Combining Different Knowledge Sources for Text Understanding

Igor Boguslavsky
Computational Linguistics Dept.
IITP RAS / UPM
Moscow/Madrid, Russia/Spain
0000-0003-3390-1449

Vyacheslav Dikonov
Computational Linguistics Dept.
IITP RAS
Moscow, Russia
sdiconov@mail.ru

Tatiana Frolova
Computational Linguistics Dept.
IITP RAS
Moscow, Russia
tfrolova@gmail.com

Leonid Iomdin
Computational Linguistics Dept.
IITP RAS
Moscow, Russia
iomdin@gmail.com

Alexander Lazoursky
Computational Linguistics Dept.
IITP RAS
Moscow, Russia
lazursky@mail.ru

Ivan Rygaev
Computational Linguistics Dept.
IITP RAS
Moscow, Russia
irygaev@jent.ru

Svetlana Timoshenko
Computational Linguistics Dept.
IITP RAS
Moscow, Russia
nyrestein@gmail.com

***Abstract*— We describe a project under development whose objective is to build a model of natural language text understanding. This model has a form of a converter of a natural language text to its semantic structure. The project basically consists in enriching the ETAP-4 linguistic processor, developed at the Institute for Information Transmission Problems (IITP) of the Russian Academy of Sciences, with a new module – that of semantic analysis. An important feature of this module is that it uses 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 repository of individuals. We developed an ontology that serves as the semantic metalanguage for semantic structures. This knowledge permits to make different types of inferences, which extract implicit information. Several examples are given that show how all these knowledge sources can be used, how the meaning is represented and what inferences can be made.**

Keywords—semantic parsing, semantic structure, inference, ontology, linguistic knowledge, background knowledge

I. INTRODUCTION

One of the prerequisites for the full-fledged human-computer interaction is that the computer has to understand natural language (NL). By natural language understanding one usually means the capacity of the computer to transform the text into its semantic representation, which can subsequently be used for different purposes – for question answering, text summarization, information extraction, translating into another natural language, conducting dialogue, and others. Many applications need a deeper semantic analysis of the text than is usual in the state-of-the-art systems.

We describe a project in progress whose objective is to develop an advanced semantic analyser, i.e. a converter of a natural language (Russian) text to semantic structures (SemS). This task is conceived as a deep NL understanding system that uses extralinguistic knowledge and common-sense reasoning. We proceed from the assumption that the more inferences from the text the system can draw the fuller the understanding.

The SemETAP general purpose semantic analyzer is under development at IITP RAS. It is a new module of ETAP-4, which is currently able to analyze and generate texts in a number of languages. The semantic module reutilizes the morphological and syntactic analyzers of ETAP-4 and interprets the meaning of the text by constructing its semantic structure. An important feature of this module is that it not only represents the meaning of the text but also enriches it with a series of inferences, which permits to extract implicit information.

As opposed to several semantic parsers based on machine learning, our parser is strictly rule-based. 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 our parser reaches maturity, it will open the possibility to develop such a corpus, which could then be used for refining and evaluating the parser, as well as for developing other semantic parsers. The second, and more important, reason for our approach is our firm belief that modelling deep understanding of texts requires knowledge-intensive methods.

Although our system is under construction, it has been tested in various experiments designed to test the capacity of the system to extract implicit information and make inferences based on background knowledge. In [1] we tested the analyser on the task of interpreting high spots of the football match. In [2] it was used for resolving the Winograd Schema Challenge. [3] describes how SemETAP resolves the Triangle-COPA dataset, which tests the interpretation of social scenarios. In all the experiments reasonably good results have been obtained.

In this short paper, we cannot describe all the components of the model. They are described in our previous publications ([1, 2, 3 and references therein]).

II. Related Work

There are several directions in which semantic processing relying on ontologies is currently carried out. Some researchers understand semantic analysis as tagging the text by semantic elements, such as WordNet synsets, ontology classes or individuals, semantic roles, or FrameNet frames [4], [5], [6]. Others view their task in abstracting away from syntactic details, i.e. getting rid of grammatical words, such as auxiliary verbs, prepositions and conjunctions, and establishing semantic relations between semantically loaded words [7], [8]. In a different approach, some authors attempt to translate natural language sentences into logical formulae [9], [10], [11], [12]. Many papers focused on semantic analysis tend to use, in addition to linguistic data, also background information contained in the ontology [13], [14]. Most of the successful semantic analyzers are developed within the machine learning paradigm, especially under supervised learning [15], [16], [17], [18], [19]. 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 [20].

III. SemETAP system

Our approach differs from previous research in several important respects. Its salient features include:

1) The system is based on explicit knowledge. An important advantage of this approach (wrt machine learning approaches) is that it gives an opportunity to provide an explanation of the result understandable for humans.

2) Intensive use of both linguistic and background knowledge.

3) Semantic analysis with inference allows us to extract implicit information. Inference rules applied by the reasoner are written in a new logical formalism Etalog [21].

4) Many words and concepts of the ontology are supplied with explicit decompositions for inference purposes.

5) 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. Going beyond the sentence boundaries is essential for finding antecedents of pronouns which are very often located in one of the preceding sentences.

6) 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 inferences.

7) Two types of inferences are carried out: 100%-true logical entailments, and implications that implement plausible expectations.

IV. Semantic analysis.

SemETAP has a large amount of knowledge which is distributed between several resources. Language knowledge is stored in the Morphological and the Combinatorial dictionaries and in several sets of rules. World (background) knowledge is concentrated in the Ontology, Repository of individuals and inference rules.

Before being passed to semantic analysis, the text is subjected to morphological analysis, dependency parsing and normalization. Semantic analysis consists in two major steps. First, Normalized Syntactic Structures of all the sentences of the text are individually transformed into Basic Semantic Structures (BSemS), which reflect the meaning directly conveyed in the sentence. At this canonization stage, missing arguments are restored, some syntactic constructions are further normalized and semantically void collocates are eliminated. Then all meaningful Russian words are replaced by their semantic definitions.

Second, BSemSs are enriched by inference rules which contain the detailed information on the concepts and world knowledge and thus convert BSemS to Enhanced Semantic Structures (EnSemS).

V. Inferences

The main task of SemETAP consists in constructing a semantic representation of the text that contains all the inferences triggered by the system knowledge. The ability to draw inferences is the main manifestation of our understanding of the text. The more inferences we can make, the better we understood the text. The inferences we can make can be classified by two categories - by the knowledge source and by the reliability degree. In section VI we will show how a series of inferences can help us obtain the semantic interpretation of the sentence.

A. Inference knowledge source

The knowledge needed to make an inference may come from different sources – from the Combinatorial dictionary of language, Ontology and Repository of Individuals. First, it may be contained directly in the lexical meaning of the word. For example, if "somebody was told something", then "he has this information". If "somebody was made to go away", then "he went away". If "somebody kept his promise to come", then "he came". If "somebody missed an opportunity to go to the concert", then "he did not go there".

Second, the knowledge needed for the inference may belong to the category of common sense. Here are some examples: "People usually tend to avoid what makes them feel negative emotions", "If somebody is hungry, he/she has a goal to eat", "A precondition of eating is having food", "If X defeated Y, this is good for X and bad for Y", "If X and Y are adversaries, they have contrary goals". Such information is part of the ontological description of the corresponding concepts (`NegativeFeeling`, `BeHungry`, `Eating`, `WinEvent`, `BeAdversaries`).

Third, the knowledge needed may be related to concrete individuals and be stored in the Repository of Individuals. Let us show by an example how all the three resources – the Combinatorial Dictionary, the Ontology and the Repository of Individuals – contribute to the interpretation process. Let us look at the sentence (1) extracted from a commentary on a football match:

(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 resulted in the kick point blank by Necid, but Dikan was up to the mark'.

Suppose we want to know if a goal has been scored. Obviously, this sentence does not give a direct answer to this question. To answer this question, the reasoner will have to recur to three sources of information mentioned above – the Combinatorial dictionary, the Ontology and the Repository of Individuals:

- The Combinatorial Dictionary tells us that the expression *byt' na vysote* 'be up to the mark' corresponds to the concept `EqualToOccasion`.
- The Ontology interprets this concept as 'do well what one is expected to do' (concept descriptions are done in a specialized logical language – Etalog, but for readers' convenience, we render them here and below by means of a NL gloss);
- The Repository of Individuals contains the information that Andrei Dikan is a goalkeeper of Spartak Football Club;
- The Ontology describes the goalkeeper role as preventing the ball from penetrating the goal of his team.
- Placing the ball into the goal of the opposite team is qualified by the Ontology as a goal.

These five pieces of information allow the reasoner to infer that Dikan, being a goalkeeper, has the function of preventing the ball from entering his goal. Since he performed his function well, the ball did not get into the 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.

It's worth empathizing that a conclusion concerning scoring a goal has been made in the context which does not mention the word *goal* nor any of its synonyms.

B. *Inference reliability degree.*

When it comes to making inferences from NL sentences, usually the first order logic is used. However, it cannot account for all the types of reasoning used in everyday communication. It is not sufficient if only because first order logic does not allow for exceptions, while everyday communication is permeated with them. The most obvious example of this is the universal quantifier. The words that correspond to the universal quantifier are used in NL by far more freely than in the language of logic. From the point of view of logic, *always* just means 'always', i.e. 'it can never be otherwise'. But in NL one can easily say *I always get up at 7, but today my alarm clock broke and I got up at 9.*

In SemETAP, we distinguish two degrees of confidence of the speaker in the inference reliability. There are 100% reliable implications and there are plausible expectations (implicatures). The latter are natural to expect in the given situation, but they can turn out false. For example, if somebody says *John broke the cup,* we can safely infer that the cup lost integrity. However, if somebody says *John dropped the cup,* it would be natural to expect that the cup is broken, but this expectation may be not confirmed.

It should be noted that a sentence may give rise to both types of inference at a time. For example, the BSemS of the sentence *John went to the university (at moment t)* claims that 'at moment t John began moving towards the university with the aim of being there'. One can make two inferences from this BSemS with different reliability. The first one is a logical implication and therefore is definitely true: 'at t John finished being at the starting point of his movement'. The second inference is simply a plausible expectation: 'one can expect that at some t_1>t John will be at the university'. This does not exclude the possibility that on the way John changed his mind and went to the movies.

The role of plausible expectations in the text interpretation is difficult to overestimate. As is known, the speaker does not express explicitly all the information that the addressee extracts from the text. He often omits pieces of information that can be easily recovered by the addressee. It is often by means of plausible expectations that such information is recovered.

VI. Question Answering

Inferences described above open the possibility to extract a large amount of implicit knowledge. This knowledge may be used for different purposes, one of which is question answering. We implemented a question answering option that can process questions even if they do not contain the same words as the text and require deep semantic analysis of both the text and the question. For example: Text: *Peter took an opportunity to leave.* Question: *Did Peter leave?* Answer: Yes. Text: *Peter is Mary's husband.* Question: *Who is married to Peter?* Answer: *Mary.*

This option is not only interesting in itself. It is very useful in evaluating semantic analysis. 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. We will illustrate this option below, in section VII.

VII. Semantic interpretation of a sentence by means of a series of inferences.

In this section we will give some examples that show how inferences can help obtain semantic interpretation of the sentence. First, we will take one more sentence from the football commentaries with which we tested our semantic analyzer:

(2) *Ronaldu tak i ne smog spasti matč* 'Ronaldo could not save the match'.

Let us see how the analyzer comes to the conclusion that in sentence (2) the team for which Ronaldo was playing was defeated.

In Fig.1, we show the BSemS obtained for sentence (2).

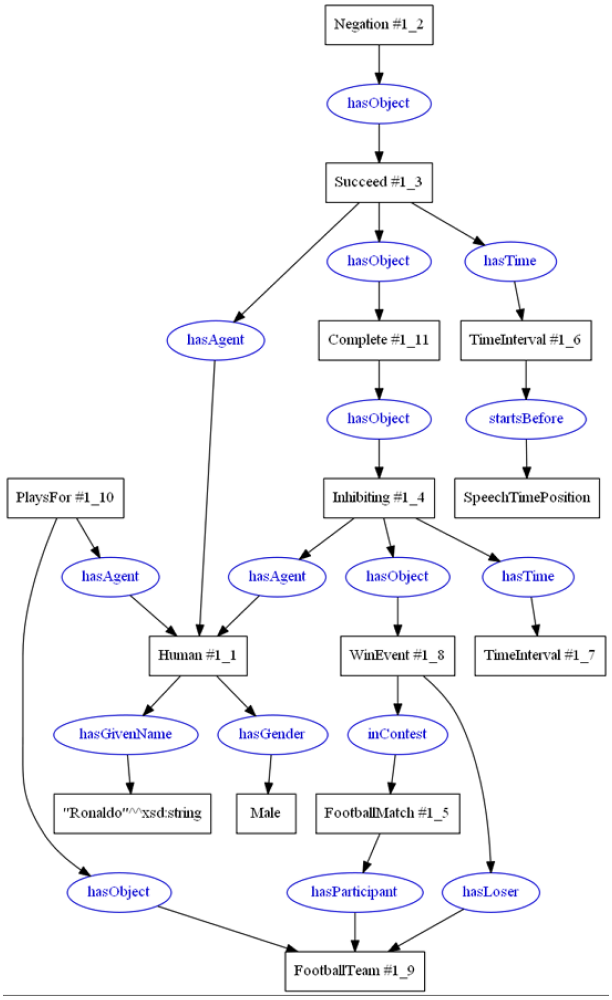

Fig. 1. BSemS of the sentence *Ronaldu tak i ne smog spasti matč* 'Ronaldo could not save the match'.

This structure can be "read" as follows: "the human #1_1, whose name is Ronaldo, who is a male and plays for team #1_9, did not succeed in inhibiting defeat of team #1_9 in a football match in which it participated. All this happened in the past (= before time of speech)".

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

(3a) The verb *smoč* 'be able', in the perfective aspect, is implicative in the sense of [22]. This means that it activates two implications: *X smog sdelat' P* 'X could do P' implies that P took place, while *X ne smog sdelat' P* 'X could not do P' implies that P did not take place.

(3b) The phrase *spasti matč* 'save the match' is interpreted as 'prevent the defeat of one's team'.

(3c) 'Prevent' is also an implicative predicate, but of a different type than 'be able'. *X prevented P* implies that P did not take place.

(3d) Double negation: 'it is not true that P does not take place' implies 'P takes place'.

These facts underlie the following inference chain performed by the reasoner:

(2) *Ronaldo could not save the match*

⇒ does not take place: Ronaldo saved the match [from (3a)]

⇒ does not take place: Ronaldo prevented the defeat of his team [from (3b)]

⇒ does not take place: Ronaldo's team was not defeated [from (3d)]

⇒ Ronaldo's team was defeated.

In Fig. 2 one can see the result of processing sentence (2) in the question-answering mode. As mentioned above, this mode is used to make sure that the inference produced the desired result.

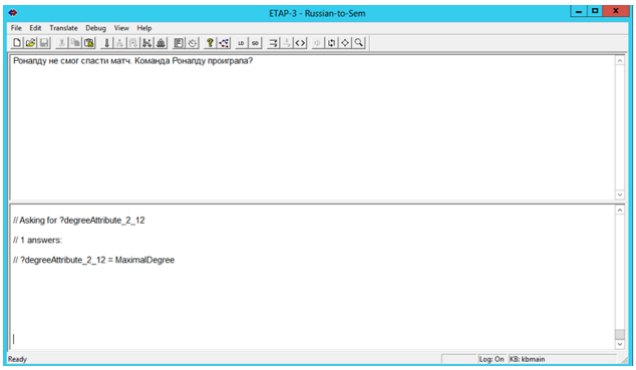

Fig. 2. Sentence (2) and the question *Was Ronaldo's team defeated?*

In the upper window of Fig. 2 is the text (*Ronaldo could not save the match*) and the diagnostic question (*Did Ronaldo's team lose the match?*). The lower window contains the answer returned. Let us explain this answer. For SemETAP, to find if a proposition is true means to discover the value of the epistemic modality of this proposition. The meaning of the question is: what is the value of the attribute ?degreeAttribute_2_12, which is the value of the epistemic modality of the statement "the team for which Ronaldo was playing was defeated"? In plain words, it means: is it true that Ronaldo'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 is answered in the affirmative.

VIII. Conclusions

The SemETAP semantic analyzer is 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, the Repository of Individuals and inference rules. 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. Satisfactory results obtained in the experiments ([1, 2, 3]) prove that a general scope semantic analyzer can solve specific

problems, provided it is supplied with good-quality knowledge. Explicit knowledge based on the concepts meaning and common sense knowledge plays a key role.

ACKNOWLEDGEMENTS

The work has been supported by the Ministry of Science and Higher Education of the Russian Federation within the Agreement No 075-15-2020-793.

REFERENCES

[1] Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S. Semantic analysis with inference: high spots of the football match. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”, 2018, Moscow, May 30—June 2.

[2] Boguslavsky I. M., Frolova T. I., Iomdin L. L., Lazursky A. V., Rygaev I. P., Timoshenko S. P. Knowledge-based approach to Winograd Schema Challenge. Computational linguistics and intellectual technologies. Papers from the Annual International Conf. “Dialogue”, 2019, 18(25): 86–103.

[3] Boguslavsky I. M., Dikonov V. G., Frolova T. I., Iomdin L. L., Lazursky A. V., Rygaev I. P., Timoshenko S. P. Full-fl edged semantic analysis as a tool for resolving Triangle-COPA social scenarios. Computational linguistics and intellectual technologies. Papers from the Annual International Conf. “Dialogue”, 2020, 19(26): 106–118.

[4] Lei Shi and R. Mihalcea. “Open Text Semantic Parsing Using FrameNet and WordNet”. Proceedings HLT-NAACL--Demonstrations '04 Demonstration Papers at HLT-NAACL 2004. p. 19-22

[5] B. Coppola and A. Moschitti. “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), 2010.

[6] Z. Azmeh, Jean-Rémy Falleri, Marianne Huchard, Chouki Tibermacine. “Automatic Web Service Tagging Using Machine Learning and WordNet Synsets”. Web Information Systems and Technologies. Lecture Notes in Business Information Processing. Vol. 75, 2011, pp. 46-59.

[7] L. Banarescu, C. Bonial, Shu Cai, M. Georgescu et al. “Abstract meaning representation for sembanking”. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria, 2013, pages 178–186.

[8] L. Banarescu, C. Bonial, Shu Cai, M. Georgescu et al. Abstract Meaning Representation (AMR) 1.2.6 Specification. 2019. https://github.com/amrisi/amr-guidelines/blob/master/amr.md#special-frames-for-roles

[9] J. Bos. “Wide-Coverage Semantic Analysis with Boxer”. In: Semantics in Text Processing. STEP 2008 Conference Proceedings. W08-2222.

[10] J. Bos. “A Survey of Computational Semantics: Representation, Inference and Knowledge in Wide-Coverage Text Understanding”. Language and Linguistics Compass, Vol. 5, Issue 6, 2011, pp. 336–366.

[11] Copestake, A., D. Flickinger, C. Pollard, and I. Sag. “Minimal recursion semantics: An introduction”. Research on Language and Computation 3 (4), 2006, 281–332.

[12] J. Allen, M. Swift, W. de Beaumont. “Deep Semantic Analysis of Text”. In: Symposium on Semantics in Systems for Text Processing (STEP), volume 2008.

[13] Nirenburg, S., and Raskin, V. “Ontological Semantics”. The MIT Press. Cambridge, Massachusetts. London, England. 2004.

[14] S. Nirenburg and M. McShane. “A knowledge representation language for natural language processing, simulation and reasoning”. International Journal of Semantic Computing. Vol. 6, No. 1 (2012) 1-21.

[15] Ruifang Ge, R. Mooney. “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] Poon, H., and Domingos, P. “Unsupervised semantic parsing”. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 EMNLP 09.

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

[18] I. Titov, A. Klementiev. “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. 2011. pp. 1445-1455.

[19] P. Liang, M. Jordan, D. Klein. “Learning Dependency- Based Compositional Semantics”. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - v. 1, 2011, p. 590599.

[20] Moldovan, D., Tatu, M., Clark, Ch. “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, 2010, pp. 373-420.

[21] Rygaev I. Etalog - a natural-looking knowledge representation formalism // Proceedings of ITaS 2018 School and Conference, 2018, (http://itas2018.iitp.ru/media/papers/1570472169.pdf).

[22] Karttunen L. “Implicative verbs”. Language, 1971, 47(2): 340–358.