts `Hall` and `Wall` are linked in the ontology by a short semantic chain consisting of relations "is a" и "part of": a `Wall` is a typical part of `Premises`, and a `Hall` is a kind of `Premises` and therefore inherits all its properties. For two other pairs of words – *walls* / *reception* и *walls* / *delegations,* one cannot find in the ontology such a natural connection.

7 NL Words and Semantic Elements

As mentioned above, the ontology serves as a semantic metalanguage, which means that SemSs are composed of the ontology elements – concepts, instances and properties (henceforth, semantic elements). All meaning bearing NL units are defined in terms of these elements. The correlation between NL words and semantic elements is far from being straightforward. Here several situations are possible. Let us look at some of them.

1. The simplest case: a NL word corresponds directly to an ontology element. For example, such words as *to win, to defeat, to lose, victory*, etc. all correspond to concept `WinEvent`.

2. A NL word does not have any ontological equivalent and simply disappears in SemS. In particular, this is the case of Lexical Functions of Oper-Func-Labor family. For example, sentence

(6) *Messi scored a goal*

has SemS

(6a) `hasAgent(GoalEvent, Messi)`

3. A NL word generates a fragment of SemS but it has no fixed ontological equivalent. Its interpretation depends on the context. Such words are particularly difficult for the semantic analysis. Let's consider the word *local*. In one of its meanings it denotes a place which is activated in the given context. For example, in the sentence

(7) *In Malaga we defeated the local team 2:1*

the *local team* is a team from Malaga. In the sentence

(8) *A deadly accident involved a local team at the Arizona Nationals*

the same phrase refers to a team from Arizona. In each case, the semantic analysis should discover which place is referred to by *local*. In sentences (7)-(8) above this place is directly mentioned in the sentence. However, this is not always the case. Sometimes the situation is more involved. The place referred to by *local* may be absent from the sentence altogether. Sentence

(9) *Charlemagne's remains were interred in the local cathedral*

is only appropriate if the place (the German city of Aachen) has already been activated in the previous text. Often, the place referred to by *local* is the place where the situation described in the sentence takes place.

(10) *The forum will allow local dealers to meet with the representatives of the manufacturer.*

Although this sentence doesn't provide any information on where the forum will take place, we know that the dealers are from the same place, and this fact can be used for inference. SemS of this sentence will contain the following fragment:

```
hasLocation(Forum01, Place01)
basedIn(Dealer03, Place01).
```

Let's give two examples more.

(11) *In the quarters Ahli Tripoli defeated Qatari 3:1. Their namesakes from Benghazi lost 0:3 to Saudi Arabia's Hilal.*

The first sentence contains a reference to the team whose name is *Ahli* and which is based in the city of Tripoli. The second sentence speaks of another team whose name is given by a reference to the first team by means of a shifter (*namesake*):

```
hasName(Team01, Ahli)
basedIn(Team01, Tripoli)
hasName(Team02, Ahli)
basedIn(Team01, Benghazi).
```

(12) *On May 28 Chelsea received Juventus. The visitors won 1:0.*

To understand where the match took place and who won, the system has to manipulate both semantic and encyclopaedic knowledge. First, it has to know that if a team receives another team, the match is played in a place which is the home for the first team but not for the second. This fact is accounted for by the semantic definition of *receive,* which is a rule that operates between NormSyntS y SemS:

Receives(X, Y) ⇔
```
hasParticipant(Match01, X)
hasParticipant(Match01, Y)
```

```
hasLocation(Match01, Place01)
basedIn(X, Place01)
basedIn(Y, Place02)
defferentFrom(Place01, Place02)
```

Besides that, there is a rule stating that a team is called *visitors* if it is not playing at home:

Visitors(X) ⇔
```
hasParticipant(Match01, X)
hasLocation(Match01, Place01)
basedIn(X, Place02)
defferentFrom(Place01, Place02)
```

Second, the system should possess the encyclopaedic information that Chelsea is based in London, and Juventus in Turin. With this information available, the system will construct SemS of the text and restore the following implicit information:

```
hasLocation(Match01, London)
hasWinner(Match01, Juventus)
```

4. An ontological concept does not correspond to any word of the sentence. This happens when the meaning is not expressed lexically but syntactically. For example, sentence (13) gives information on the match score but does not contain the word *score:*

(13) *FC Barcelona defeated Manchester United 3:1*

SemS of this sentence looks as follows:

```
hasWinner(WinEvent01, FC_Barcelona)
hasLoser(WinEvent01, Manchester_United)
inMatch(WinEvent01, Match01)
inMatch(MatchScore01, Match01)
hasValue1(MatchScore01, 3)
hasValue2(MatchScore01, 1)
inFavorOf(MatchScore01, FC_Barcelona)
```

8 Future research

As mentioned above, the project is at the initial stage. The coverage of the semantic rules is still small, although the syntactic parser, which constitutes an integral part of SemAnOn, has a very large coverage. In the near future we are planning to extend the coverage of the semantic analyzer to different types of constructions, in particular to the ones expressing intervals, temporal expressions and quantified noun phrases, to complete the football domain ontology and combine it with an upper/middle ontology, such as SUMO. Later, we hope to reformulate semantic restrictions in the subcategorization frames in terms of the ontology and use it widely for disambiguation tasks. Another direction of future research is connected with the construction of the fact repository.

Acknowledgements

The work reported in this paper has been partly supported by grants RFBR № 11-06-00405 and RSFH № 10-04-00040a which is gratefully acknowledged.

Bibliography

Apresjan, Ju. D. (2000). *Systematic Lexicography*. Oxford University Press, 2000, XVIII p., 304 p.

Apresjan, Jury, I. Boguslavsky, L. Iomdin, A. Lazursky, V.Sannikov, V.Sizov, L.Tsinman. (2003). ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. In: *MTT 2003, First International Conference on Meaning – Text Theory (June 16-18 2003).* Paris: Ecole Normale Supérieure, P. 279-288.

Apresjan, Yu., Tsinmam L. (2002). Formal'naja model' perifrazirovanija predlozhenij dlja sistem pererabotki tekstov na estestvennyx jazykax. *Russkij jazyk v nauchnom osveshenii,* № 2 (4), pp. 102-146.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American,* pp. 96-101, May 2001.

Boguslavsky, I., Grigoriev, N., Grigorieva, S., Kreidlin, L., Frid, N. (2000a). Dependency Treebank for Russian: Concept, Tools, Types of Information. *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, , p. 987-991

Boguslavsky, I., Frid, N., Iomdin, L., Kreidlin, L., Sagalova, I., Sizov, V. (2002b). Creating a Universal Networking Module within an Advanced NLP system. *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, vol. 1, 83-89.

Boguslavsky, I., Iomdin, L., Sizov, V., Timoshenko, S. (2010). Interfacing the Lexicon and the Ontology in a Semantic Analyzer. In: Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), Beijing, August 2010, pp. 67–76.

Clarke, J., D. Goldwasser, M. Chang and D. Roth. (2010). Driving Semantic Parsing from the World's Response. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010).*

Horrocks, I., P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof and M. Dean. (2004). SWRL: A Semantic Web Rule Language Com-bining OWL and RuleML. In: *W3C Member Submission 21 May 2004.*

Moldovan, D., Tatu, M., Clark, Ch. (2010). Role of Semantics in Question Answering. In: *Phillip C.-Y. Sheu, Heather Yu, C. V. Ramamoorthy, Arvind K. Joshi, Lotfi A. Zadeh (Eds.) Semantic Computing, pp. 373-420.*

Nirenburg, S., and Raskin, V. (2004). *Ontological Semantics.* The MIT Press. Cambridge, Massachusetts. London, England.

Pall, Barney. (2006). POWERSET - Natural Language and the Semantic Web, invited lecture at *The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*, (http://videolectures.net/barney_pell/)

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).

Ruifang Ge, Raymond J. Mooney. (2005). A Statistical Semantic Parser that Integrates Syntax and Semantics. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning.* Ann Arbor, MI, pp. 9--16, June 2005.