Using Multiple Semantic Measures in a Framework of Coreference Resolution in the Process of Ontology Population*

Natalia Garanina, Elena Sidorova, Irina Kononenko, and Sergei Gorlatch

A.P. Ershov Institute of Informatics Systems,
Lavrent'ev av., 6, Novosibirsk 630090, Russia
University of Muenster, Germany
{garanina,lsidorova}@iis.nsk.su, irina_k@cn.ru, 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 novel contribution is a formal framework which extends the state-of-the-art approaches by using multiple semantic similarity properties in the coreference resolution process. Using the additional semantic similarity measures for evaluating coreference candidates improves the quality of the resolution process, especially for complex objects.

1 Introduction

The process of ontology population is the widely discussed problem of adding new instances of concepts to the ontology. This process is a part of ontology acquisition [11] from a domain-specific content, which is mostly represented in a natural language. In this context, the solution for 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, 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 research has been supported by Russian Foundation for Basic Research (grant 15-07-04144, grant 17-07-01600) and Siberian Branch of Russian Academy of Science (Integration Grant n.15/10 "Mathematical and Methodological Aspects of Intellectual Information Systems").

The contribution of this paper is a 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. The proposed framework is used to improve our coreference resolution algorithm [5] for making the decision on the candidate admissibility which is used in our general approach to text analysis and information extraction for populating a subject domain ontology. In our approach, the following IE tasks are performed: the preliminary extraction of subject domain terms from a given text [8]; the segmentation of the text into formal and genre fragments (sentences, sections, headlines, etc) [13]; the construction of objects, corresponding to instances of a subject domain ontology, from the terms [3]; the coreference resolution [5]; the lexical and syntactic disambiguation [4]; and the update of the ontology with the processed objects (in plans). In our framework, the coreference resolution problem is to detect if some group of retrieved objects refer to the particular ontology instance.

There are several basic approaches to coreference resolution covered in the literature. The most important trends in the field can be found in the comprehensive surveys [10, 1, 12]. These trends can be categorized into rule-based and machine learning approaches. The rule-based approaches rely on hand-coded heuristics that specify whether two expressions can or cannot corefer, and exploit a lot of domain and linguistic knowledge. The machine learning approaches use corpora annotated with coreference information as a training data for automatic recognizing coreference. Unfortunately, in limited subject domains (for example, particular science and industrial domains such as technical documentation) representative training text corpora do not usually exist. In this cases, it is reasonable to use classical rule-based methods.

In the context of ontology population, the rule-based approaches 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 and, on the other hand, the ontology structure and knowledge represented in it help to solve IE domain-specific subtasks [9]. In [14, 7] the coreference 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 [7] seems somewhat limited as it 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, a special hand-crafted database is used. This identification knowledge base 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 [14], 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 which are mapped into an RDF graph. Af-

ter that 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. This case corresponds to using only the identity property of ontology elements.

Our approach to coreference resolution [5] 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), as 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 ontology properties of the 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.).

In this paper, we suggest to extend the list of the ontological properties used for coreference resolution with several other 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. In the concluding Section 4, we discuss future work.

2 Problem Statement and Base 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 [6].

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_α} 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_c^K. The key attributes can be data as well as relation attributes. We say that a is *an instance of the class* $c_a = (Dat_{c_a}, Rel_{c_a})$ $(a \in c_a)$ iff $a = (c_a, Dat_a, Rel_a)$, where every data attribute in Dat_a has a name $\alpha_a \in Dat_{c_a}$ with the values V_{α_a} from $V_{d_{\alpha_a}}$ and every relation attribute in Rel_a has a name $\rho_a \in Rel_{c_a}$ with the values V_{ρ_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 can 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 $\forall c_1, c_2 \in C_O : Atr_{c_1} \neq Atr_{c_2}$. *An 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 processes 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 sub-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*: $\rho = \rho' \triangleright \rho''$ 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*: $\rho \in Rel_O^{\mathrm{t}}$;
– *the symmetry*: $\rho \in Rel_O^{\mathrm{s}}$.

We extend the list of standard 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 goals of this paper – the 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 $\gamma, \gamma' \in Atr_O$, $c, c' \in C_O$, and $C, C' \subseteq C_O$.

– *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 $Hi(C) = \bigcup_{c' \in C} Hi(c')$.
– *Hierarchical inclusion*: $c \in^i C$ iff $c \in Hi(C)$.
– *Hierarchical-subset 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 $Hi(c) \cap Hi(c') \neq \emptyset$;
  • $\gamma \simeq^i \gamma'$ 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 $\alpha, \beta \in Dat_O$, $\rho, \xi, \pi \in Rel_O$.

– $\alpha \simeq^i \beta \Rightarrow \mathcal{N}^{\mathrm{d}} = (V_{d_{\alpha_a}} \subseteq V_{d_{\beta_b}} \vee V_{d_{\beta_b}} \subseteq V_{d_{\alpha_a}})$;
– $\rho \simeq^i \xi \Rightarrow \mathcal{N}^{\mathrm{r}} = (C_\rho \subseteq^i C_\xi \vee C_\xi \subseteq^i C_\rho)$;
– $\rho \in Rel_O^{\mathrm{t}} \Rightarrow \mathcal{N}^{\mathrm{t}} = (c^\rho \in^i C_\rho)$;
– $\rho \in Rel_O^{\mathrm{s}} \Rightarrow \mathcal{N}^{\mathrm{s}} = (c^\rho \in^i C_\rho)$.
– $\rho = \pi^{\smile} \Rightarrow \mathcal{N}^{\smile} = (c^\rho \in^i C_\pi \wedge c^\pi \in^i C_\rho)$;
– $\pi = \rho \sqcap \xi \Rightarrow \mathcal{N}^{\sqcap} = (c^\pi \in^i \{c^\rho\} \wedge c^\pi \in^i \{c^\xi\} \wedge C_\pi \subseteq^i C_\rho \cap^i C_\xi)$;
– $\xi = \rho \sqcup \pi \Rightarrow \mathcal{N}^{\sqcup} = (c^\rho \in^i \{c^\xi\} \wedge C_\rho \subseteq^i C_\xi)$;
– $\rho \sqsubseteq \xi \Rightarrow \mathcal{N}^{\sqsubseteq} = (c^\rho \in^i \{c^\xi\} \wedge C_\rho \subseteq^i C_\xi)$;
– $\rho = \xi^* \Rightarrow \mathcal{N}^{*} = (c^\rho = c^\xi \wedge C_\rho \subseteq^i C_\xi \wedge c^\xi \in^i C_\xi)$;
– $\rho = \xi \triangleright \pi \Rightarrow \mathcal{N}^{\triangleright} = (c^\xi \in^i \{c^\rho\} \wedge c^\pi \in^i C_\xi \wedge C_\pi \subseteq^i C_\rho)$;
– $\rho = \xi \circ \pi \Rightarrow \mathcal{N}^{\circ} = (c^\xi = c^\rho \wedge c^\pi \in^i C_\xi \wedge C_\pi = C_\rho)$;

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 $\alpha \in Dat_{c_a}$, and
  - Val_{α_a} is the set of information values $\bar{v} = (v_{\bar{v}}, s_{\bar{v}})$ with
    - the data value $v_{\bar{v}} \in d_\alpha$,
      a set of values of α_a is $V_{\alpha_a} = \{v_{\bar{v}} \mid \bar{v} \in Val_{\alpha_a}\}$,
    - $s_{\bar{v}}$ 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 $\rho \in Rel_{c_a}$, and
  - V_{ρ_a} is the set of i-objects of a class $c_{\bar{o}}$ from C_{ρ_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 $Atr_a = Dat_a \cup Rel_a$ as 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_{α_a} and every $\rho \in Rel_{a'}$ has values in V_{ρ_a}.

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

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

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

- *The coreferential group of the i-object* a is $cR(a) = \{a\} \cup \{x \in A \mid x \approx a\}$.
- *The coreferential group of the set* X is $cR(X) = \bigcup_{x \in X} cR(x)$.
- *Coreferential inclusion*: $a \in^r X$ iff $a \in cR(X)$.
- *Coreferential-subset inclusion*: $X \subseteq^r Y$ iff $\forall x \in X : x \in^r Y$.
- *Coreferential intersection*: $X \cap^r Y = cR(X) \cap cR(Y)$.
- *Coreferential conflict*: i-objects a and b are in the coreferential conflict with respect to i-object c ($a \overset{c}{\rightsquigarrow} b$) iff $a \approx c \wedge b \approx c \wedge a \notin cR(b)$. The coreferential conflict means that some i-object is a coreferential 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 constructs conflict-free groups of coreferential 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 $a \overset{c}{\rightsquigarrow} b$, 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 semantic $S(a,b)$, context $C(a,b)$, position $P(a,b)$ and grammar measures $G(a,b)$: $cs(a,b) = \frac{1}{4}(S(a,b) + C(a,b) + P(a,b) + G(a,b))$. We leave for future work a more precise estimation of the contribution of each component to this measure which 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, coreferential 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 [5].

3 The semantic measure of the coreference similarity.

The semantic measure of the coreference similarity takes into account the attribute similarity of i-objects. In Table 1, we summarize 11 types of the similarity, corresponding to the properties of 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: $S(a,b) = \frac{1}{|Sim_a^b|} \sum_{(\gamma_a,\delta_b)\in Sim_a^b} sim(\gamma_a,\delta_b)$, 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, the letter x denotes the type of a similarity: $x \in \{\mathrm{d}, \mathrm{r}, \sqcap, \sqcup, \circ, \triangleright, \smile, \sqsubseteq, *, \mathrm{t}, \mathrm{s}\}$. *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 [6]. 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 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, $\alpha_a \in Dat_a, \beta_b \in Dat_b, \rho_a \in Rel_a, \xi_b \in Rel_b$, $i(\gamma) = ic(\gamma)^{-1}$, $i(\rho_a, \xi_b) = (ic(\rho_a) \cdot ic(\rho_a))^{-1}$ and the normalizing coefficients are $Norm(\alpha_a, \beta_b) = \frac{1}{2}(\frac{1}{|V_{\alpha_a}|} + \frac{1}{|V_{\beta_b}|})$ and $Norm(\rho_a, \xi_b) = \frac{1}{2}(\frac{1}{|cR(V_{\rho_a})|} + \frac{1}{|cR(V_{\xi_b})|})$.

Table 1. The types of the semantic similarity

Similarity	$\mathcal{O}^x$	$\mathcal{V}^x = (S^x \neq \emptyset \wedge E^x = \emptyset)$	$sim(\gamma_a, \delta_b)$
Data $\alpha_a \sim_{\mathrm{d}} \beta_b$	$\alpha \simeq^i \beta \vee \mathcal{N}^{\mathrm{d}}$	$S^{\mathrm{d}} = V_{\alpha_a} \cap V_{\beta_b}$	$\lvert S^{\mathrm{d}}\rvert \cdot Norm(\alpha_a, \beta_b)$
Relation $\rho_a \sim_{\mathrm{r}} \xi_b$	$\rho \simeq^i \xi \vee \mathcal{N}^{\mathrm{r}}$	$S^{\mathrm{r}} = V_{\rho_a} \cap^r V_{\xi_b}$	$\lvert S^{\mathrm{r}}\rvert \cdot Norm(\rho_a, \xi_b)$
Transitive $\rho_a \sim_{\mathrm{t}} \xi_b$	$\rho = \xi$, $\rho \in Rel_O^{\mathrm{t}} \vee \mathcal{N}^{\mathrm{t}}$	$E^{\mathrm{t}} = V_{\rho_a} \cap^r V_{\xi_b}$, $S^{\mathrm{t}} = \{(o,p) \lvert o \in^r V_{\rho_a}, p \in^r V_{\xi_b},$ $p \in^r V_{\rho_o} \vee o \in^r V_{\rho_p}\}$	$\frac{\lvert S^{\mathrm{t}}\rvert \cdot i(\rho_a)}{\lvert cR(V_{\rho_a})\rvert \cdot \lvert cR(V_{\xi_b})\rvert}$
Symmetric $\rho_a \sim_{\mathrm{s}} \xi_b$	$\rho = \xi$, $\rho \in Rel_O^{\mathrm{s}} \vee \mathcal{N}^{\mathrm{s}}$	$E^{\mathrm{s}} = V_{\rho_a} \cap^r V_{\xi_b}$, $S^{\mathrm{s}} = \{o \lvert o \in^r V_{\rho_a} \wedge b \in^r V_{\rho_o}$ or $o \in^r V_{\xi_b} \wedge a \in^r V_{\xi_o}\}$	$\frac{\lvert S^{\mathrm{s}}\rvert \cdot i(\rho_a)}{\lvert cR(V_{\rho_a} \cup V_{\xi_b})\rvert}$
Inverse $\rho_a \sim_{\smile} \xi_b$	$\rho = \xi$, $\exists \pi \in Rel_O$: $\rho = \pi^{\smile} \vee \mathcal{N}^{\smile}$	$E^{\smile} = V_{\rho_a} \cap^r V_{\xi_b}$, $S^{\smile} = \{o \lvert o \in^r V_{\rho_a} \cup V_{\xi_b}$ and $a \in^r V_{\pi_o} \vee b \in^r V_{\pi_o}\}$	$\frac{\lvert S^{\smile}\rvert \cdot i(\rho_a)}{\lvert cR(V_{\rho_a} \cup V_{\xi_b})\rvert}$
Intersection $\rho_a \sim_{\sqcap} \xi_b$	$\exists \pi \in Rel_O$: $\pi = \rho \sqcap \xi \vee \mathcal{N}^{\sqcap}$	$S^{\sqcap} = \bigcup_{o \in c^\pi} V_{\pi_o} \cap^r V_{\rho_a} \cap^r V_{\xi_b}$	$\frac{\lvert S^{\sqcap}\rvert \cdot i(\rho_a, \xi_b)}{\lvert V_{\rho_a} \cap^r V_{\xi_b}\rvert}$
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 $\rho_a \sim_{\sqsubseteq} \xi_b$	$\rho \sqsubseteq \xi \vee \mathcal{N}^{\sqsubseteq}$	$S^{\sqsubseteq} = V_{\rho_a} \cap^r V_{\xi_b}$	$\lvert S^{\sqsubseteq}\rvert \cdot i(\rho_a) \cdot Norm(\rho_a, \xi_b)$
Closure $\rho_a \sim_{*} \xi_b$	$\rho = \xi^* \vee \mathcal{N}^*$	$S^* = V_{\rho_a} \cap^r V_{\xi_b}$	$\lvert S^*\rvert \cdot i(\rho_a) \cdot Norm(\rho_a, \xi_b)$
Refinement $\rho_a \sim_{\triangleright} \xi_b$	$\exists \pi \in Rel_O$: $\rho = \xi \triangleright \pi \vee \mathcal{N}^{\triangleright}$	$S^{\triangleright} = \{(o,p) \lvert o \in^r V_{\rho_a}$ and $p \in V_{\xi_b} \cap^r V_{\pi_o}\}$	$\frac{\lvert S^{\triangleright}\rvert \cdot i(\rho_a, \xi_b)}{\lvert cR(V_{\rho_a})\rvert \cdot \lvert cR(V_{\xi_b})\rvert}$
Composition $\rho_a \sim_{\circ} \xi_b$	$\exists \pi \in Rel_O$: $\rho = \xi \circ \pi \vee \mathcal{N}^{\circ}$	$S^{\circ} = \{o \lvert o \in c^\pi$ and $o \in^r V_{\xi_b} \wedge V_{\rho_a} \cap^r V_{\pi_o} \neq \emptyset\}$	$\frac{\lvert S^{\circ}\rvert \cdot i(\rho_a, \xi_b)}{\lvert cR(V_{\xi_b})\rvert}$

Proposition 2. Let $X \subseteq \{\mathrm{d}, \mathrm{r}, \sqcap, \sqcup, \circ, \triangleright, \smile, \sqsubseteq, *, \mathrm{t}, \mathrm{s}\}$. If for the attributes of coreferential candidates a and b the semantic similarity condition $\bigwedge_{x \in X} \mathcal{A}^x$

holds, then these coreferential 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.

Let us illustrate our introduced framework by the examples of coreferential candidates whose attributes are related by the composition, refinement, inverse and symmetric relations. The ontology's domain of our examples of i-objects is the area of Technical Documentation. We consider a technical document X on the development of some system Foo, which includes a database Bar with a query language Baz. The document describes modules Qux and Quux, and service files Grault and Garply.

In the document, we consider two mentions of a programm module: one mention concerns the interaction of a programm module with the database Bar, the other mention is about query generations in the language Baz by a programm module. For this document, our algorithm of text analysis creates the following i-objects:

$a =$ module(*name* = Qux, *parent* = Quz $\ldots \rho_a =$ interaction (database: Bar)),

$b =$ module(*name* = Qux, $\ldots \xi_b =$ query_generation (language: Baz)),

Bar = database($\ldots \pi =$ query_language (language: Baz)).

These i-objects are coreferential candidates, because they have the identical class module, the key data attribute *name* is the same, and the key relation attribute *parent* is not defined for b. The compositional similarity of these i-objects is $sim(\rho_a, \xi_b) = 1$, because the values of the attributes are consistent ($S^{\circ} \neq \emptyset$), the ontology of the document X contains the following compositional relation: query_generation = interaction $\circ$ query_language, and the inverse cardinalities of interaction and query_generation are equal to 1.

In the document, we consider also two mentions of a user: one mention states that a user uses system Foo, the other mention states that a user can read database Bar. For this document, our algorithm creates the following i-objects:

$a =$ user($\ldots \rho_a =$ use (system: Foo)),

$b =$ user($\ldots \xi_b =$ read (database: Bar)),

Bar = database($\ldots \pi_c =$ part_of (system: Foo)).

These i-objects are coreferential candidates, because they have identical class user and their key attributes are not defined. The refinement similarity of these i-objects is $sim(\rho_a, \xi_b) = 1$, because the values of the attributes are consistent ($S^{\triangleright} \neq \emptyset$), in the ontology of the document X the following refinement relation is given: use = read $\triangleright$ part_of, and the the inverse cardinalities of use and read are equal to 1.

We consider two mentions of a program module: one mention concerns sending messages to the module Qux, the other is about communication with the module Quux. The module Qux can receive messages from the module of the latter mention as written in the document X. For this document, our algorithm of text analysis creates the following i-objects:

$a =$ module ($\ldots \rho_a =$ send_message_to (module: Qux)),

b = module $(\ldots \xi_b$ = send_message_to (module: Quux)),
Qux = module $(\ldots \pi$ = receive_message_from (module: b)).
These i-objects a and b are coreferential candidates, because they have identical class module and their key attributes are not defined. The inverse similarity of these i-objects is $sim(\rho_a, \xi_b) = 1$, because the values of the attributes are consistent ($E^{\smile} = \emptyset \wedge S^{\smile} \neq \emptyset$), in the ontology of the document X the following inverse relation is given: send_message_to = (receive_message_from)$^{\smile}$, and the inverse cardinality of send_message_to is equal to 1.

In the document, we can finally consider two mentions of a service files: one mention concerns synchronization with the file Grault, the other is about the synchronization with the file Garply. The file Grault must be synchronized with the file of the latter mention as written in the document X. For this document, our algorithm of text analysis creates the following i-objects:
a = file($\ldots \rho_a$ = synchronization_with (file: Grault)),
b = file($\ldots \xi_b$ = synchronization_with (file: Garply)),
Grault = file($\ldots \pi$ = synchronization_with (file: b)).
These i-objects a and b are coreferential candidates, because they have identical class file and their key attributes are not defined. The symmetry similarity of these i-objects is $sim(\rho_a, \xi_b) = \frac{1}{3}$, because the values of the attributes are consistent ($E^{\mathrm{s}} = \emptyset \wedge S^{\mathrm{s}} \neq \emptyset$), in the ontology of the document X the relation synchronization_with is stated to be symmetric, and the inverse cardinality of send_message_to is equal to 3.

4 Conclusion

In this paper, we suggest a novel formal framework that allows to use a number of properties of ontology classes and relations in the approach to coreference resolution in the process of ontology population. These properties include class and attribute hierarchy, intersection, union, composition, refinement, inverse, inclusion, reflexive-transitive closure, transitivity, and symmetry. We show how they can be efficiently used in the evaluation of the semantic similarity of coreferential candidates. This evaluation is integrated into a single estimation of coreferential similarity. Using these properties give more precise and complete coreference identification due to taking into account more similarity factors than just the elements' identity.

In the nearest future 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 propose to specify them for n-ary ontology relations which represent situations and events of the real world. All these additional properties will 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.

References

1. **Elango P.** *Coreference Resolution: A Survey* // Technical Report, UW-Madison, 2006, available at: https://ccc.inaoep.mx/ villasen/index_archivos/cursoTATII/EntidadesNombradas/Elango-SurveyCoreferenceResolution.pdf
2. **N. O. Garanina, E. V. Bodin.** *Distributed Termination Detection by Counting Agent* // Proc. of the 23nd International Workshop on Concurrency, Specification and Programming (CS&P 2014), Chemnitz, Germany, 29. September - 01. Oktober 2014. Humboldt-Universitat zu Berlin, 2014, pp. 69–79.
3. **Garanina N., Sidorova E., Bodin E.** *A Multi-agent Text Analysis Based on Ontology of Subject Domain* // In: Perspectives of System Informatics. LNCS Vol .8974, 2015, pp. 102-110.
4. **Garanina N., Sidorova E.** *Context-dependent Lexical and Syntactic Disambiguation in Ontology Population* // Proc. of the 25th International Workshop on CS&P. Rostock, Germany, Sep. 28-30, 2015. – Humboldt-Universitat zu Berlin, 2016, pp. 101–112.
5. **Garanina N., Sidorova E., and Kononenko I.** *A Distributed Approach to Coreference Resolution in Multiagent Text Analysis for Ontology Population* // To appear in Perspectives of System Informatics 2017, the series Lecture Notes in Computer Science, 2018.
6. *Handbook on Ontologies.* // Edt.: Staab S., Studer R., International Handbooks on Information Systems, Springer Berlin Heidelberg, 2009, 808 p.
7. **Hladky D., Ehrlich C., Efimenko I., Vorobyov V.** *Discover Shadow Groups from the Dark Web* // 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.
8. **Kononenko I.S., Sidorova E.A.** *Language Resources in Ontology-Driven Information Systems* // First Russia and Pacific Conference on Computer Technology and Applications, 6-9 September, 2010, Vladivostok, Russia. -P.18-23.
9. **Motta E., Siqueira S., Andreatta A.** *An Unsupervised Rule-Based Method to Populate Ontologies from Text* // Proc. of International Conference on Web Information Systems and Technologies WEBIST 2009: Web Information Systems and Technologies, pp 157-169
10. **Mitkov R.** *Anaphora resolution* // Mitkov Ruslan (ed.) The Oxford handbook of computational linguistics, ch.14, pp.266-283, N.Y.: Oxford university press, 2003.
11. **Petasis G., Karkaletsis V., Paliouras G., Krithara A., and Zavitsanos E.** *Ontology Population and Enrichment: State of the Art.* // In Knowledge-driven multimedia information extraction and ontology evolution, LNAI 6050, Springer-Verlag Berlin, pp. 134-166, 2011
12. **Prokofyev R., Tonon A., Luggen M., Vouilloz L., Difallah D.E., and Cudre-Mauroux P.** *SANAPHOR: Ontology-Based Coreference Resolution* // In Lecture Notes in Computer Science, Volume 9366, 14th International Semantic Web Conference, 2015, Proceedings, Part I, pp. 458-473. 2015
13. **Sidorova E.A., Kononenko I.S.** Representation and use of the jenre structure of documentation in text processin. // Proc. of the Science-Intensive Software (SIS-09) - PSI 2009 Satellite Workshop, Novosibirsk, Russia, June, 2009. – Novosibirsk: Siberian Science Publisher, 2009, pp. 248–254. (In Russian)
14. **Yatskevich M., Welty C., and Murdock J.W.** *Coreference resolution on RDF Graphs generated from Information Extraction: first results.* // Proc. of ISWC'06 Workshop on Web Content Mining with Human Language Technologies. 2006.