computing@computingonline.net
www.computingonline.net

Print ISSN 1727-6209
On-line ISSN 2312-5381
International Journal of Computing

USING MULTIPLE SEMANTIC MEASURES FOR COREFERENCE RESOLUTION IN ONTOLOGY POPULATION

Natalia Garanina [1], Elena Sidorova [1], Irina Kononenko [1], Sergei Gorlatch [2]

[1] A. P. Ershov Institute of Informatics Systems, Siberian Branch of Russian Academy of Science,
Lavrent'ev av., 6, Novosibirsk 630090, Russia,
garanina@iis.nsk.su, lsidorova@iis.nsk.su, irina_k@cn.ru
[2] University of Muenster, Germany, gorlatch@uni-muenster.de

Abstract: The problem of populating an ontology consists in adding to it some new, domain-specific content from an input expressed, in particular, in a natural language. We focus on an important aspect in the ontology population process – finding and resolving *coreferences*, i.e., similar mentions of entities in the input text. Our contribution is a novel formal framework that extends the state-of-the-art approaches to coreference resolution by using multiple semantic similarity properties in the resolution process, i.e., we extend the list of the ontological properties used for coreference resolution with additional properties such as inverse, symmetry, intersection, union, etc. We use the proposed framework to improve our previously proposed algorithm for coreference resolution used in our general approach to text analysis and information extraction for populating subject domain ontologies. We describe a multi-agent implementation of our information extraction system and we show that using additional semantic similarity measures for evaluating coreferential candidates improves the quality of the coreference resolution process, especially for complex objects whose coreferencing has not been yet studied in detail. *Copyright © Research Institute for Intelligent Computer Systems, 2017. All rights reserved.*

Keywords: coreference resolution, ontology properties, semantic measure, ontology population, semantic text analysis.

1. INTRODUCTION

The process of ontology population is the actively studied problem of adding new instances of concepts to the ontology. This process is a part of ontology acquisition [18] from a domain-specific content that is most often represented in a natural language. In this context, the solution for the ontology population task is interrelated with the elaboration of natural language processing (NLP) techniques applied in the process of information extraction (IE), with coreference resolution as one of the most challenging NLP tasks.

In linguistics, a reference is a relation of a text expression with some non-linguistic object or circumstances in the real or abstract world. The *coreference resolution* problem is to identify a particular text mention of a non-linguistic entity to its other mentions in this text. Traditionally, the process of coreference resolution consists of two main tasks: 1) the detection of entity mentions that are candidates for coreference, and 2) the pairwise comparison of candidate mentions in order to make the decision on candidate admissibility (whether the pair is valid or not) using some criteria.

The contribution of this paper is a formal framework for the broad use of properties of ontology classes and relations in the coreference resolution process. We exploit these properties for evaluating the semantic coreference similarity in the integral evaluation of coreference similarity. We use the proposed framework to improve our coreference resolution algorithm suggested in [6] for making the decision on the candidate admissibility, which is used in our general approach to text analysis and information extraction for populating subject domain ontology. In our approach, the following IE tasks are performed: the preliminary extraction of subject domain terms from a given text [14]; the segmentation of the text into formal and genre fragments (sentences, sections, headlines, etc) [22]; the construction of objects corresponding to instances of a subject domain ontology, from the terms [4] and the coreference resolution [6]; the lexical and syntactic disambiguation [5]; the update of the ontology with the processed objects (planned as future work). In our framework, the coreference resolution problem means detecting if some group of retrieved objects refers to the particular ontology instance.

There are several basic approaches to coreference resolution proposed in the literature. The most important trends in the field can be found in the comprehensive surveys [3, 16, 17, 19]. These trends can be categorized into rule-based and machine learning approaches. Early coreference resolution systems (dating back to 1970s and 1980s) are called “rule-based” as they rely on hand-coded heuristics that specify whether two expressions can or cannot corefer [1, 2, 11, 24]. The better term for this trend is “linguistic approach” [3] as it incorporates a lot of domain and linguistic knowledge: syntactic constraints, semantic features and preferences, and discourse-oriented theories, such as Centering model [7], which can predict the focus of attention and the choice of a referring expression for a sentence. Theoretical models consider integrated knowledge sources and reveal factors that help to remove unlikely candidates until the minimal set of plausible candidates is obtained, and then make use of the center or focus, or other preferences. Modern theories investigating multiplicity of factors involved in the coreference phenomenon (such as the notion of Referent activation based on a discourse structure, antecedent syntactic or semantic role, animacy, etc. proposed in [12]) were used directly or indirectly in [13]. The mid-to-late 90's gave rise to “corpus-based” (a.k.a. machine learning) approaches which were inspired by the emergence of more powerful automatic parsers and taggers, and corpora annotated with coreference information to be used as a training data [8, 23]. Paper [3] gives a survey of machine learning based techniques with respect to the coreference resolution task starting from a simple statistical naive Bayes-based model to methods using decision trees and conditional random fields and others. Unfortunately, in limited subject domains (for example, technical documents) representative training text corpora do not exist usually. In these cases, it is reasonable to use classical rule-based methods.

In the context of ontology population, the rule-based approach called “ontology-driven IE” is of particular significance. In this approach, IE and ontology population are closely interrelated: an ontology is used to represent the IE process output while the ontology structure and knowledge represented in it help to solve IE domain-specific subtasks [15]. In [10, 25] the coreference resolution task is discussed with respect to both intra- and cross-document analysis. In both papers the ontology-level information is used to determine ontology object identity and similarity: they can be calculated using the object's own features' values and the values of features of other objects that are connected with this object by semantic relations. The approach to coreference resolution in [10] allows only certain types of named entities (persons, organizations, etc.), and the feature values comparison is made by direct string matching without use of any similarity measure. To avoid identification errors, they use a special hand-crafted database that contains validated objects with no duplicates: the identifiers (feature values) of the extracted objects are compared with the identifiers of objects in the base. In [25], the process consists of two consecutive steps. The first step deals with the coreference factors at the text level (such as string similarity) and produces typed entity and relation instances that are mapped into an RDF graph. Afterwards a semantic coreference algorithm runs on the RDF graph to revise the results of the text-based step: instances are merged if they belong to the same class in the domain ontology and their string similarity is higher than a predefined threshold. However, these approaches to coreference resolution provide insufficient completeness, in particular, due to the poor use of the features of ontology classes and relations. They take into account coincidence of classes and relations of coreferential candidates for the resolution, i.e. they use only the identity property of ontology elements.

There exist several attempts to apply distributed or agent-based techniques to the coreference resolution task, in particular [20, 26]. In [20] the coreference resolution factors (recency, number agreement, gender agreement etc.) are grouped in sets as constraint sources corresponding to the known partial theories of coreference. In [26] a common constraint agent allows for morphological agreement and semantic consistency, while different coreference types (where a candidate may be a name alias, nominal predicate, appositional, definite, demonstrative, or bare noun phrase) are charged to special agents. In both papers, agents correspond to the coreference resolution factors. The detection of coreference candidates is done sequentially. In [20] the agents make the decision about admissibility of a particular candidate in parallel, and in [26] the agents compose the system of sequential decision filters. Unfortunately, due to a low degree of concurrency, the performance of the coreference resolution in these agent systems is close to the performance of the sequential resolution process.

Our approach to coreference resolution [6] is rule-based, because we deal with limited subject domains. Our proposed algorithm is ontology-driven as it strongly relies on the structure of the underlying predefined domain ontology. We focus on full lexical items (nominals and names), since they bear more semantic clues than pronominals for making comparisons with ontology classes and instances. Ambiguities occurring at the linguistic level are resolved at the ontology level. We use a similarity measure to compare potential coreferential objects within the group. The detection and resolution of the coreference use the ontology properties of classes and the similarity measure. Unlike the previous ontology-driven approaches, our evaluation of the measure is not limited to string similarity and the identity property of ontology elements: the notion of similarity integrates textual factors (such as text distance and context dependence) with the factors based on the ontological properties of instances' attributes (class hierarchy, composition, transitivity, etc.). We use the special class agents to detect and resolve the coreference candidates using the similarity measure. Our agents work in parallel, which speeds up the process in comparison with the sequential and multi-agent approaches mentioned above.

In this paper, we suggest to extend the list of the ontological properties used for coreference resolution with additional properties such as inverse, symmetry, intersection, union, etc. Using these extra properties for evaluating coreference similarity improves the quality of the resolution process. Such evaluation method can be applied to any ontology-driven approach. Our way of using the ontology structure allows one to resolve coreferences more precisely even for complex objects such as descriptions of events and situations presented as ontology polyadic relations. To the best of our knowledge, coreferencing such complex objects has not been studied in detail yet.

The rest of the paper is organized as follows. In Section 2, we give some background definitions and formally state the problem of coreference resolution. Section 3 defines the semantic similarity measure in detail and gives some examples of its evaluation. Section 4 outlines our approach to multi-agent information extraction, gives the description of the process of the coreference resolution and presents the algorithm of computing the combined semantic similarity measure. In the concluding Section 5, we discuss future work.

2. BASIC DEFINITIONS

Let us consider an ontology of some particular subject domain, together with the ontology population rules, semantic and syntactic models for the language of the subject domain, and the term vocabulary. We assume that input data are provided as a finite natural language text, information from which is used for populating our ontology. We consider an OWL-like ontology representation [9]. *An ontology O of a subject domain* includes the following elements:

- a finite nonempty set C_O of *classes* for representing the concepts of the subject domain,
- a finite set D_O of *data domains*, and
- a finite set of *attributes* with names in $Atr_O = Dat_O \cup Rel_O$, each of which has values in some data domain from D_O (*data attributes* in Dat_O) or has values as instances of some classes (*relation attributes* in Rel_O, which model binary relations).

Every class $c \in C_O$ is defined by the tuple of attributes: $c = (Dat_c, Rel_c)$, where every data attribute $\alpha \in Dat_c \subseteq Dat_O$ has the domain $d_\alpha \in D_O$ with values in $V_{d\alpha}$ and every relation attribute $\rho \in Rel_c \subseteq Rel_O$ has values from classes $C_\rho \subseteq C_O$. We denote the class of an attribute γ by c^γ. The set of all class attributes is denoted by $Atr_c = Dat_c \cup Rel_c$. This set includes the nonempty set of *key attributes* $Atr^K{}_c$. The key attributes can be data or relation attributes. We say that a is *an instance of the class* $c_a = (Dat_{ca}, Rel_{ca})$ $(a \in c_a)$ iff $a = (c_a, Dat_a, Rel_a)$, where every data attribute in Dat_a has a name $\alpha_a \in Dat_{ca}$ with the values $V_{\alpha a}$ from $V_{d\alpha a}$ and every relation attribute in Rel_a has a name $\rho_a \in Rel_{ca}$ with the values $V_{\rho a}$ as instances of the classes from C_ρ. The data key attributes are always one-valued, i.e. every key attribute of every ontology instance may have only a single value. The relation key attributes correspond to bijective relations. We consider an ontology without data and class synonyms, i.e. $\forall\, \alpha_1, \alpha_2 \in Dat_O$: $d_{\alpha 1} \neq d_{\alpha 2}$ and $\forall\, c_1, c_2 \in C_O : Atr_{c1} \neq Atr_{c2}$. *The information content* IC_O of the ontology O is a set of instances of the classes from O. *The ontology population problem* is to compute an information content for a given ontology from the given input data.

In the following, we list some properties of classes and attributes which are well-known in the area of ontology and description logics. We will use them in the process of detection and resolution of coreferences. This list does not claim to be comprehensive. The use of these properties for evaluating the semantic coreferential similarity improves the precision and recall of coreference resolution. We can evaluate the degree of identity/similarity of coreferential candidates using the fact that the data/relation attributes of these coreferential candidates are related by some of these relations and their values are consistent. In this paper, combinations of the properties are not considered, except the refinement relation which is the combination of the composition and inclusion relations. We use the standard notions of class and attribute inheritance relations. The relations on relation attributes correspond to the standard definitions of ontology relations between classes.

Definition 1. Let $c, c' \in C_O$, $\gamma, \gamma' \in Atr_O$, and $\rho, \rho', \rho'' \in Rel_O$. We define the following properties:

- the single *inheritance class relation*: $c < c'$;

- the single *inheritance attribute relation*: $\gamma \ll \gamma'$;
- the ternary *intersection relation*: $\rho = \rho' \sqcap \rho''$;
- the ternary *union relation*: $\rho = \rho' \sqcup \rho''$;
- the ternary *composition relation*: $\rho = \rho' \circ \rho''$;
- the ternary *refinement relation*: $\rho = \rho' \rhd \rho''$ iff $\rho' \circ \rho'' \sqsubset \rho$;
- *the inverse relation*: $\rho = \rho'^{\smile}$;
- *the inclusion relation*: $\rho \sqsubseteq \rho'$;
- *the transitive-reflexive closure relation*: $\rho = \rho'^*$;
- *the transitivity*: $\rho \in Rel^t{}_O$;
- *the symmetry*: $\rho \in Rel^s{}_O$.

We extend the standard list of properties with the refinement relation as the combination of the composition and inclusion relation, because in many practical cases of ontology relations the strict inclusion of the relation composition is required in coreferential candidates' comparison. For example, using the attribute *relation live_in$\circ$ include* $\sqsubset$ *appear_in* we can deduce that if somebody lives in a house then the one can appear in a room of the house, but the opposite assertion does not hold, i.e. in some sense, attribute *include* refines *live_in*.

For the specific goal of this paper – evaluating the semantic coreference similarity – we introduce the following new notions. For classes and attributes, we take into account the hierarchical structure implied by the inheritance relation. Let $\gamma, \gamma' \in Atr_O$, $c, c' \in C_O$, $C, C' \subseteq C_O$, then:

- *The hierarchical group of the class c* is $Hi(c) = \{c\} \cup \{ c' \mid c' < c \vee c' > c \}$.
- *The hierarchical group of the set C* is $Hi(C) = \cup_{c' \in C} Hi(c')$.
- *Hierarchical membership*: $c \in^i C$ iff $c \in Hi(C)$.
- *Hierarchical inclusion*: $C \subseteq^i C'$ iff $\forall c \in C : c \in^i C'$.
- *Hierarchical intersection*: $C \cap^i C' = Hi(C) \cap Hi(C')$.
- *Hierarchical consistency* $\simeq^i$:
    - $c \simeq^i c'$ iff $Hi(c) \cap Hi(c') \neq \emptyset$;
    - $C \simeq^i C'$ iff $C \cap^i C' \neq \emptyset$;
    - $\gamma \simeq^i \gamma'$ iff $\gamma = \gamma' \vee \gamma \ll \gamma' \vee \gamma \gg \gamma'$.

For cases when properties of attributes in Definition 1 are unknown for a given ontology to be populated, we use the necessary conditions of the properties for evaluating the semantic coreferential similarity. The following proposition formulates these conditions in a constructive way. We denote the necessary condition of a property x by $\mathcal{N}^x$. The proof follows from Definition 1.

Proposition 1. Let $\alpha, \beta \in Dat_O, \rho, \xi, \pi \in Rel_O$.

- $\alpha \simeq^i \beta \Rightarrow \mathcal{N}^d = (V_{d\alpha} \subseteq V_{d\beta} \vee V_{d\beta} \subseteq V_{d\alpha})$;
- $\rho \simeq^i \xi \Rightarrow \mathcal{N}^r = (C_\rho \subseteq^i C_\xi \vee C_\xi \subseteq^i C_\rho)$;
- $\rho \in Rel^t{}_O \Rightarrow \mathcal{N}^t = (c^\rho \in^i C_\rho)$;
- $\rho \in Rel^s{}_O \Rightarrow \mathcal{N}^s = (c^\rho \in^i C_\rho)$.
- $\rho = \pi^{\smile} \Rightarrow \mathcal{N}^{\smile} = (c^\rho \in^i C_\pi \wedge c^\pi \in^i C_\rho)$;
- $\pi = \rho \sqcap \xi \Rightarrow \mathcal{N}^{\sqcap} = (c^\pi \in^i \{c^\rho\} \wedge c^\pi \in^i \{c^\xi\} \wedge C_\pi \subseteq^i C_\rho \cap^i C_\xi)$;
- $\xi = \rho \sqcup \pi \Rightarrow \mathcal{N}^{\sqcup} = (c^\rho \in^i \{c^\xi\} \wedge C_\rho \subseteq^i C_\xi)$;
- $\rho \sqsubseteq \xi \Rightarrow \mathcal{N}^{\sqsubseteq} = (c^\rho \in^i \{c^\xi\} \wedge C_\rho \subseteq^i C_\xi)$;
- $\rho = \xi^* \Rightarrow \mathcal{N}^* = (c^\rho = c^\xi \wedge C_\rho \subseteq^i C_\xi \wedge c^\xi \in^i C_\xi)$;
- $\rho = \xi \rhd \pi \Rightarrow \mathcal{N}^{\rhd} = (c^\xi \in^i \{c^\rho\} \wedge c^\pi \in^i C_\xi \wedge C_\pi \subseteq^i C_\rho)$;
- $\rho = \xi \circ \pi \Rightarrow \mathcal{N}^{\circ} = (c^\xi = c^\rho \wedge c^\pi \in^i C_\xi \wedge C_\pi = C_\rho)$.

We define a set A of *information objects (i-objects)* retrieved from input data and corresponding to ontology instances. Every information object $a \in A$ has the form $(c_a, Dat_a, Rel_a, G_a, P_a)$, where

- the class $c_a \in C_O$;
- Dat_a is the set of data attributes $\alpha_a = (\alpha, Val_{\alpha a})$, where
    - the name $\alpha \in Dat_{ca}$, and
    - $Val_{\alpha a}$ is the set of information values $\boldsymbol{v} = (v_v, s_v)$ with
        - the data value $v_v \in d_\alpha$, a set of values of α_a is $V_{\alpha a} = \{ v_v \mid \boldsymbol{v} \in Val_{\alpha a}\}$,
        - s_v is structural information (a position in input data);
- Rel_a is the set of relation attributes $\rho_a = (\rho, V_{\rho a})$, where
    - the name $\rho \in Rel_{ca}$, and $V_{\rho a}$ is the set of i-objects of a class $c_{\rho a}$ from $C_{\rho a}$;
- G_a is the grammar information (morphological and syntactic features);
- P_a is the structural information (a set of positions in the input data).

We denote by $Atr_a = Dat_a \cup Rel_a$ the set of all attributes. Note that the properties of natural language processing may cause assigning key attributes of i-objects with many values. Such ambiguities are resolved after the coreference resolution process is finished.

Every i-object corresponds to some ontology instance in a natural way as follows. Let $a = (c_a, Dat_a, Rel_a, G_a, P_a)$ be an i-object, then its corresponding ontology instance is $a' = (c_a, Dat_{a'}, Rel_{a'})$, and every $\alpha \in Dat_{a'}$ has value(s) in $V_{\alpha a}$ and every $\rho \in Rel_{a'}$ has values in $V_{\rho a}$.

For defining the problem of coreference resolution formally, we introduce the following collative relations on i-objects $a, b \in A$:

- *duplication*: a and b are duplicates ($a = b$) iff $Atr^K{}_a = Atr^K{}_b$ and $P_a = P_b$;
- *ontological equivalence*: a and b are ontological equivalents ($a \equiv b$) iff $Atr^K{}_a = Atr^K{}_b$, and $P_a \neq P_b$;
- *coreference*: a and b are coreferential candidates ($a \approx b$) iff $c_a \simeq^i c_b$, and $Atr^K{}_a \subseteq Atr^K{}_b \vee Atr^K{}_b \subseteq Atr^K{}_a$, where $Atr^K{}_a \subseteq Atr^K{}_b$ iff $\forall \gamma_a \in Atr^K{}_a : V_{\gamma a} \neq \emptyset \Rightarrow \exists \delta_b \in Atr^K{}_b : \gamma_a \subseteq \delta_b)$, where $\gamma_a \subseteq \delta_b$ iff $(\gamma_a, \delta_b \in Dat_O \wedge V_{\gamma a} \subseteq V_{\delta b}) \vee (\gamma_a, \delta_b \in Rel_O \wedge V_{\gamma a} \subseteq^r V_{\delta b})$, where $\subseteq^r$ is defined in the next paragraph.

Further we say just *co-candidates* instead of coreferential candidates.

We define for i-objects the following notions, taking into account i-objects' co-candidates. Let a, b, $c \in A$, and X, $Y \subset A$.

- *The coreferential group of the i-object a (co-group)* is $c(a) = \{ x \in A \mid x \approx a \}$.
- *The co-group of the set X* is $c(X) = \bigcup_{x \in X} c(x)$.
- *Coreferential membership*: $a \in^r X$ iff $a \in c(X)$.
- *Coreferential inclusion*: $X \subseteq^r Y$ iff $\forall x \in X : x \in^r Y$.
- *Coreferential intersection*: $X \cap^r Y = c(X) \cap c(Y)$.
- *Coreferential conflict*: i-objects a and b are in the coreferential conflict with respect to i-object c ($a \leftrightsquigarrow^c b$) iff $a \approx c \wedge b \approx c \wedge a \notin c(b)$.

The coreferential conflict means that some i-object is a co-candidate for two non-coreferential i-objects.

The coreference resolution problem is to detect if given i-objects correspond to the same ontology instance. Our algorithm for coreference resolution discussed in Section 4 constructs conflict-free co-groups of co-candidates. This construction uses *the coreference similarity* of i-objects for resolving coreferential conflicts. The measure of coreference similarity for i-objects a and b is denoted as $cs(a,b)$. If $a \leftrightsquigarrow^c b$, then we say that *the coreferential conflict is resolved to a* iff $cs(a,c) > cs(b,c)$.

The measure of the coreference similarity $cs(a,b)$ is calculated as the normalized sum of four measures – semantic $S(a,b)$, context $C(a,b)$, position $P(a,b)$ and grammar measures $G(a,b)$ – as follows: $cs(a,b) = ¼ (S(a,b) + C(a,b) + P(a,b) + G(a,b))$. We leave for future work a more precise estimation of the contribution of each component to this measure that may change the corresponding coefficients in the formula.

The semantic measure is discussed in the next section in detail, while the other three measures are briefly explained here. *The context measure of similarity C(a,b)* takes into account the information connectivity of i-objects in a given text. This measure depends on the number of i-objects which directly or indirectly use (1) attribute values from both a and b, and (2) attribute values borrowed by a from b, and by b from a, for the evaluation of their own attributes. *The position measure of similarity P(a,b)* takes into account various forms of closeness of i-objects in an input text. This measure depends on the number of segments, co-candidates in the conflict, and lexemes placed between the positions of a and b. *The grammar measure of similarity G(a,b)* is based on the standard linguistic features such as gender, number, person, etc. The details of these measures' definitions can be found in [6].

3. THE SEMANTIC MEASURE OF COREFERENCE SIMILARITY

The semantic measure of coreference similarity takes into account the attribute similarity of i-objects. This measure combines 11 types of the similarity which we summarize in Table 1. These types correspond to the properties introduced in Definition 1. Here $a, b \in A$, $\gamma_a \in Atr_a$, $\delta_b \in Atr_b$, and $a \approx b$. The measure of semantic similarity is defined by the normalized sum of all attribute similarity powers: $S(a,b) = |Sim(a,b)|^{-1} \sum_{(\gamma a,\ \delta b) \in Sim(a,b)} sim(\gamma_a, \delta_b)$, where $Sim(a,b) = \{(\gamma_a, \delta_b) \mid sim(\gamma_a, \delta_b) \neq 0\}$ is the set of similar attributes with the non-zero similarity power $sim(\gamma_a, \delta_b)$.

In Table 1, letter x denotes the type of similarity: $x \in \{d, r, \sqcap, \sqcup, \circ, \rhd, \smile, \sqsubseteq, *, t, s\}$. *The ontology condition* $\mathcal{O}^x$ is composed of the condition on the attributes and the corresponding necessary condition $\mathcal{N}^x$ from Proposition 1. This necessary condition is used when the properties of attributes in Definition 1 are unknown for a given populating ontology. *The value condition* $\mathcal{V}^x = (S^x \neq \emptyset \wedge E^x = \emptyset)$, where S^x is the set of similar values and E^x is the set of common values in the three cases of similarity (in other cases E^x is not necessary to define). *The x-similarity condition* is $\mathcal{A}^x = \mathcal{O}^x \wedge \mathcal{V}^x$. *The power of similarity* with respect to attributes γ_a and δ_b is $sim(\gamma_a, \delta_b)$. For a relation attribute γ, we introduce *the inverse cardinality* $ic(\gamma) = cardinality(\gamma^{\smile})$, where cardinality is the standard numeric property of ontology relations [9]. The value of $ic(\gamma)$ characterizes the number of how many distinct instances may or must be related with the same instance by the relation corresponding to γ. This value is used in the computation of the power of similarity.

Following Table 1, we consider that for the i-objects a and b *the attribute* γ_a *is x-similar to attribute* δ_b iff $\mathcal{A}^x$ holds, and the power of the x-similarity is $sim(\gamma_a, \delta_b)$. In the table, $\alpha_a \in Dat_a$, $\beta_b \in Dat_b$, $\rho_a \in Rel_a$, $\xi_b \in Rel_b$, $i(\gamma) = ic(\gamma)^{-1}$, $i(\rho_a, \xi_b) = (ic(\rho_a) \cdot ic(\xi_b))^{-1}$ and the normalizing coefficients are $Norm(\alpha_a,\beta_b) = ½(|V_{\alpha a}|^{-1} + |V_{\beta b}|^{-1})$ and $Norm(\rho_a,\xi_b) = ½(|c(V_{\rho a})|^{-1} + |c(V_{\xi b})|^{-1})$.

Proposition 2. Let $X \subseteq \{d, r, \sqcap, \sqcup, \circ, \rhd, \smile, \sqsubseteq, *, t, s\}$. If for the attributes of co-candidates a and b the semantic similarity condition $\bigwedge_{x \in X} \mathcal{A}^x$ holds, then these co-candidates correspond to the same ontology instance with the integral accuracy $cs(a,b)$ which uses the semantic similarity powers sim^x through the semantic similarity measure $S(a,b)$.

The proof of the proposition is based on Definition 1 and Proposition 1.

Table 1. The types of semantic similarity

Similarity	$\mathcal{O}^x$	$\mathcal{V}^x = (S^x \neq \emptyset \wedge E^x = \emptyset)$	$sim(\gamma_a, \delta_b)$
Data $\alpha_a \sim_d \beta_b$	$\alpha \simeq^j \beta \vee \mathcal{N}^d$	$S^d = V_{\alpha a} \cap V_{\beta b}$	$\lvert S^d\rvert \cdot Norm(\alpha_a,\beta_b)$
Relation	$\rho \simeq^j \xi$	$S^r =$	$\lvert S^r\rvert \cdot$

$\rho_a \sim_r \xi_b$	$\vee \mathcal{N}^r$	$V_{\rho a} \cap^r V_{\xi b}$	$Norm(\rho_a, \xi_b)$
Transitive $\rho_a \sim_t \xi_b$	$\rho = \xi$, $\rho \in Rel^t_O$ $\vee \mathcal{N}^t$	$E^t=V_{\rho a}\cap^r V_{\xi b}$, $S^t = \{ (o, p) \mid o \in^r V_{\rho a}, p \in^r V_{\xi b}, p \in^r V_{\rho o} \vee o \in^r V_{\rho p}\}$	$\lvert S^t\rvert \cdot i(\rho_a) \cdot (\lvert c(V_{\rho a})\rvert \cdot \lvert c(V_{\xi b})\rvert)^{-1}$
Symmetric $\rho_a \sim_s \xi_b$	$\rho = \xi$, $\rho \in Rel^s_O$ $\vee \mathcal{N}^s$	$E^s=V_{\rho a}\cap^r V_{\xi b}$, $S^s=\{o \mid o \in^r V_{\rho a} \wedge b \in^r V_{\rho o}$ or $o \in^r V_{\xi b} \wedge a \in^r V_{\xi o}\}$	$\lvert S^s\rvert \cdot i(\rho_a) \cdot \lvert c(V_{\rho a} \cup V_{\xi b})\rvert^{-1}$
Inverse $\rho_a \sim_{\smile} \xi_b$	$\rho = \xi$, $\exists \pi \in Rel_O$: $\rho = \pi^{\smile}$ $\vee \mathcal{N}^{\smile}$	$E^{\smile}=V_{\rho a}\cap^r V_{\xi b}$, $S^{\smile} = \{ o \mid o \in^r V_{\rho a} \cup V_{\xi b}$ and $a \in^r V_{\pi o} \vee b \in^r V_{\pi o}\}$	$\lvert S^{\smile}\rvert \cdot i(\rho_a) \cdot \lvert c(V_{\rho a} \cup V_{\xi b})\rvert^{-1}$
Intersection $\rho_a \sim_{\sqcap} \xi_b$	$\exists \pi \in Rel_O$: $\pi = \rho \sqcap \xi$ $\vee \mathcal{N}^{\sqcap}$	$S^{\sqcap} = \bigcup_{o \in c\pi} V_{\pi o}\cap^r V_{\rho a}\cap^r V_{\xi b}$	$\lvert S^{\sqcap}\rvert \cdot i(\rho_a, \xi_b) \lvert V_{\rho a} \cap^r V_{\xi b}\rvert^{-1}$
Union $\rho_a \sim_{\sqcup} \xi_b$	$\exists \pi \in Rel_O$: $\xi = \rho \sqcup \pi$ $\vee \mathcal{N}^{\sqcup}$	$S^{\sqcup} = V_{\xi b} \cap^r V_{\rho a}$	$\lvert S^{\sqcup}\rvert \cdot i(\rho_a) \cdot Norm(\rho_a, \xi_b)$
Inclusion $\rho_a \sim_{\sqsubseteq} \xi_b$	$\rho \sqsubseteq \xi$ $\vee \mathcal{N}^{\sqsubseteq}$	$S^{\sqsubseteq} = V_{\xi b} \cap^r V_{\rho a}$	$\lvert S^{\sqsubseteq}\rvert \cdot i(\rho_a) \cdot Norm(\rho_a, \xi_b)$
Closure $\rho_a \sim_* \xi_b$	$\rho = \xi^*$ $\vee \mathcal{N}^*$	$S^* = V_{\xi b} \cap^r V_{\rho a}$	$\lvert S^*\rvert \cdot i(\rho_a) \cdot Norm(\rho_a, \xi_b)$
Refinement $\rho_a \sim_{\rhd} \xi_b$	$\exists \pi \in Rel_O$: $\rho = \xi \rhd \pi$ $\vee \mathcal{N}^{\rhd}$	$S^{\rhd}=\{ (o, p) \mid o \in^r V_{\rho a}$ and $p \in V_{\xi b}\cap^r V_{\pi o}\}$	$\lvert S^{\rhd}\rvert \cdot i(\rho_a,\xi_b) \cdot (\lvert c(V_{\rho a})\rvert \cdot \lvert c(V_{\xi b})\rvert)^{-1}$
Composition $\rho_a \sim_{\circ} \xi_b$	$\exists \pi \in Rel_O$: $\rho = \xi \circ \pi$ $\vee \mathcal{N}^{\circ}$	$S^{\circ}=\{o \mid o \in c^{\pi}$ and $o \in^r V_{\xi b} \wedge V_{\rho a}\cap^r V_{\pi o} \neq \emptyset\}$	$\lvert S^{\circ}\rvert \cdot i(\rho_a,\xi_b) \cdot \lvert c(V_{\xi b})\rvert^{-1}$

Let us illustrate our introduced framework by a practice-relevant example with co-candidates whose attributes are related by the refinement and composition relations. The ontology's domain of our example is the area of Technical Documentation for Industrial Process Control (TDIPC). We consider the natural language description of a bottle-filling system example from [21].

Let us discuss the following fragment of the text that demonstrates the refinement similarity in the description:

> *A filler tank holds fluid. In this system, the fluid is heated and maintained at 100 degrees Celsius. Although this might typically be performed with a PID implementation, in this case the steam valve is opened and steam is inserted into the tank when the temperature falls below 100 degrees, and closes when the temperature reaches 110 degrees Celsius.*

For this text fragment, our algorithm creates the following i-objects:

a = heater(type = $\emptyset$, … ρ_a = heat (object: *fluid*)),
b = heater(type = *steam*, … ξ_b = heat (reservoir: *tank*)),
fluid = object(… π = inside (reservoir: *tank*)).

These i-objects are co-candidates, because they have the identical class heater, and the key data attribute type is not defined for a. The refinement similarity of these i-objects is $sim(\rho_a, \xi_b) = 1$, because the values of the attributes are consistent ($S^{\rhd} \neq \emptyset$), the ontology of TDIPC contains the refinement relation: heat = heat ⊳ inside, and the inverse cardinality of heat is equal to 1.

Note that the previous approaches for coreference resolution, e.g., from [10, 25], would miss this coreference, because they consider coreferential candidates only with identical (may be after some normalization) key attributes, but the key attributes of the example i-objects are different.

Our next example text fragment illustrates the composition similarity:

> *There is a valve in the bottom of the filler tank that is opened when an empty bottle is present, the fluid is present, and the fluid is at or above 100 degrees. A photosensor attached to the filler tank determines when the bottle is full.*

For this text fragment, our algorithm of text analysis creates the following i-objects:

a = bottle(… ρ_b = open (gate: *valve*)),
b = bottle(… ξ_b = fill_from (reservoir: *tank*)),
valve = gate(… π = in_bottom (reservoir: *tank*)).

These i-objects are co-candidates, because they have the identical class bottle and their key attributes are not defined. The compositional similarity of these i-objects is $sim(\rho_a, \xi_b) = 1$, because the values of the attributes are consistent ($S^{\circ} \neq \emptyset$), in the ontology of TDIPC the following compositional relation is given: fill_from = open ∘ in_bottom, and the inverse cardinalities of fill_from and open are equal to 1.

The approach to coreference resolution from [25] can consider the example i-objects in this case as potential coreferents due to the coincidence of their classes which are treated as key characteristics, but this coreference will not be established, because the example i-objects have different names and values of the attribute relations. The approach to coreference resolution from [10] would not consider the example i-objects as potential coreferents, because they have no key attributed defined.

Summarizing, the previously suggested approaches would miss some coreferents which our approach would consider; this demonstrates the higher degree of completeness of our approach to the coreference resolution as compared to related work.

4. A MULTI-AGENT APPROACH TO COREFERENCE RESOLUTION IN THE INFORMATION EXTRACTION

The coreference similarity measures, including the semantic measure from the previous section, are used in our coreference resolution algorithm which is the part of our general approach to information extraction (IE) for the ontology population outlined below. In this section, we sketch the approach as a whole, and we provide informal descriptions of the actions of agents that execute our multi-agent algorithm of the coreference resolution.

The input of our IE-system comprises: an ontology of some particular subject domain, the ontology population rules, semantic and syntactic models for the language of the subject domain, the term vocabulary, and input data as a finite natural language text. The output is the ontology populated by information from the text.

Our IE-system consists of the following five sequential modules.

1. The module of *lexical analysis* executes a preliminary extraction of subject domain terms from a given text [14]. This module takes the semantic and syntactic models, the term vocabulary, and the input text, and it produces *the terminological cover* (the set of lexical objects without structural information). Every lexical object has the same structure as i-objects: it stores the grammatical and structural information, but it has exactly one data value in a data domain from D_O, and its class is a semantic class of the term vocabulary.
2. *The segmentator* module performs segmentation of the text into formal and genre fragments (sentences, sections, headlines, etc) [22]. This module receives the semantic and syntactic models, and the input text as the input, and its output is *the segment cover* representing text decomposition into formal and genre subunits.
3. *The main analysis* module constructs objects, corresponding to instances of subject domain ontology, from the terms [4], and resolves coreference [6]. The input for this module is the terminological cover with the structural information from the segment cover, and *the analysis rules* which implement semantic and syntactic models and ontology population rules. They are formulated by experts taking into account the ontology and language of subject domain. This module produces the set of i-objects with resolved coreference and unresolved lexical/syntactical ambiguity.
4. *The disambiguation* module resolves lexical and syntactic ambiguity [5]; it takes the output of the main analysis module as its input, and yields the set of i-objects without ambiguities.
5. *The population* module updates the ontology with the processed objects (planned as future work). The module's input is the output of the disambiguation module and the given ontology, and its output is this ontology populated by information from the text.

Let us describe the main analysis module which performs the coreference resolution. The main analysis module performs two tasks in parallel: construction of i-objects and coreference resolution.

In the constructing process, the module generates new information based on information (attribute values) taken from i-objects and lexical objects using the analysis rules. This information is used to define new attribute values of existing i-objects and to generate new i-objects. Following the analysis rules, the module takes into consideration only linguistically and ontologically compatible sets of i-objects. Using information from one i-object for another i-object sets *the information connection* between these i-objects labeled by this information. These connections keep the history of the evolution of an i-object. They are used by the disambiguation module for evaluating the integration of the i-object, i.e. amount of information related to the i-object in the text. This construction process terminates when new information cannot be generated.

For the coreference resolution task the main analysis module constructs and updates the co-groups of i-objects in parallel with constructing i-objects. The coreferential conflict resolution in the co-groups based on the similarity of i-objects is performed after the termination of constructing i-objects. The result of this process are the conflict-free co-groups. The attribute values and information connections of i-agents in these co-groups are joined in the main analysis module for the further processes of the lexical/syntactical disambiguation and ontology population. One advantage of our approach to coreference resolution is that this joining improves the quality of the disambiguation and population processes, because it allows the corresponding modules to take into consideration all information about objects accessible from the text. Another advantage is that using the multiple similarity measures of the coreferents allows us to more precisely estimate the integration of i-objects into a given text than in our previous work [5].

In our multi-agent framework, we assign a separate agent for every i-object, every analysis rule, and every ontology class. These agents perform the following tasks of text analysis for ontology population: creating/updating i-objects and coreference resolution in parallel by i-agents, rule agents, and class agents, and then the ambiguity resolution by i-agents. These agents communicate and exchange data for executing their tasks. There is also an auxiliary agent: the master-agent detects terminations and coordinates all other agents in the disambiguation process. For the details of creating i-objects and disambiguation, see [4,5]. The result of agent interactions is the system of i-objects without the coreferences, lexical, and syntactical ambiguities. All agents execute their protocols in parallel until, from time to time, it happens that none of the agents can proceed. Such termination events are detected by the master agent. We use our original algorithm for termination detection, which is based on activity counting. After detecting termination, the master agent sends coordination signals, depending on the task performed, to other agents. Our system of agents is dynamic: the rule agents can create new information agents, the class agents can kill the i-agents by joining duplicates and ontological equivalents, and co-candidates (at the end of the coreference resolution process), and, in the disambiguation process, the master agent can kill the i-agents whose i-objects are weakly integrated in a given text. The agents are connected by duplex channels. The master agent is connected with all agents, the i-agents are connected with their rule agents, class agents, and successors/predecessors by information connections. We assume that messages are transmitted instantly via a reliable medium and stored in channels until being read.

Let us briefly describe the process of coreference resolution by the class agents. Here, we do not distinguish an i-agent from its i-object if there is no ambiguity. Every class agent performs the following tasks:

1) creating the co-group for every newly born i-agent;
2) updating the co-group for every i-agent in a case of its key attribute update;
3) regulating the attribute exchange between i-agents;
4) computing the measures of the coreference similarity for i-objects in co-groups by formulas from Sections 2 and 3;
5) resolving the coreferential conflicts using the calculated measures; and
6) generating for every conflict-free co-group the integrating i-object (with the corresponding i-agent) by joining the i-objects from the co-group.

Every class agent acts at its level of the class hierarchy, i.e. in processing pairs of i-agents (testing for collative relations in creation/update co-groups, computing the similarity measures etc.); at least one i-agent must be in the class of this class agent. The higher class agents use results of constructing conflict-free groups from the lower class agents. The details of the coreference resolution process are described in [6], where we use a simpler semantic measure of coreference similarity in the task 4 and the computation of the measure is not discussed. Here we use more complex and precise semantic measure, so it is reasonable to describe its computation in detail.

Let us describe how the semantic similarity measure is computed. In this paper we consider the relation attributes which have exactly one property from Definition 1. If these properties are given for the given ontology (for example, they are summarized in some table RP) then the algorithm of computing the semantic similarity measure is trivial: it simply checks in the table RP if given relations ρ and ξ satisfy some property and computes the corresponding similarity measure $sim(\rho,\xi)$ by formulas from Table 1. If the table of properties RP does not exist, then the algorithm of computing the measure must check the necessary conditions from Proposition 1. In order to reduce the computation of the necessary conditions, it is reasonable to organize them into "implication" chains. The following proposition describes three such chains. The proof follows directly from the definitions of the necessary conditions.

Proposition 3. Let $\rho, \xi, \pi \in Rel_O$. Then it holds:

- $\mathcal{N}^{\smile} \wedge (\pi = \rho) \Rightarrow \mathcal{N}^{t} \vee \mathcal{N}^{s} \Rightarrow \mathcal{N}^{r}$;
- $\mathcal{N}^{*} \Rightarrow \mathcal{N}^{\sqsubseteq} \vee \mathcal{N}^{\sqcup}$ *and* $(\mathcal{N}^{\sqsubseteq} \vee \mathcal{N}^{\sqcup}) \wedge (\pi=\rho) \Rightarrow \mathcal{N}^{\sqcap}$;
- $\mathcal{N}^{\circ} \Rightarrow \mathcal{N}^{\triangleright}$.

This proposition is the base for the following algorithm of computing the semantic similarity measure in the case of absence of a specification for the ontology properties of relation attributes. Using the implication chains from the proposition allows us not to compute many times the truth values of the same conditions. We introduce the following notation for conjunctions of Boolean formulas: $\varphi^{x} = \varphi^{y} \wedge \varphi^{x/y}$. The following procedure `SiMeasure(a,b)` returns the semantic similarity measure $S(a,b)$ = `SimMes` for i-objects $a = (c_a, Dat_a, Rel_a)$ and $b = (c_b, Dat_b, Rel_b)$. In the procedure, function $\mathtt{sim}^{x}(\gamma,\delta)$ returns the power of the corresponding similarity using the formulas from Table 1.

```
SiMeasure(a,b) ::
int S = 0; N = 0;
1. forall $\alpha \in Dat_a$, $\beta \in Dat_b$
2.     if $\mathcal{N}^{d}(\alpha,\beta)$ then S = S + $sim^{d}(\alpha, \beta)$; N++;
3. forall $\rho \in Rel_a$, $\xi \in Rel_b$
4.     if $\mathcal{N}^{r}(\rho, \xi)$ then
               S = S + $sim^{r}(\rho, \xi)$; N++; continue;
5.     else if $\mathcal{N}^{t/r}(\rho, \xi)$ then
               S = S + $sim^{t}(\rho, \xi)$; N++; continue;
6.         else if $\mathcal{N}^{s/r}(\rho, \xi)$ then
                   S=S+$sim^{s}(\rho,\xi)$; N++; continue;
7.             else if $\mathcal{N}^{\smile/t}(\rho, \xi)$ then
                       S = S + $sim^{\smile}(\rho,\xi)$;
                           N++; continue;
8.     if $\mathcal{N}^{\sqcap}(\rho,\xi)$ then
             S = S + $sim^{\sqcap}(\rho, \xi)$; N++; continue;
9.     else if $\mathcal{N}^{\sqcup/\sqcap}(\rho, \xi)$ then
             S = S + $sim^{\sqcup}(\rho, \xi)$; N++; continue;
10.        else if $\mathcal{N}^{\sqsubseteq/\sqcap}(\rho, \xi)$ then
                 S=S+$sim^{\sqsubseteq}(\rho,\xi)$; N++; continue;
11.            else if $\mathcal{N}^{*/\sqsubseteq}(\rho, \xi)$ then
                       S = S + $sim^{*}(\rho, \xi)$;
                           N++; continue;
12.    if $\mathcal{N}^{\triangleright}(\rho, \xi)$ then
```

```
                  S = S + $sim^{\triangleright}(\rho, \xi)$; N++;
     else if $\mathcal{N}^{\circ/\triangleright}(\rho, \xi)$ then
                  S = S + $sim^{\circ}(\rho, \xi)$; N++;
return SimMes = S/N;
```

5. CONCLUSION

Our main contribution in this paper is a formal framework for coreference resolution in the process of ontology population. The novelty of the suggested framework is the use of multiple properties of ontology classes and relations for solving the coreference resolution problem. Using multiple properties provides a significantly more precise and complete coreference identification due to taking into account more similarity factors than just the elements' equality as done in previous work. The properties used in our framework include class and attribute hierarchy, intersection, union, composition, refinement, inverse, inclusion, reflexive-transitive closure, transitivity, and symmetry. We describe in detail how these properties are efficiently used in evaluating the semantic similarity of coreferential candidates. This evaluation is integrated into our multi-agent system of information extraction (IE) from texts in a natural language, which significantly speeds up the IE process as compared to a sequential implementation.

As shown in Sections 3 and 4, our approach has several advantages over the previous work: 1) it provides a higher degree of completenes regarding the considered coreferents, 2) it improves the quality of the disambiguation and population processes, because it allows the corresponding modules to take into consideration all information about objects accessible from the text, and 3) using the multiple similarity measures of the coreferents allows us to more precisely estimate the integration of i-objects into a given text than in our own previous work [5].

In the future work, we plan to extend the above list of properties used in our framework with their meaningful combinations which appear in the practice of information extraction. While the presented properties are defined for binary ontology relations, we intend to specify them for n-ary ontology relations which represent situations and events of the real world. These properties will additionally improve the quality of coreference resolution. For the better estimation of the impact of the semantic similarity on the integrated evaluation of coreference similarity, we will investigate the frequency and significance of using particular ontology properties for defining the corresponding coefficients in the similarity evaluation formula.

6. ACKNOWLEDGEMENTS

This work has been supported by the Russian Foundation for Basic Research (grant 15-07-04144, grant 17-07-01600), by the Siberian Branch of the Russian Academy of Science (Integration Grant n.15/10 “Mathematical and Methodological Aspects of Intellectual Information Systems”), as well as by the DFG Cluster of Excellence CiM and the BMBF project HPC2SE at the University of Muenster, Germany.

7. REFERENCES

[1] S. E. Brennan, M. W. Friedman, and C. J. Pollard, “A centering approach to pronouns,” in *Proceedings of the 25th annual meeting on Association for Computational Linguistics,* Stanford, California, USA, July 06-09, 1987, pp. 155-162.

[2] J. G. Carbonell and R. D. Brown, “Anaphora resolution: a multi-strategy approach,” in *Proceedings of the 12 International Conference on Computational Linguistics*, Budapest, Hungary, August 22-27, 1988, pp. 96-101.

[3] P. Elango, *Coreference Resolution: A Survey*, Technical Report, UW-Madison, 2005, 8 p.

[4] N. O. Garanina and E. A. Sidorova, “Ontology population as algebraic information system processing based on multi-agent natural language text analysis algorithms,” *Programming and Computer Software*, vol. 41, issue 3, pp. 140-148, 2015.

[5] N. Garanina, E. Sidorova, “Context-dependent lexical and syntactic disambiguation in ontology population,” in *Proceedings of the 25th International Workshop on Concurrency, Specification and Programming*, Rostock, Germany, September 28-30, 2015, pp. 101-112.

[6] N. Garanina, E. Sidorova, and I. Kononenko, “A distributed approach to coreference resolution in multiagent text analysis for ontology population,” in *Proceedings of the Ershov Informatics Conference (PSI’2017)*, June 27-29, 2017, Moscow, Russia.

[7] B. J. Grosz, S. Weinstein, and A. K. Joshi, “Centering: A framework for modeling the local coherence of discourse,” *Journal Computational Linguistics*, vol. 21, issue 2, pp. 203-225, 1995.

[8] S. M. Harabagiu, R. C. Bunescu, and S. J. Maiorano, “Text and knowledge mining for coreference resolution,” in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, Pittsburgh, PA, USA, June 2-7, 2001, pp. 1-8.

[9] *Handbook on Ontologies*, Edt.: S. Staab, R. Studer, International Handbooks on

Information Systems, Springer Berlin Heidelberg, 2009, 808 p.

[10] D. Hladky, C. Ehrlich, I. Efimenko, V. Vorobyov, *Discover Shadow Groups from the Dark Web*, in: M. Last, A. Kandel (Eds.), Web Intelligence and Security: Advances in Data and Text Mining Techniques for Detecting and Preventing Terrorist Activities on the Web, IOS Press, 2010, pp. 67-81.

[11] J. Hobbs, *Resolving Pronoun References*, in: B. J. Grosz, B. L. Webber, K. S. Jones (Eds.), Readings in Natural Language Processing, Morgan Kaufmann Publishers Inc., San Francisco,1986, pp. 339-352.

[12] A. A. Kibrik, *Anaphora in Russian Narrative Discourse: A Cognitive Calculative Account*, in: B. Fox (ed.) Studies in anaphora, J. Benjamins Pub., Amsterdam, 1996, pp. 255-304.

[13] A. A. Kibrik, G. B. Dobrov, M. V. Khudyakova, N. V. Loukachevitch, A. Pechenyj, “A corpus-based study of referential choice: Multiplicity of factors and machine learning techniques,” in *Proceedings of the 13th International Conference Cognitive Modeling in Linguistics*, Corfu, Greece, September 22-29, 2011, pp. 118-126.

[14] I. S. Kononenko, E. A. Sidorova, “Language resources in ontology-driven information systems,” in *Proceedings of the First Russia and Pacific Conference on Computer Technology and Applications*, Vladivostok, Russia, September 6-9, 2010, pp. 18-23.

[15] E. Motta, S. Siqueira, A. Andreatta, “An unsupervised rule-based method to populate ontologies from text,” in *Proceedings of the 5th International Conference on Web Information Systems and Technologies*, Lisboa, Portugal, March 23-26, 2009, pp. 157-169.

[16] R. Mitkov, *Anaphora Resolution: The State of the Art*, Technical report, University of Wolverhampton, 1999, 34 p.

[17] R. Mitkov, *Anaphora resolution*, in R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics, Oxford University Press, New York, 2003, pp. 266-283.

[18] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos, *Ontology Population and Enrichment: State of the Art*, in G. Paliouras, C. D. Spyropoulos, G. Tsatsaronis (Eds.), Knowledge-driven Multimedia Information Extraction and Ontology Evolution, Springer-Verlag, Berlin, 2011, pp. 134-166.

[19] R. Prokofyev, A. Tonon, M. Luggen, L. Vouilloz, D. E. Difallah, and P. Cudre-Mauroux, “SANAPHOR: Ontology-based coreference resolution,” in *Proceedings of the 14th International Semantic Web Conference*, Bethlehem, Pennsylvania, USA, October 11-15, 2015, pp. 458-473.

[20] E. Rich and S. Luper Foy, “An architecture for anaphora resolution,” in *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, USA, February 9-12, 1988, pp. 18-24.

[21] S. G. Shanmugham, C. A. Roberts, “Application of graphical specification methodologies to manufacturing control logic development: a classification and comparison,” *Int. J. Computer Integrated Manufacturing*, vol. 11, issue 2, pp. 142-152, 1998.

[22] E. A. Sidorova, I. S. Kononenko, “Representation and use of the jenre structure of documentation in text processing,” in *Proceedings of the Science-Intensive Software Workshop*, Novosibirsk, Russia, June 15-19, 2009, pp. 248-254. (in Russian)

[23] W. M. Soon, H. T. Ng, and D. C. Y. Lim, “A machine learning approach to coreference resolution of noun phrases,” *Journal Computational Linguistics*, vol. 27, issue 4, pp. 521-544, 2001.

[24] Y. Wilks, *Preference Semantics*, ed. by E. Keenan (Ed.), The Formal Semantics of Natural Language, Cambridge University Press, 1975, pp. 329-348.

[25] M. Yatskevich, C. Welty, and J. W. Murdock, “Coreference resolution on RDF Graphs generated from information extraction: first results,” in *Proceedings of ISWC’06 Workshop on Web Content Mining with Human Language Technologies*, Athens, GA, USA, November 6, 2006.

[26] G. D. Zhou and J. Su, “A high-performance coreference resolution system using a constraint-based multi-agent strategy,” in *Proceedings of the 20th International Conference on Computational Linguistics,* Geneva, Switzerland, August 23-27, 2004, pp. 522-528.

Natalia Garanina *has been senior researcher of the Laboratory of Theoretical Programming at the A.P. Ershov Institute of Informatics Systems (Novosibirsk, Russia) since 2014 and lecturer at Novosibirsk State University since 2007. She holds MSc degree from the Novosibirsk State University and PhD degree from the A.P. Ershov Institute of Informatics Systems. Dr. Garanina has about 50 peer-reviewed publications*

in international journals and conferences. Her research interests include Distributed Systems, Formal Verification, Artificial Intelligence, Non-classical Logics, and Domain theory.

***Elena Sidorova** has been senior researcher of the Laboratory of Artificial Intelligence at the A.P. Ershov Institute of Informatics Systems (Novosibirsk, Russia) since 2014, and was a leader of several projects in computational linguistics. She holds MSc degree from the Novosibirsk State University and PhD degree in Computer Science. Dr. Sidorova has about 50 peer-reviewed publications in international journals and conferences. She has research interests in areas of NLP Systems, Multi-agent Systems, Knowledge Representation, and Ontology Engineering.*

***Irina Kononenko** has been researcher of Laboratory of Artificial Intelligence at the A. P. Ershov Institute of Informatics Systems (Novosibirsk, Russia) since 1990 and principal investigator of several projects in computational linguistics. She graduated from the humanitarian faculty of Novosibirsk State University, specializing in "Mathematical Linguistics" in 1975. I. Kononenko has about 40 peer-reviewed publications in international journals and conferences. She has research interests in areas of NLP Systems, Linguistic models of the Natural Language, Knowledge Representation, Question-and-answer Systems and Human-computer Interaction.*

***Sergei Gorlatch** has been Full Professor of Computer Science at the University of Muenster, Germany since 2003. He holds MSc degree from the State University of Kiev, PhD degree from the Institute of Cybernetics in Ukraine, and the Habilitation degree from the University of Passau, Germany.*

Prof. Gorlatch has about 200 peer-reviewed publications in international journals and conferences. He has been principal investigator in several projects on parallel, distributed, Grid and Cloud computing, funded by the European Commission and by German national bodies.