Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2019”

Moscow, May 29—June 1, 2019

KNOWLEDGE-BASED APPROACH TO WINOGRAD SCHEMA CHALLENGE

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

We propose a method to resolve anaphoric pronouns in the framework of Winograd Schema Challenge (WSC) by means of SemETAP—a knowledge-based semantic analyzer. WSC is a modern version of the famous Turing test. Its objective is to check a machine’s ability to exhibit intelligent behavior indistinguishable from that of a human. In contrast to other approaches to WSC, which are based on machine learning, our method uses explicit knowledge. An important advantage of this approach is that it gives an opportunity to provide an explanation of the result understandable for humans. SemETAP interprets the text using both linguistic and extralinguistic (background) knowledge. The former is stored in the grammar and the dictionary of the ETAP-4 system, and the latter is provided by the SemETAP ontology, inference rules and the repository of individuals. We show how this knowledge is used for resolving WSC. At the moment, the performance of the algorithm is not high—54%. This is due to the incompleteness of the background knowledge supplied to the system. It is shown, however, that if the background knowledge is complete and accurate enough, the WSC test is resolved well and it is easily understandable why the system arrived at a particular conclusion.

Предлагается метод разрешения анафоры в рамках теста Winograd Schema Challenge (WSC) с помощью семантического анализатора SemETAP, основанного на знаниях. Тест WSC представляет собой современный вариант теста Тьюринга и предназначен для проверки того, в какой степени компьютер владеет фоновыми знаниями и некоторыми мыслительными операциями, свойственными человеку. В отличие от других подходов к WSC, использующих машинное обучение, наш

метод основан на эксплицитных знаниях. Важное преимущество такого подхода состоит в том, что он позволяет дать обоснование полученного результата, понятное человеку. Для интерпретации текста SemETAP использует как лингвистические, так и внелингвистические (фоновые) знания. Лингвистические знания собраны в словарях и грамматике системы ETAP-4, а фоновые знания — в онтологии, массиве правил вывода и в базе индивидов. Мы показываем, какие знания и как используются для WSC-теста. Проведенная оценка алгоритма показала невысокий результат — 54%. Это объясняется недостаточно полными фоновыми знаниями, вложенными в систему. Тем не менее, показано, что, если фоновые знания системы достаточно детальны, WSC-тест дает хороший результат, обоснование которого легко понимается человеком.

Keywords: Winograd Schema Challenge, knowledge-based approach, knowledge representation, inference, Etalog language, anaphora resolution

1. Introduction

In this paper, we propose a knowledge-based method to tackle the problem known as Winograd Schema Challenge (WSC). This test is a modern version of the famous Turing test proposed in 1950 and since then playing an important role in the philosophy of artificial intelligence. Turing test is intended to check a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. It consists in maintaining free conversation between a human and a computer through a text-only communication channel. The computer is considered as having passed the test if after a 5-minute conversation the human cannot reliably determine with whom he has conversed—with another human or with a computer. This test was strongly (and fairly) criticized for the central role of deception and trickery incorporated into the system. To pass the test, it was sufficient for the computer to fool the human into thinking he is dealing with another human by means of various tricks, puns, jokes, clever asides, emotional outbursts and the like. The weakness of the Turing test became especially obvious after it was successfully passed in 2014 by the chatbot Eugene Goostman who assumed a false identity of a 13-year old boy from Odessa. The ability to mislead the interlocutor, especially in the course of a short conversation, can hardly qualify as the most natural manifestation of computer intellect.

A better test was proposed in 2011 by Hector Levesque [Levesque 2011], cf. also [Levesque et al. 2011]. As opposed to the Turing test, the WSC test by Levesque requires an unambiguous answer to a series of questions that represent no difficulty for humans but need elementary background knowledge (or, in other words, that what is called "naïve picture of the world") and commonsense reasoning. These questions require to select the correct antecedent of an anaphoric pronoun in a sentence of a particular structure. Let us give two examples that rely on different types of knowledge.

(1) *The trophy does not fit into the brown suitcase because it's too [small/large].* What is too [small/large]?

(2) *Joan made sure to thank Susan for all the help she had [given/received].* Who had [given/received] help?

The questions are compiled in such a way that it is sufficient to replace one word with the other and the correct answer will be different. "With a very high probability, anything that answers correctly a series of these questions… is thinking in the full-bodied sense we usually reserve for people" [Levesque 2011]. However simple the test may seem, it is impossible to pass it if the computer does not possess basic knowledge and reasoning capacities that any normal adult should possess. Hence, being able to make progress on this task enables us to move one step closer to building a machine that can truly understand natural language.

The rest of the paper is organized as follows. We review related work in **Section 2**. Most of the attempts to tackle WSC we are aware of are based on machine learning and statistics. In contrast, we are trying to advance within the knowledge-based framework. In **Section 3** we describe our approach, in **Section 4** we present semantic resources we are using. The key concept used for resolving WSC is the concept of semantic consistency, which is explained in **Section 5**. We report our experiments and analysis in **Section 6** and conclude in **Section 7**.

2. Related work

After the WSC was proposed in [Levesque 2011], it generated a lot of interest and in 2016 a first international WSC competition was organized within International Joint Conference on Artificial Intelligence in New York. The participants were offered a series of sentences similar to (1) and (2) above and related to different aspects of background knowledge. The prize could not be awarded to anybody. Most of the participants showed a result close to the random choice or even worse. The second competition scheduled for 2018 was canceled due to the lack of prospective participants. Thus, WSC sets a very high bar, which the current state-of-the-art can hardly overcome.

As of today, there have been several attempts to resolve WSC. All the authors recognize the necessity to take background knowledge into account but do it in very different ways. Most of the papers recur to some variant of machine learning including deep learning. Most attempts on solving WSC involve heavy utilization of annotated knowledge bases, rule-based reasoning, or hand-crafted features.

[Rahman and Ng 2012] employs the largest set of features derived from a variety of sources: narrative chains [Chambers and Jurafsky 2008], Google API, FrameNet information, heuristic polarity, machine-learned polarity, connective-based relations, semantic compatibility and lexical features.

[Haoruo Peng et al. 2015] develop the notion of Predicate Schemas, instantiate them with automatically acquired knowledge, and compile it into constraints that are used to resolve coreference. Specifically, two types of Predicate Schemas are introduced that cover a large fraction of the challenging cases. The first one specifies one predicate with its subject and object, thus providing information on the subject and object preferences of a given predicate. Example:

(3) *[The bee]*$_{e1}$ *landed on [the flower]*$_{e2}$ *because [it]*$_{pro}$ *had pollen.*
(The flower had pollen) IS MORE PROBABLE THAN (The bee had pollen).

The second type specifies two predicates with a semantically shared argument (either subject or object), thus specifying role preferences of one predicate, among roles of the other. Example:

(4) *[Bill]*$_{e1}$ *was robbed by [John]*$_{e2}$*, so the officer arrested [him]*$_{pro}$.
(Bill was robbed by John, The officer arrested John) IS MORE PROBABLE THAN (Bill was robbed by John, The officer arrested Bill).

These schemas are instantiated by acquiring statistics in an unsupervised way from multiple resources including the Gigaword corpus, Wikipedia, Web Queries and polarity information.

Trinh and Le 2018 propose a method for commonsense reasoning with neural networks, using unsupervised learning. No annotated knowledge bases or hand-engineered features are used. Large RNN language models are built that operate at word or character level. They are trained on unlabeled data taken in a number of massive and diverse text corpora, such as LM-1-Billion, CommonCrawl, SQuAD, Gutenberg Books. The authors claim that the diversity of training data plays an important role in test performance. The system successfully discovers important features of the context that decide the correct answer, indicating a good grasp of commonsense knowledge.

In contrast to this, the system described in [Quan Liu et al 2016], that obtained the highest score at the 2016 challenge, includes both supervised and unsupervised models. It makes use of the skip-gram model to learn word representations. The model incorporates several knowledge bases to regularize its training process, resulting in Knowledge Enhanced Embeddings (KEE). A semantic similarity scorer and a deep neural network classifier are then combined on top of KEE to predict the answers. The commonsense knowledge used is constituted by cause-effect relationship pairs automatically extracted from a large corpus. The pairs are composed of a verb and an adjective and belong to 4 types: “active V—positive A” (*win—happy*), “active V—negative A” (*rob—be arrested*), “passive V—positive A” (*be confident—not afraid*), and “passive V—negative A” (*be restricted—unable*). To combine context and commonsense knowledge for solving the WSC, the paper proposes to treat the commonsense knowledge as semantic constraints and learn KEE based on the generated constraints.

There are several papers that do not use machine learning but apply explicit knowledge representation and reasoning. The focus of the approach proposed in Schüller 2014 is on knowledge representation. To represent both the meaning of the text and the background knowledge, Roger Schank’s graph framework is used. Inference is based on pragmatic effects described in Relevance Theory.

[Bailey et al. 2015] uses a series of axioms and inference rules. A mathematical framework for reasoning is introduced, based on the notion of correlation between the events. F and G are correlated if message F would cause the hearer to view G as more plausible, and message G would cause the hearer to view F as more plausible. For example, if we learn that “A fits into B”, then it is more plausible that “B is big”. It is supposed that such axioms can be acquired automatically from existing lexical and commonsense knowledge bases, such as WORDNET [Fellbaum 1998], FRAMENET [Baker, Fillmore, and Lowe 1998], VERBNET [Kipper-Schuler 2005], PROPBANK [Palmer,

Gildea, and Kingsbury 2005], CONCEPTNET [Liu and Singh 2004), KNEXT [Schubert 2002], and the OPENCYC project (http://www.opencyc.org/doc).

The approach of [Sharma et al. 2015] is close to ours in the sense that it relies on a semantic parser to represent the meaning of the text and the background knowledge. The parser produces graphs on which reasoning is performed. The concept of background knowledge adopted in the paper is somewhat unconventional. Knowledge is extracted from the web (or any other large text repository) on demand, individually for any processed sentence. The idea here is to extract sentences which contain commonsense knowledge required to answer the question about a given Winograd sentence. For example, for the initial sentence

(5) *The man couldn't lift his son because he was so weak*

and the question

(6) *Who was so weak?*

one needs to acquire the knowledge of the type "if X could not lift Y then X may be weak". This knowledge is looked for by creating string queries from the concepts in the sentence and the question and using them as queries to retrieve sentences from a text repository. For this example the query would be: ".*could not lift.*because.*weak.*". It is supplemented by another query obtained by substituting the verb for its synonym: ".*could not raise.*because.*weak.*" Practically, the process of knowledge hunting is nothing more than searching for sentences similar to the initial text in the hope to find a sentence in which the ambiguity in question would be resolved. In our example, such a sentence was found:

(7) *She could not lift it because she is a weak girl*,

in which a coreference resolver can determine that the two occurrences of *she* are coreferent. Each given sentence and the corresponding commonsense knowledge sentences are translated into semantic representation graphs, by using the K-Parser system. Semantic representation of the initial sentence (5) is compared with the semantic representation of the sentence (7) found in the corpus, which results in the inference that *he* = *man*. Rules and constructs for reasoning are formulated within the Answer Set Programming framework.

One could doubt that what the system extracts from the text repository is really deep background knowledge and not just a text similar to the initial one. Regardless, this method is closer to commonsense reasoning than many statistical or machine learning methods commonly used in NLP and in particular Natural language understanding (NLU).

3. Our approach

Nowadays, when computational linguistics is dominated by a powerful machine learning mainstream, the frameworks proposed beyond this mainstream have to justify their choice. We assume that computational linguistics is a fundamental branch of science at the intersection of linguistics and artificial intelligence. Its aim is to describe natural language by means of computer modeling. This is a contrast to NLP, which

primarily aims at the development of useful applications. We believe that in many cases the choice of the paradigm is determined by the task the researchers are facing. In many cases, a linguistic model based on knowledge has a higher explanatory power than a model obtained by machine learning, at least given the current state of technology. If we wish to learn which words are closer in meaning and which are farther, distributive semantics will be a good choice. However, if our task needs defining what the adjective *intelligent* exactly means and how it differs from its synonyms *clever* or *smart* (and any complete model of semantics should necessarily include this information), then machine learning will hardly solve our problem. If our target is to build a concrete useful application (e.g. a syntactic parser or a system of machine translation), it is quite probable that machine learning should be a paradigm of choice. If, however, what we want is not just to describe a certain linguistic phenomenon or a fragment of language but also to understand it and for instance to be able to compare it to a similar phenomenon of another language or of the same language but at a different period of time, then a machine learning-based model will not be the optimal one, since it will not be as transparent as a knowledge-based model can be.

Similar considerations can be heard from the computer science camp. Erik Mueller, a well-known specialist in artificial intelligence and commonsense reasoning and one of the key authors of IMB Watson, in his recent book "Transparent Computers: Designing Understandable Intelligent Systems" insists that computers should be more transparent, open and understandable. We should be able to understand why they arrived at a particular conclusion or why they behaved in a certain way. Accordingly, "intelligent systems should be able to reason like people, they should use concepts familiar to people and combine them in ways that make sense to people" [Mueller 2016]. Computers should explain their reasoning so as to help us decide whether to accept or reject the system's advice. It is important for computers to be transparent, because transparency promotes understanding, is educational, makes it easier to fix problems, improves customer satisfaction, and builds trust. Neural networks are like black boxes. You don't know what they are doing and how they come to the conclusion. At the same time, symbolic techniques, such as Cyc, Event Calculus or OntoSem, are close to be transparent. They represent knowledge symbolically, and they reason symbolically, in a way people can understand.

Coming back to modeling commonsense reasoning and, in particular, WSC, we are entering the field where transparency of the solution is especially desirable. Let us look at one of the examples mentioned above. To resolve the pronoun in sentence (2) above, [Haoruo Peng et al. 2015] uses the background information that the agent of robbing is often an object of arrest. Such information can be extracted from large corpora by means of statistical processing. However, even if this information is available, it only tells us what happens more often but does not provide any explanation or hint of why this should be the case. If a system is modeling commonsense reasoning, it is reasonable to expect some motivation of its decisions, which presumably requires availability of explicit knowledge. For example: ARREST is a curtailment of freedom of a person who committed a criminal action or is suspected thereof. On the other hand, ROBBING is a criminal action. This knowledge makes the assumption of coreference between the agent of ROBBING and the object of ARREST very plausible.

This is the way that we are trying to follow. We adopted what is often called *knowledge-based approach*: explicitly representing knowledge in a formal language, and providing procedures to reason with that knowledge. A major obstacle is that the bulk of background knowledge of humans is not formalized. As of now, it is impossible to build a unified model to cover all this knowledge due to its boundlessness. The attempts to automatically extract background knowledge needed for reasoning are worthy of respect but, as far as we know, they did not deliver tangible result so far, if we are speaking about minimally non-trivial knowledge.

We assume that the formalization of this knowledge can be achieved along the line of incrementally building different fragments of the "naïve picture of the world" so that in the long run the whole picture is covered. This is a long way to go, and we are at its beginning.

This scenario does not seem unfeasible to many researchers. Cf. [Levesque et al. 2012]: "WSC allows for incremental progress: we can begin with simple lexical analyses of the words in the sentences, and then progress all the way to applying arbitrary amounts of world knowledge to the task … In addition, the schema can be grouped according to domain. Some examples involve reasoning about knowledge and communication; others involve temporal reasoning or physical reasoning. Researchers can choose to work on examples in a particular domain, and to take a test restricted to that domain". "Hand crafting microtheories" is one of the methods for creating commonsense knowledge bases, along with statistical and corpus-based machine learning techniques and crowd sourcing, which were invited for presentation at the Twelfth International Symposium on Logical Formalization of Commonsense Reasoning (http://commonsensereasoning.org/2015/cfp.html).

As of today, there have been several attempts of formalization of different fragments of human background knowledge (microtheories), beginning with very narrow scenarios (such as breaking an egg and pouring it into a bowl [Morgenstern 2001]) and ending with larger pieces of the "naïve picture of the world", such as emotions, interpersonal relations, naïve psychology, causality, change of state, etc.—[Gordon and Hobbs 2004]; [Gordon and Hobbs 2011]; [Gordon, Hobbs et al. 2011]; [Hobbs and Gordon 2008]; [Hobbs and Gordon 2010]; [Hobbs and Gordon, 2014]; [Hobbs, Sagae et al. 2012]; [Montazeri and Hobbs, 2011], [Montazeri and Hobbs 2012]). However, so far their number is absolutely insufficient for covering texts of general semantics.

We made one more step along this road. Out of a collection of WSC texts[1], we extracted a subcollection which requires knowledge on mental predicates, translated it into Russian and built a semantic description of the corresponding fragment of the human naïve picture of the world. This description is implemented as a set of inference rules (see **Section 4** below). The experiments that we carried out showed that in many cases this description ensures correct resolution of the WSC test, provided that the key elements of the text are covered by the description. It is essential that this formalization is not geared specifically to the WSC test. It is applicable in a wide range of tasks, such as question answering, extraction of implicit knowledge from the text, recognition of textual entailment, machine comprehension, etc.

[1] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.html

4. Inference rules

In our experiments, we used the SemETAP semantic analyzer. In previous publications, we described its different aspects [Boguslavsky et al. 2015], [Boguslavsky 2017], [Rygaev 2017], [Boguslavsky et al. 2018] and will not repeat them here. Let it only be reminded that:

- SemETAP is an option of the ETAP-4 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, Repository of Individuals and the set of inference rules SemRule.
- Inference rules is a crucial component of SemETAP. 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 decomposition of the concept meaning helps produce additional inferences and thus achieve a deeper understanding.
- Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological concepts. 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 (*Ontoelement-1 relation Ontoelement-2*), where *Ontoelement-i* is a variable or a constant denoting a concept or an instance and *relation* is an object or data property of the ontology that holds between *Ontoelement-1* and *Ontoelement-2*.

For the purposes of this paper, EnSemS is the most important representation, since it makes explicit all the inferences that the knowledge available permits to make from the text and the context. The inferences, in their turn, may help select the more appropriate antecedent in the WSC sentence.

The rules that generate EnSemS, which we call inference rules, are mostly written in the Etalog language [Rygaev 2018]. At the time of writing this paper (February 2019), the number of Etalog rules has reached 408. There are two major types of inference rules—rules for concepts and rules for relations. Concept rules mostly decompose the meaning of the concept explicating components that are relevant from the inference perspective. For example, in describing events, special attention is paid to the following aspects:

a) preconditions that need to have taken place; for example, preconditions of the event “Ivan bought a book from Masha” are Masha’s having the book and Ivan’s having money. If Ivan refused to close the door, the precondition is that someone had asked him to do so.
b) objectives of its participants; for example, Ivan’s goal in the event “Ivan asks Masha what time it is it” is to make Masha tell him the time.
c) results of the event, both obligatory and possible; for example, the result of the event “Ivan bought a book from Masha” is Ivan’s having the book. If Ivan lost the book, it results in Ivan’s not having the book any more, the precondition being that he had it before.

d) if the event is complex, what are its subevents; for example, the event "Peter exchanged his book for Ivan's apple" consists of two acts of givings: Peter gave his book to Ivan and Ivan gave his apple to Peter.
e) presuppositions that the event may have; for example, if Ivan does not know that Peter failed the exam, it is still the case that Peter's failure at the exam did take place.
f) assessment of different components of the event from the point of view of its participants or the speaker; for example, the event "Ivan defeated Peter" is beneficial for Ivan and unbeneficial for Peter.

Fig. 1 presents the screenshot of the rule for the concept `Refusing` written in Etalog.

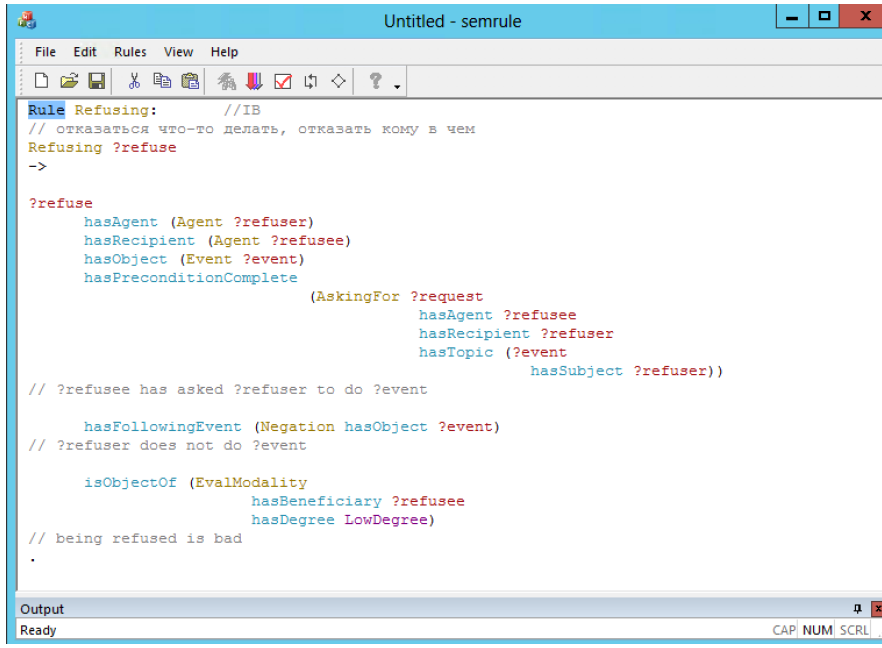

Fig. 1. Inference rule `Refusing`

It says that `Refusing` has an agent (`?refuser`), a recipient (`?refusee`) and an object (`?event`) that the agent refused to do. The precondition of `Refusing` is that `?refusee` has previously asked `?refuser` to do `?event`. The result of `Refusing` is that `?refuser` does not do `?event`. Besides, being refused is bad for `?refusee`.

Among the inference rules that describe relations, noteworthy is the group of rules that define modal and temporal relationships between events. These are: `hasResult`, `hasPossibleResult`, `hasPrecondition`, `hasPossiblePrecondition`, `hasSpeakerCommitment` (introduces factive complements: *Ivan does not know that Peter failed the exam*), `hasPreventedEvent` (introduces the consequence that did not take place: *Ivan forgot to call a taxi*), `hasSubEvent`,

`hasPossibleSubEvent` (introduces a subevent that may, but not necessarily, take place: *chewing* is a possible subevent of *eating*), `hasSubEventFinal` (introduces the final stage of a complex event), `hasSyncEvent` (the event is synchronized in time and modality with another event), `hasPossibleSyncEvent` (the event may be synchronized in time and modality with another event), `hasSyncAntiEvent` (the event excludes another event at the same time: *sleep—be awake*), `simultaneously` (the event is synchronized with another event in time but not in modality).

5. In search of semantic consistency

One of the first approaches to semantic consistency was proposed in [Apresjan 1974/1995:13–15]. It was applied to the task of word sense disambiguation. A sentence interpretation is considered semantically consistent when the repetition of semantic elements is maximal. This idea was illustrated by example (8)

(8) *Xorošij konditer ne žarit xvorost na gazovoj plite* 'a good pastry-cook does not fry pastry straws in a gas-stove'

Some of the words here are ambiguous. The word *xvorost* is ambiguous between 'pastry straws' and 'dry tree branches fallen on the ground'. *Plita* means either 'heating device for cooking food' or 'flat piece of solid material'. *Konditer* may mean 'a person that cooks pastry' or 'an owner of the confectionary'. Obviously, in all the three cases one should select the first of the senses indicated ('pastry straws', 'heating device for cooking food', and 'a person that cooks pastry'), because all these senses contain the semantic element 'cooking/cooked food'.

For WSC, one cannot use this approach directly. Often, competing interpretations differ not so much in the composition of semantic elements as in their organization. The same elements are organized in different predications, that is structures composed of a predicate and its arguments. Therefore, we modified slightly the notion of semantic consistency. We will consider sentence interpretation *Int1* more consistent than *Int2*, if it contains more identical predications. It is essential that we check consistency not on the initial text, nor on its syntactic structure and not even on the Basic Semantic Structure. We are searching the Enhanced Semantic Structure, because it contains the full body of inferred predications, and all of them should be taken into account. In more detail, the algorithm for determining consistency will be presented below. Here, we will illustrate its idea with a concrete example.

Here is one of the classical WSC pairs:

(9) *James asked Robert for a favor but he refused.*

(10) *James asked Robert for a favor but he was refused.*

We will proceed as follows. For each sentence, we will build two variants which differ in the antecedent selection. Then we will produce EnSemS for both and check which of them manifests a higher degree of consistency in the sense defined above. Let us begin with sentence (9).

(9a) *James asked Robert for a favor but Robert refused.*

(9b) *James asked Robert for a favor but James refused.*

Sentences (9a) and (9b) are composed of the same semantic elements, but organized differently. The key component for comparing (9a) and (9b) is *refusing*. As we saw above (Fig. 1), its meaning contains a reference to a precondition. A1's refusal to do A2 presupposes that A3 asked A1 to do A2 before. Hence, (9a) contains the element 'ask for' twice: 1) it makes part of the first part of the sentence ('James asked Robert for a favor') and 2) it is part of the precondition of refusal in the second part of the sentence ('somebody asks Robert for something'). These predications are identical up to unification. This means that they coincide, if the same variables are instantiated by the same expressions (James --> somebody, favor --> something). It's easy to see that in (9b) the corresponding predications do not unify: the addressee of the first occurrence of 'asking for' is Robert, while in the second occurrence it is James. So, (9a) has identical (up to unification) propositions, and (9b) does not have them. Consequently, (9a) is more consistent than (9b) in the sense indicated above. In the same way, one can show that (10a) is less consistent than (10b):

(10a) *James asked Robert for a favor but Robert was refused.*

(10b) *James asked Robert for a favor but James was refused.*

However, this time the difference between the predications in the less consistent sentence is due to different agents of 'asking for' and not addressees: in (10a) it is James in the first request, and Robert in the second one.

5.1. Algorithm for consistency check

The algorithm for determining consistency is based on the idea of similarity between two nodes in the graph. Similarity is a numerical value which can range from −1 (two nodes are absolutely different) to 1 (two nodes are identical).

The algorithm takes a pronoun node and calculates its similarity to each potential antecedent node. The one with a higher similarity is selected. Similarity is calculated in the following way:

1. Similarity of a node to itself is 1.
2. Similarity of two different constant nodes is −1.
3. Similarity of two nodes which have incompatible classes is −1.
4. Similarity of two nodes which have compatible classes is calculated based on their environment in the graph, namely:
    a. For each incoming and outgoing functional relation of the pronoun node find a corresponding relation of the antecedent node and match their values using the same algorithm,
    b. If there is no corresponding relation on the antecedent node then assume the similarity is 0.
    c. Return a total similarity of all the relations divided by the number of the relations.

Here is the (simplified) calculation for the case in (9):

Pronoun node	Antecedent node	Calculation
(Agent isRecipientOf(AskingFor hasAgent(Agent)))	(Human hasGivenName "Robert" isRecipientOf(AskingFor hasAgent(Human hasGivenName "James")))	Classes match. The value of the only relation (isRecipientOf) match with similarity = 1. So the total similarity is 1/1 = 1.
(Agent isRecipientOf(AskingFor hasAgent (Agent)))	(Human hasGivenName "James" isAgentOf(AskingFor hasRecipient(Human hasGivenName "Robert")))	Classes match. There is not match for the only relation (isRecipientOf) hence its value is 0. So the total similarity is 0/1 = 0.

6. Experimental results

We carried out two series of experiments. In the first series, we processed the sentences of the WSC mental subcollection. These sentences were open to us when we were writing inference rules. They served as the testing bed of the model. In most of these sentences all the antecedents were identified correctly. Here are some of these sentences:

Okun' proglotil červja, on byl golodnyj.
The perch swallowed the worm, it was hungry. It = the perch

Okun' proglotil červja, on byl vkusnyj.
The perch swallowed the worm, it was tasty. It = the worm

Petr dal konfetu Ivanu, potomu čto on byl goloden.
Peter gave Ivan a candy, because he was hungry. He = Ivan

Petr dal konfetu Ivanu, potomu čto on byl ne goloden.
Peter gave Ivan a candy, because he was not hungry. He = Peter

Ivan vo vsem podražaet Petru, on ego obožaet.
Ivan imitates Peter in everything, he_1 adores him_2. He_1 = Ivan, he_2 = Peter

Ivan vo vsem podražaet Petru, on sil'no na nego vlijaet.
Ivan imitates Peter in everything, he_1 has a strong influence upon him_2.
He_1 = Peter, he_2 = Ivan

Petr postučal v dver' Ivana, no on ne otvetil.
Peter knocked at Ivan's door, but he did not reply. He = Ivan

Petr postučal v dver' Ivana, no on ne polučil otveta.
Peter knocked at Ivan's door, but he did not receive a reply. He = Peter

Ivan poprosil Petra ob odolženii, no on otkazal.
Ivan asked Peter for a favor, but he refused. He = Peter.

Ivan poprosil Petra ob odolženii, no emu otkazali.
Ivan asked Peter for a favor, but he was refused. He = Ivan.

Ivan oral na Petra, potomu čto on byl rasstroen.
Ivan was shouting at Peter, because he was upset. He = Ivan

Ivan utešal Petra, potomu čto on byl rasstroen.
Ivan was comforting Peter, because he was upset. He = Peter

Ivan obižal Petra, tak čto my ego zaščitili.
Ivan offended Peter so we defended him. He = Peter.

Ivan obižal Petra, tak čto my ego nakazali.
Ivan offended Peter so we punished him. He = Ivan.

We can draw two conclusions from this experiment. First: if the background knowledge provided is detailed and accurate, the WSC test can be passed in most cases, and the explanation of the result is easily understandable by humans. Second: it is time- and effort-consuming to compile all the information needed, including a) entering new ontology concepts, b) linking Russian words with the ontology, and c) formulating inference rules. As is often the case with rule-based systems, the result is rather fragile. The rules should be complete and accurate, in order to achieve the expected result.

The second series of experiments was an evaluation proper. We compiled a test corpus of 20 new sentences belonging to the same mental domain. Many of the mental predicates of the test corpus were not represented in the concept dictionary or in the inference rules. These were added to the concept dictionary and supplied with semantic descriptions before running the test, but—naturally—without having access to the test corpus. The result of the testing was rather low—54%, slightly above the random benchmark (50%).

The evaluation results are summarized in Table 1. The table shows for each test sentence whether an antecedent for a pronoun was identified correctly (1) or not (0). For sentences with two pronouns each antecedent identification result is displayed separately. We compare two approaches: the one based on the syntactic constraints and the semantic one described in this paper.

The syntactic approach for anaphora resolution developed in our labs [Inshakova 2019] gives quite good results (precision and recall around 70% depending on the test corpus). But, as was explained above, purely syntactic approach is not suitable for WSC sentences, which are specifically built in such a way that syntactic constraints are helpless and background knowledge is required. Hence the syntactic approach

on our test WSC sentences gives us 50% of correctly identified antecedents (exactly the chance level). This was expected and can be an indication that the WSC sentences are composed correctly.

On the other hand we expected that the semantic approach would give us noticeably better results. Regretfully, it did not happen. The result of 54% is only marginally better than the syntactic approach and a random selection. In some cases the system could not decide between two antecedents and a random choice had to be applied (those are denoted by 0.5 in the table). In other cases the antecedent was selected incorrectly.

The error analysis shows that in all the cases the failure to identify the antecedent correctly was due to knowledge incompletely provided to the system. This can be partly explained by limited time for the preparation of the evaluation. But, more importantly, it is not clear yet what is the amount of effort needed to produce a description complete enough.

Table 1. Evaluation results

Test sentence	Anaphora resolution approach	
	Syntactic	Semantic
Vrač propisal Petru očki, potomu čto on ploxo vidit. The doctor prescribed Peter glasses, because he has poor eyesight	0	0
Vrač propisal Petru očki, potomu čto on proveril ego zrenie. The doctor prescribed Peter glasses, because he checked his eyesight	1 + 1	1 + 1
Kolja posovetoval Petru otdoxnut', potomu čto on očen' ustal. Kolya advised Peter to have a rest, because he was very tired	1	0.5
Kolja posovetoval Petru otdoxnut', no potom on peredumal. Kolya advised Peter to have a rest, but later on he changed his mind	0	0.5
Petr dal den'gi Ivanu, potomu čto on bogat. Peter gave money to Ivan, because he was rich	1	1
Petr dal den'gi Ivanu, potomu čto on beden. Peter gave money to Ivan, because he was poor	0	0
Petr odolžil den'gi Ivanu, no on ix ne vernul. Peter lent money to Ivan, but he did not give it back	0	0.5
Petr odolžil den'gi Ivanu, potomu čto on xotel pomoč emu. Peter lent money to Ivan, because he wanted to help him	1 + 1	0.5 + 0.5
Petr pobedil Kolju, potomu čto on xorošo igral. Peter defeated Kolya because he played well	1	0.5
Petr pobedil Kolju, potomu čto on ploxo igral. Peter defeated Kolya because he played poorly	0	0.5
Petr pomog Kole s zadaniem, potomu čto on dobryj. Peter helped Kolya with the task, because he is kind	1	0

Test sentence	Anaphora resolution approach	
	Syntactic	Semantic
Petr pomog Kole s zadaniem, potomu čto on poprosil ego pomoč. Peter helped Kolya with the task, because he asked him to help	1 + 0	1 + 1
Petr obvinil Ivana, no ego opravdali. Peter accused Ivan, but he was acquitted	0	0.5
Petr obvinil Ivana, no potom on požalel ob etom. Peter accused Ivan, but later he regretted it	1	1
Petr zaviduet Ivanu, potomu čto on xorošo tantsuet. Peter is envious of Ivan, because he is a good dancer	0	0.5
Petr ne zaviduet Ivanu, xotja on xorošo tantsuet. Peter is not envious of Ivan, although he is a good dancer	0	0
Džon rasserdilsja na Billa, xotja on dobryj. John got angry at Bill, although he is kind	1	0.5
Džon rasserdilsja ne Billa, xotja on ne vinovat. John got angry at Bill, although he was not guilty	0	0.5
Vasja umoljal Ivana ostat'sja doma, no on ne soglasilsja. Vasya begged Ivan to stay at home, but he refused	1	1
Vasja umoljal Ivana ostat'sja doma, no on ne dobilsja uspexa. Vasya begged Ivan to stay home, but he was unsuccessful	0	0
Average	**0.50**	**0.54**

7. Conclusion

This paper proposes a knowledge-based framework for solving Winograd Schema Challenge (WSC). We use the general semantic parser SemETAP that represents the sentence at two semantic levels: Basic Semantic Structure shows the semantics of the isolated sentence, and the Enhanced Semantic Structure supplies it with inferences made on the basis of available knowledge. Background knowledge is implemented by means of inference rules written in the Etalog inference language. SemETAP can be used for a wide range of tasks that require explicit representation of implicit knowledge. Experiments show that if the background knowledge provided is detailed and accurate, the WSC test can be passed in most cases, and the explanation of the result is easily understandable by humans.

8. Acknowledgements

This work was supported by the RSF grant No. 16-18-10422-P, which is gratefully acknowledged.

References

1. *Apresjan Yu. D.* (1974), Leksicheskaya semantika. Sinonimicheskie sredstva yazyka [Lexical semantics. Synonimic means of language]. Moscow: Nauka; Second edition: Selected works: In 2 vol. Vol. I. Moscow: Shkola «Yazyki Russkoi Kul'tury», 1995.
2. *Bailey D., A. Harrison, Yu. Lierler, V. Lifschitz, and J. Michael* (2015), The Winograd schema challenge and reasoning about correlation. In: Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning.
3. *Baker, C. F.; Fillmore, C. J.; and Lowe, J. B.* (1998), The Berkeley FrameNet Project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1 (ACL/COLING), 86–90.
4. *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.
5. *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.
6. *Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S.* (2018), 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.
7. *Chambers N., D. Jurafsky* (2008), Unsupervised learning of narrative event chains. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies , pages 787–797.
8. *Fellbaum, C., ed.* (1998), WordNet: An Electronic Lexical Database . MIT Press.
9. *Gordon, A. and Hobbs, J.* (2004), Formalizations of Commonsense Psychology. AI Magazine 25(4):49–62.
10. *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.
11. *Haoruo Peng, Daniel Khashabi, and Dan Roth.* (2015), Solving hard coreference problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 809–819.
12. *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.
13. *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.
14. *Hobbs, J., and Gordon, A.* (2014), Axiomatizing Complex Concepts from Fundamentals (invited paper). Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014), April 6–12, 2014, Kathmandu, Nepal.

15. *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.
16. *Inshakova, E.* (2019), An anaphora resolution system for Russian based on ETAP-4 linguistic processor (this volume).
17. *Kipper-Schuler, K.* (2005), VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon . Ph.D. Dissertation, University of Pennsylvania.
18. *Levesque H.* (2011), The Winograd Schema Challenge. In: AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning.
19. *Levesque, H., Davis, E., Morgenstern, L.* (2011), "The Winograd Schema Challenge", In: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning.
20. *Liu, H., and Singh, P.* (2004), Conceptnet: A practical commonsense reasoning toolkit. BT Technology Journal 22:211–226.
21. *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.
22. *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.
23. *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
24. *Schubert, L.* (2002), Can we derive general world knowledge from texts? In: Proceedings of the Second International Conference on Human Language Technology Research (HLT), 94–97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
25. *Mueller E.* (2016), Transparent Computers: Designing Understandable Intelligent Systems. Createspace Independent Publishers.
26. *Palmer, M.; Gildea, D.; and Kingsbury, P.* (2005), The proposition bank: An annotated corpus of semantic roles. Computational Linguistics 31(1):71–106.
27. *Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, Yu Hu.* (2016), Combing Context and Commonsense Knowledge Through Neural Networks for SolvingWinograd Schema Problems. arXiv:1611.04146v1 [cs.AI] 13 Nov 2016.
28. *Rahman A., V. Ng.* (2012), Resolving complex cases of definite pronouns: the Winograd schema challenge. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 777–789. Association for Computational Linguistics.
29. *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).

30. *Rygaev I.* (2018), Etalog—a natural-looking knowledge representation formalism // Proceedings of ITaS 2018 School and Conference (http://itas2018.iitp.ru/media/papers/1570472169.pdf).
31. *Schüller P.* (2014), Tackling Winograd schemas by formalizing relevance theory in knowledge graphs. In: Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning.
32. *Sharma A., Nguyen Ha Vo, Somak Aditya, and Chitta Baral.* (2015), Towards addressing the Winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In IJCAI, pages 1319–1325.
33. *Trieu H. Trinh, Quoc V. Le.* (2018), A Simple Method for Commonsense Reasoning. arXiv:1806.02847v1 [cs.AI] 7 Jun 2018.