Logical-Ontological Approach to Coreference Resolution

Elena Sidorova[1], Natalya Garanina[1], Irina Kononenko[1], and Alexey Sery[1]

[1] A.P. Ershov Institute of Informatics Systems SB RAS, 6, Acad. Lavrentjev pr., Novosibirsk 630090, Russia
{lsidorova,garanina,alexey.seryj}@iis.nsk.su
irina_k@cn.ru

Abstract. We suggest a logical-ontological approach to the coreference resolution in the process of text analysis and information extraction. Our approach solves the problem of comparing objects found in the text – instances of ontology classes — using the evaluation of the similarity of attributes and relations of objects. In object comparison, we take into account the discourse factors associated with the text and the extra-textual characteristics presented in the ontology of the subject domain. Particularly, we consider polyadic relations which may represent the situations found in the text (events, processes, actions). We propose the ontological interpretation of polyadic relations as classes with single-valued object properties. For coreference resolution we use information about objects and their relations. We propose the corresponding measures for evaluating the semantic similarity of the participant objects in the relations.

Keywords: ontology population, text analysis, information extraction, coreference resolution, referential factors, polyadic relations.

1 Introduction

Identification of referential relations in discourse is one of the most vital but difficult for modeling problems of automatic text analysis. Reference is a relation between some text unit (language expression) and non-linguistic object, which is called a referent. Correct interpretation of an utterance in the text under analysis involves identification of the object mention referent, i.e. reference resolution. There is a range of language means to mention certain referent in the text, and a speaker (text author) makes choice between two opposite types of language expressions: full noun phrases (proper names and descriptions) and reduced means of reference (pronouns and anaphoric zeroes). Processing expressions of the first type requires direct comparison of extracted objects. In the second case, an anaphoric relation of the reduced expression to antecedent expression is detected with respect to a number of text-structure, syntactic, semantic and pragmatic conditions.

The anaphora and coreference resolution is an important task within the framework of automatic discourse analysis: machine translation, text summarization and information extraction. The latter can be performed by natural language processing in which certain types of information must be recognized and extracted from the text (named entities recognition and fact extraction tasks, in particular). We consider the coreference resolution within the framework of information extraction for ontology population. In this framework, an ontology is used to represent the results of information extraction, and knowledge presented in the ontology helps to solve specific information extraction tasks.

Solving the task of automatic ontology population involves addition of information to the ontology repository. In [1] we consider mentions of simple entities and propose an approach to their coreference resolution in the process of information extraction for ontology population. An ontology structure allows to take into account implicit information in the input text due to detecting relations between objects. In this paper, we suggest coreference resolution for new objects with a complex structure including situations (events, actions, processes), which are represented by polyadic relations in an ontology. These situations extend the domain knowledge used for solving coreference resolution problem. The new knowledge improves the quality of coreference resolution.

In Section 2 we give a brief review of modern trends in the coreference problem definition and the present research. In Section 3 we describe our basic approach to ontology-based information extraction with formal definitions of and ontology and polyadic relations. Section 4 presents ontological factors relevant for coreference resolution illustrated by text examples and revises the similarity measure of objects. In Section 5 we consider features of experiments in our approach. We conclude with the base characteristics and advantages of the proposed approach and outline the directions for future research.

2 Coreference in Information Extraction Tasks

We observe several classification aspects of problems related to the reference identification.

- First aspect is the way of presenting references in the text: full lexical expressions (noun phrases – proper names, descriptions, descriptions combined with proper names) or reduced expressions using anaphoric means (pronouns, determiners) or anaphoric zero. In the first case, for noun phrases based on proper names, the problem is detecting identical references to named entities. In the second case, the problem is identification of the antecedent, i.e. anaphora resolution [2, 3].
- Second aspect is the type of the referenced object: referential identity of entities or situations (events).
- Third aspect is the search area and type of context: the context of a single document (simple and complex sentences or chains of sentences in one text) opposes to cross-document analysis, in which references to the same object are looked for in the corpus or document flow.

The traditional problem of anaphora and coreference resolution within a coherent text remains to be relevant. Many early and modern researches solve the problem using linguistic methods based on rules and methods of machine learning. R. Mitkov's reviews [4, 5] and later [6, 7] consider the basic approaches to this problem. Recently, there has been a growing interest in solving the problem in a broader perspective: not only entities but also events or situations have been considered [8 – 12]. A cross-document reference analysis that is an important approach for populating knowledge bases and ontologies is used for the problem as well [8, 13 – 15]. The complexity of the problem of coreference resolution requires an integrated approach, involving both knowledge about the structure of the text (the level of discourse) and knowledge about the subject area, which are determined by the classes of entities in a specific ontology and their ontological structure (ontological level). In [16] the authors consciously abstract away from the discourse factors of coreference in order to investigate the role of subject knowledge. Discourse features represent the structural and textual properties of mentions (similarity of sub-chains, position, distance), grammatical and lexical features. Obviously, new tasks require a revision of the role of discourse features in comparison with ontological ones. Thus, cross-document analysis does not consider pronominal anaphora and hardly takes into account such discourse factors as the order of appearance of mentions in the text, and the distance (linear or rhetorical).

Theories of discourse analysis distinguish several types of discourse connectivity: referential (identity of participants), spatial, temporal and event-triggered ones [17]. In applied research, there are two approaches to understanding the coreference of events. In the first approach, two mentions of an event are considered coreferent if they are characterized by the same set of properties (such as time or place of the event) and the same set of participants [9 – 11]. In the second approach, only the referential identity of participants is considered for referential identity of events [3]). In [12] a broader set of referential relations between two mentions of events is considered: complete coreference, subevents for vertices of the parent and child layers, subevents for a descendant vertex of a single layer.

We consider the problem of information extraction as a task of detecting all references to objects of a given domain: entities and situations (events, states, actions, processes). In the ontology population task, the found objects should be represented as instances of concepts and relations of the ontology. It is necessary to establish referential relations between all instances found in the process of text analysis and instances of the ontology information content (which does not exclude the possibility of adding new instances to the ontology).

3 The Model of Information Extraction

Consider the environment in which our approach to coreference resolution is being developed. Fig. 1 shows the general scheme of the information extraction system (IE-system) with the emphasized module of coreference resolution.

The input of our IE-system comprises: the ontology of a subject domain, the ontology population rules and the results of preliminary text processing including the terminological, thematic, and segment coverings of an input text.

A terminological covering is the result of lexical text analysis which extracts terms of a subject domain from a text and forms lexical objects using semantic vocabularies. A segment text covering is a division of the text into formal fragments (clauses, sentences, paragraphs, headlines, etc.) and genre fragments (document title, annotation, glossary, etc.). A thematic covering selects text fragments of a particular topic. A construction of a thematic covering is based on the thematic classification methods.

The module of information extraction constructs objects representing instances of concepts and relations of the domain ontology from the lexical objects [18]. This module uses the ontology population rules which are automatically generated from fact schemes. The fact schemes are formulated by experts taking into account the ontology and language of a subject domain. These fact schemes constrain morphological, syntactic, structural, lexical, and semantic characteristics of the objects.

The coreference resolution module [19] runs in parallel with the information extraction module. This module forms hypotheses about coreference relations, and calculates their weights using various factors discussed below.

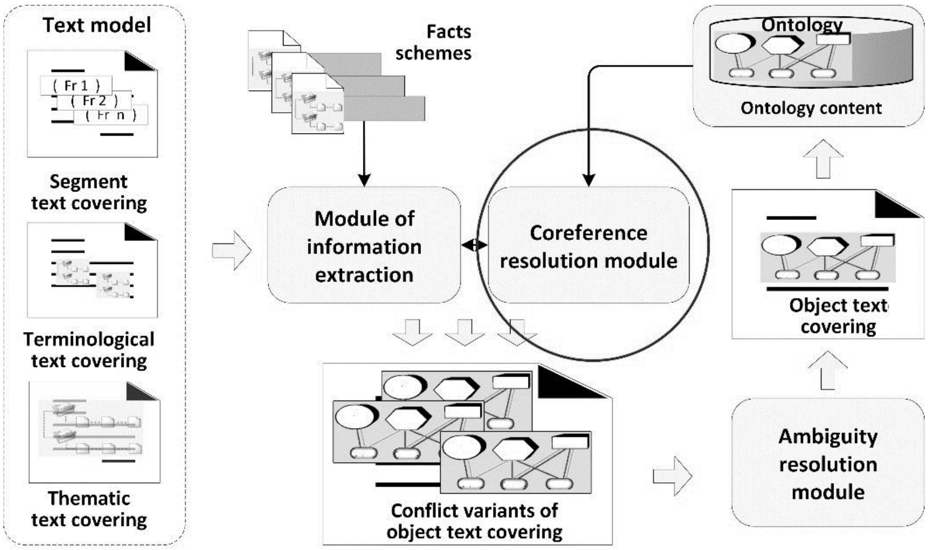

Fig. 1. The scheme of the system of information extraction and ontology population.

The ambiguity resolution module resolves all types of conflicts which are the result of various interpretations of the input text — different object text coverings for the same text fragment. This module chooses the most informative variant from the set of possible interpretations (the variant with the highest weight) [20].

The result of the work of our IE-system is the population of ontology content by instances of concepts and relations of the subject domain found in the input text.

3.1 The Ontology of a Subject Domain

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* or *datatype properties* in Dat_O) or has values as instances of some classes (*object attributes* or *object properties* in Rel_O, which model binary relations).

Each class $c \in C_O$ is defined by the set of its 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 object attribute $\rho \in Rel_c \subseteq Rel_O$ has values from the subset $C_\rho \subseteq C_O$. The set of all class attributes is denoted by $Atr_c = Dat_c \cup Rel_c$. 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 $\forall c_1, c_2 \in C_O : Atr_{c1} \neq Atr_{c2}$.

We denote the class of an attribute γ by c^γ and the set of its values by D^γ. A set of attributes of every class must include the nonempty set of *key attributes* Atr_c^K. The key attributes can either be data or object attributes. These attributes guarantee unambiguous definition and uniqueness of the class instances.

A tuple $a = (c_a, Dat_a, Rel_a)$ is *an instance of the class* $c_a = (Dat_{c_a}, \mathrm{Rel}_{c_a})$ $(a \in c_a)$ iff every data attribute $\alpha_a \in Dat_a$ has a name $\alpha \in Dat_{c_a}$ with the values V_{α_a} from V_{d_α} and every object attribute $\rho_a \in \mathrm{Rel}_a$ has a name $\rho \in \mathrm{Rel}_{c_a}$ with the values V_{ρ_a} as instances of the classes from C_ρ.

We use the standard *class inheritance relation*: the class c_2 *is a subclass* of the class c_1 ($c_1 < c_2$) iff $\forall a \in c_2$: $a \in c_1$.

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 information content for a given ontology from the given input data.

3.2 Polyadic Relations

The notion of polyadic relation is not considered in the classical ontology theory. For example, the OWL – the standard ontology description language – has no language constructions for polyadic relations, only binary relations (Object Property) are available. On the other hand, polyadic relations frequently arise in the tasks of extracting information from texts, because they can describe the propositional content of a statement that represents an extra-linguistic situation, or state of affairs (event, action, process, etc.).

To overcome these shortcomings, we model *polyadic relations* (or just relations) by ontology classes with constraints on the set of attributes. First, relations classes have to include at least two object properties. Second, every object property of a relation has to be a key attribute. A polyadic relation may also contain datatype properties without special constraints.

Due to this definition, a polyadic relation is naturally represented by the set of binary relations. And vice versa, a binary relation can be represented by the polyadic relation with two object properties as a special case of polyadic relations.

In text processing, we consider polyadic relations correspond to descriptions of situations (actions, processes) and other objects with complex structure. The following Table 1 gives some examples of polyadic relations extracted from texts.

These examples relate to the automated control systems subject domain that includes such relation classes as *Action*, *Process*, *Function*, *Control*, *Movement*, *Change_of_state*, etc. Object properties of relation classes correspond to the hierarchy of semantic roles. The semantic role is a generalization of the functions of a participant in a range of situations denoted by a group of predicates, and hence the types of corresponding situations.

Table 1. Examples of polyadic relations.

S1 Action	**Type: information_transfer** **Sender: X** **Recipient: Y** **Message: Z** **Content: null**	The system (Y) receives commands (Z) from the operator (X)
S2 Process	**Agent: X2** **Type: processing** **Message: Z**	The command (Z) is entered by the operator (X2) through the remote operator console

3.3 The Coreference Resolution Problem

The information content of a text consists of a set of instances of ontology classes and relations found in the text, which are provided with additional information.

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

- $c_a \in C_O$ is the ontology class;
- Dat_a is the set of data attributes $\alpha_a = (\alpha, V_{\alpha_a})$, where
  - $\alpha \in Dat_{c_a}$ is the attribute name, and V_{α_a} is the set of values $v \in d_\alpha$;
- Rel_a is the set of object attributes $\rho_a = (\rho, V_{\rho_a})$, where
  - $\rho \in \mathrm{Rel}_{c_a}$ is the attribute name, and V_{ρ_a} is the set of i-objects of a class $c_{\rho_a} \in C_{\rho_a}$;
- G_a is the grammar information (morphological and syntactic features based on grammar features of lexical object);
- P_a is the structural information (a set of positions in the input data and the formal segments).

The attribute γ of the i-object *a* is *filled* if $V_{\gamma_a} \neq \varnothing$. We denote by $Atr_a = Dat_a \cup Rel_a$ the set of all attributes. Each 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}.

We assume that i-objects *a* and *b* are *possible coreferents* $a \approx b$ (*candidates for coreference*) iff their classes are transitively related by the class inheritance relation and the set of values of all filled key attributes of one i-object is included in the set of values of the corresponding key attributes of the other i-object.

The coreference resolution problem is to detect if given candidates for coreference correspond to the same ontology instance.

4 Referential Factors

In previous papers [19], we considered two types of factors that affect the evaluation of the measure of the coreferential similarity of two objects. First, discourse factors (local textual and contextual) are determined by the language means used to represent the objects in the text and by their location in the text structure. Second, semantic factors determine the similarities of objects with respect to their ontological structure and relations.

In our approach, we distinguish logical-ontological factors for considering a set of associated relations between objects. For these factors we use the properties of relations specified in the ontology.

All these factors are used to evaluate similarity of objects mentioned in the text. For each factor, we define a similarity measure. This measure corresponds to the degree of strength of the coreferent relation between the i-objects *a* and *b* with respect to the factor, without taking into account other factors.

4.1 The Coreferential Conflict and the Similarity Measure

We define *coreferential conflict* as a case when two non-coreferent i-objects *a* and *b* are possible coreferents of the third i-object *c*: $a \leftrightsquigarrow^{c} b \Leftrightarrow (a \approx c) \wedge (b \approx c) \wedge \neg(a \approx b)$.

To determine which of these i-objects are actually coreferent, we use the measure of coreference similarity of i-objects. This measure for i-objects *a* and *b* is denoted as *cs(a,b)*. If the non-coreferential i-objects *a* and *b* are possible coreferents for the i-object *c*, we say that *the coreferential conflict is resolved to a* iff $cs(a,c) > cs(b,c)$, i.e. the i-object *a* is more similar to i-object *c*, then i-object *b*.

The integral measure of similarity *cs(a,b)* is calculated as an Euclidean measure of similarity based on four measures – semantic *S(a,b)*, context *C(a,b)*, position *P(a,b)* and grammar *G(a,b)*.

$$cs(a,b) = \frac{1}{2}\sqrt{(1-S(a,b))^2 + (1-C(a,b))^2 + (1-P(a,b))^2 + (1-G(a,b))^2} \qquad (1)$$

The context similarity measure 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 a) attribute values from both *a* and *b*, and b) attribute values borrowed by *a* from *b*, and by *b* from *a*, for the evaluation of their own attributes.

The position similarity measure P(a,b) takes into account variants of location of i-objects in an input text. This measure depends on the number of segments, number of possible candidates in the conflict, and number of lexemes placed between the positions of *a* and *b*.

The grammar similarity measure G(a,b) is based on the standard linguistic features such as gender, number, person, etc.

The semantic similarity measure S(a,b) determines the degree of proximity of the corresponding attribute sets Atr_a and Atr_b. Comparing these two sets takes into account both the similarity of the values of their constituent elements and additional characteristics based on the ontological properties of attributes, including the inheritance of classes and data attributes, intersection, union, composition, refinement, inversion, inclusion, closure, transitivity and symmetry.

In [1] we consider 11 types of similarities. Below we expand this set with similarities using polyadic relations. Initially, *S(a,b)* was determined by formula (2), where $Sim_b^a = \{(\alpha_a, \beta_b) \mid sim(\alpha_a, \beta_b) \neq 0\}$:

$$S(a,b) = \frac{1}{|Sim_b^a|} \sum_{(\alpha_a, \beta_b) \in Sim_b^a} sim(\alpha_a, \beta_b) \tag{2}$$

Here, under the sign of the sum, all kinds of similarities of the attributes of the objects *a* and *b* are collected. Practical considerations and experimental data revealed particular cases in which basic formula (2) is inexact and instable with respect to adding new attribute comparison characteristics: i-objects that have a large set of comparable but actually not similar attributes can turn out to be close with each other due to just taking into account that the similarity of attributes that is greater than zero. It is worth noting that such cases are very rare due to the definition of coreference and the formulation of the problem of extracting i-objects. The second disadvantage of formula (2) is expressed by the fact that adding new terms to the sum can decrease the total value. But one should expect that positive additional information about the proximity of attributes have to always increase the similarity of the corresponding i-objects. These additional characteristics are based on the ontological properties of attributes, including, in particular, composition, transitivity, refinement, etc., and specials properties of polyadic relations described below. In view of the above, it was proposed to convert formula (2) to a formula of the following form:

$$S(a,b) = S^{EQ} + (1 - S^{EQ}) \cdot S^{\Delta} \tag{3}$$

The value $S^{EQ} \in [0;1]$ corresponds to the similarity of the values of the corresponding attributes of the objects *a* and *b* without taking into account the additional characteristics, and $S^{\Delta} \in [0;1)$ — the additional information provided by these characteristics.

S^{EQ} is calculated by formula (4), similar to formula (2), where the set of pairs of similar attributes Sim_b^a is replaced by the set of pairs of comparable attributes $Comp_b^a = \{(\alpha_a, \beta_b) \mid \alpha_a \in Atr_a, \beta_b \in Atr_b, \alpha = \beta\}$.

$$S^{EQ} = \frac{1}{|Comp_b^a|} \sum_{(\alpha_a, \beta_b) \in Comp_b^a} sim(\alpha_a, \beta_b) \tag{4}$$

Only measures of standard similarity of attributes by values stand under the sign of the sum in the formula (4) [19].

Let the total amount of additional information about the attributes of objects *a* and *b* be

$$I = \sum_{\gamma_a \in Attr_a, \delta_b \in Attr_b} sim^{\Delta}(\gamma_a, \delta_b) \tag{5}$$

Here the symbol Δ denotes additional properties of attributes, such as transitivity, composition, etc. It is obvious that *I* can take any positive values. Hence, in order to get the value of S^{Δ} varying from 0 to 1, we need a monotonic transformation defined everywhere on the positive semi-axis. Using *I*, we evaluate the additional similarity of the i-objects *a* and *b*. Really, we determine the value of the probability of this similarity S^{Δ}:

$$S^{\Delta} = \frac{I}{1+I} \tag{6}$$

We can see from formulas (3), (5) and (6) that

- $S(a,b) = 1 \Leftrightarrow S^{EQ} = 1$,
- $S^{\Delta} \in [0;1)$, and
- $S(a,b) > S^{EQ} \Leftrightarrow S^{EQ} < 1 \wedge S^{\Delta} > 0$.

In other words, when objects have incomplete similarity in the values of comparable attributes, and the additional information is available, the degree of similarity *S* is always greater than S^{EQ}, but full match is achieved only under the condition that the values of all comparable attributes are the same taking coreference into account.

4.2 Relations Factor

For evaluating similarity we consider polyadic relations in the following two aspects.

First, comparing polyadic relation instances for identification coreference between them.

Example 1. *When the bottle reaches a certain position, (the sensor^X communicates with the conveyor ^{Y})^{S1} to inform it that it should stop. For this purpose (the sensor^X sends a signal Stop^Z to the receiving device of the conveyor^Y)^{S2}*

In this example, we can distinguish two possible coreferent instances of polyadic relations *S1* and *S2*:

- S1: Contact (Originator: X, Recipient: Y)
- S2: Information_transfer (Originator: X, Recipient: Y, Content: Z)

These instances are similar because their *Originator* and *Recipient* attributes have coreferent values.

Second, using information about polyadic relations for identification coreference between i-objects participating in these relations. For this purpose, pairs of relations are considered that contain similar values (besides the objects themselves being compared). Change the example from the previous version.

<u>Example 2</u>. *(The sensor^{X1} transmits a message^{Z} to the conveyor^{Y})^{S1} to inform it that the bottle has reached a certain position. So, (it^{X2} controls the operation of the conveyor^{Y})^{S2}.*

In this example polyadic relations are represented by the following instances:

- S1: Information_transfer (Originator: X1, Recipient: Y, Content: Z)
- S2: Control (Controller: X2, Patent: Y)

We consider the instances X1 and X2 are similar because S1 and S2 have a similar value Y. Note that in the last example the relations of different classes with different sets of object attributes are compared because we allow the comparison of arbitrary relations.

We define the following formal ontological properties for object attributes. They are used for definition of object similarity measures that take into account polyadic relations. We borrow some concepts of relational algebra. We denote the set of all polyadic relations of the ontology *O* by S_O.

Definition 1. Let $\rho, \rho', \rho'' \in Rel_O$.

- The attributes ρ, ρ′ are in the projection relation $\rho =_{\pi} \rho'$ iff $C_\rho, C_{\rho'} \subseteq S_O$ and $\exists \bar{A}=(\gamma_1,\ldots,\gamma_m), \bar{A}'=(\gamma'_1,\ldots,\gamma'_m): \forall a \in c \in C_\rho \; \exists a' \in C_{\rho'}: \pi_{\bar{A}}(a) = \pi_{\bar{A}'}(a')$, i.e. $V_{\gamma i a} = V_{\gamma' i a'}$ ($i \in [1..m]$), and vice versa, $\forall a' \in c' \in C_{\rho'} \; \exists a \in C_\rho: \pi_{\bar{A}'}(a') = \pi_{\bar{A}}(a)$, i.e. the values of the attributes that are in the projection relation are instances of the polyadic relations that contain equal values.
- The attributes ρ, ρ′ and ρ″ are in the natural join relation $\rho = \rho' \bowtie \rho''$ iff $C_\rho, C_{\rho'}, C_{\rho''} \subseteq S_O$ and $\forall a' \in c' \in C_{\rho'}, \exists a \in c \in C_\rho, A \subseteq Atr_a: \pi_{Atra'}(a') = \pi_A(a), \forall a'' \in c'' \in C_{\rho''} \; \exists a \in c \in C_\rho, A \subseteq Atr_a: \pi_{Atra''}(a'') = \pi_A(a)$, and $\forall a \in c \in C_\rho, b \in Atr_a : (\exists a' \in c' \in C_{\rho'}, b' \in Atr_{a'} : b = b') \vee (\exists a'' \in c'' \in C_{\rho''}, b'' \in Atr_{a''} : b = b'')$, i.e. the instances that are the values of the object attributes ρ′ и ρ″ are complementary different views (projections) on the values of the attribute ρ.

Thus, the projection describes a subset of the common elements of the relation instances. In Example 1, the common projection of instances of the relations *S1* and *S2* is {*X, Y*}. In Example 2, the corresponding projection is the set {*Y, X1, X2*}. The natural join takes into account the presence of a third relation when comparing a pair of relation instances. This relation includes the join of the attributes of these relations.

The presence of such a third relation is an evidence of the information included in the first two ones.

The example of the ontological natural join relation is ontological description of the modules of a technological complex that execute the similar tasks. Each module is represented by a relation, including instances of the tasks: S_{Mi} $(w_1, \ldots, w_n)$. The complex performs the whole set of tasks, which is the result of the natural join of the tasks executed by the modules: $\cup\, w_{ij}$, $w_{ij} \in S_{Mi}$.

For those cases when properties of attributes in Definition 1 cannot be derived from the ontology description, there is a need to check the necessary conditions of the presence of the properties. The following proposition formulates these conditions in a constructive way. We denote the necessary condition of a property x by $\mathcal{N}^x$.

Proposition 1. Let $\rho, \rho', \rho'' \in Rel_O$.

- $\rho =_\pi \rho' \Rightarrow \mathcal{N}^\pi = (C_\rho \cap^i C_{\rho'} \neq \emptyset)$;
- $\rho = \rho' \bowtie \rho'' \Rightarrow \mathcal{N}^\bowtie = (C_{\rho'} \cup^i C_{\rho''} \subseteq^i C_\rho)$.

Here, the superscript i in the set operations means that we make the operation over the elements of the sets and over their parental classes and subclasses in the class hierarchy. The proof follows from Definition 1.

Taking into account Definition 1, we define the projection and natural join based similarities of the attributes. We also define the class similarity. In the following definition, the superscript r in comparison operations and calculation of the power of sets means that the operations consider the elements of the sets and their possible coreferents.

Definition 2. For i-objects a and b with $a \approx b$ and $c_a \leq c_b$, we compute *the power of the class similarity* as $sim^c(c_a, c_b) = |c_b|/|c_a|$, where $|c_x|$ is the number of subclasses of the class x including x itself.

Definition 3. For i-objects a and b, we consider object relation $\rho \in Rel_a$ and $\xi \in Rel_b$ with $\rho, \xi \in S_O$ is

- *projectionally similar* $\rho \sim_\pi \xi$, iff $\rho =_\pi \xi \vee \mathcal{N}^\pi$ and $S^\pi = \cup_{x \in V_{\rho a}} \{X \subseteq Atr_x \mid \exists\, y \in V_{\xi b}, Y \subseteq Atr_y : \pi_X(x) =^r \pi_Y(y)\} \neq \emptyset$. The power of the projection similarity is $sim^\pi(\rho, \xi) = ½|S^\pi|(c(V_{\rho a})^{-1} + c(V_{\xi b})^{-1})$, where $c(V_\mu) = \sum_{z \in V_\mu} \sum_{\gamma \in Atrz} |V_\gamma|^r$.
- *joinly similar* $\rho \sim_\bowtie \xi$, iff $\exists\, \mu$: $\mu = \rho \bowtie \xi \vee \mathcal{N}^\bowtie$ and $S^\bowtie = \{(x, y) \mid x \in V_{\rho a}, y \in V_{\xi b}, \exists\, z \in C^\mu, Z_x \subseteq Atr_z, Z_y \subseteq Atr_z: Atr_z \subseteq Z_x \cup Z_y, \pi_{Atrx}(x) =^r \pi_{Zx}(z) \text{ and } \pi_{Atry}(y) =^r \pi_{Zx}(z)\} \neq \emptyset$. The power of the join similarity is $sim^\bowtie(\rho, \xi) = ½|S^\bowtie|((|V_{\rho a}|^r)^{-1} + (|V_{\xi b}|^r)^{-1})$.

Thus, we can take into account the power of sim^c, sim^π and $sim^\bowtie$ of the projection and join similarity in the semantic similarity measure along with the other factors in formula (5). This allows us to take the context into account more accurately, improving the quality of information extraction.

5 Characteristic of Experimental Study

The proposed approach to resolving coreference is based on the properties of the domain concepts presented formally. Testing its implementation requires for a formally presented ontology of a subject domain, as well as text corpus annotated in accordance with the ontology. Typed coreferential relations also have to be annotated.

There exist coreferentially annotated corpora for English (MUC) and a number of other languages (Catalan, Dutch, English, German, Italian, Spanish, Czech, Chinese and Arabic). The first open corpus for Russian is RuCor (available at http://rucoref.maimbava.net/) that represents anaphorical and coreferential relations and morphological annotation. RuCor contains about 200 texts of different genres (primarily news, essays, and fiction) that do not correspond to any special subject domain [21]. The lack of appropriate datasets with deep layers of annotation is the obstacle to the study of complex cases of coreference.

Hence, for evaluation of our approach we form a corpus of examples with a complex type of coreference, which can be resolved on the basis of ontology. Several examples are selected for each type of ontological relation. The total volume of the corpus is about 50 text fragments taken from texts of technical documentation and encyclopedias. These fragments represent specifications of requirements from the subject domain of automated control systems. Each example is annotated by coreference relations with types based on ontological properties.

We consider such annotation of coreference information necessary for further linguistic research. Extending the capabilities of automatic analyzers with computational similarity models based on ontological properties improves the quality of coreference resolution. Thus, for the examples found, the use of logical-ontological measures allows to increase the measure of similarity of the “correct” variant by 0.05-0.1 (5-10%).

6 Conclusion

In the papers on the topic of coreference resolution, we proposed a formal statement of the problem and mathematically-strict definitions of the notions of coreference, coreferential conflict and ontological properties used to resolve the coreference. This is an important contribution to ensure the correct operation and improve the quality of the coreference resolution algorithms.

The main features of the proposed approach to coreference resolution are:

1. shift of the emphasis from discourse factors to the subject knowledge, primarily to the ontology of the subject domain to be populated through information extraction, disambiguation, and coreference resolution;
2. integration of computational and linguistic models and techniques of text analysis at the phase of semantic processing. Thus, weighted coreferential relations between objects are used for coreference resolution. In this process, the hypothetic coreferential relations are generated by the linguistic model, and the resolution (choice of the best hypothesis) is based on the statistical data;
3. scalability of the solution. Our approach can be enriched with new information extraction rules and referential factors.

The corpus with annotated coreference is necessary for studying different cases of repeated mentions of events that need ontological information about polyadic relations to correctly resolve coreferences. Our future research will focus on general classification of such cases. We plan to develop special case-oriented coreference resolution techniques, particularly, by considering the relevance of ontological properties for the evaluation of similarity of possible coreferents. Taking this into account, we are faced with the problem of defining ontology formal properties that provide a better solution to the tasks of extracting information from the text and, in particular, the resolution of the coreference.

Acknowledgement. The study was supported by the Russian Foundation for Basic Research, project 17-07-01600.

References

1. Garanina, N., Sidorova, E., Kononenko, I., Gorlatch, S.: Using Multiple Semantic Measures For Coreference Resolution. Ontology Population. International Journal of Computing 16(3), 166–176 (2017).
2. Dimitrov, M., Bontcheva, K., Cunningham, H., Maynard, D.: A Light-weight Approach to Coreference Resolution for Named Entities in Text. In: Branco, A., McEnery, T., Mitkov, R. (eds.) Anaphora Processing: Linguistic, Cognitive and Computational Modelling, vol. 263, pp. 97-112. John Benjamins Publ., (2005).
3. Sobha, L.: Anaphora Resolution Using Named Entity and Ontology. In: Johansson, C. (ed.) Proceedings of the Second Workshop on Anaphora Resolution (WAR II), NEALT Proceedings Series, vol. 2, pp.91-96 (2008).
4. Mitkov, R.; Anaphora resolution: the state of the art. In: Working paper based on the COLING'98/ACL'98 tutorial on anaphora resolution. Wolverhampton (1999).
5. Mitkov, R.: Anaphora resolution. In: Mitkov, R. (ed.) The Oxford handbook of computational linguistics, ch.14, pp. 266-283. Oxford university press, N.Y. (2003), https://pdfs.semanticscholar.org/e782/00b1e3ba2a72de1ca9b9b2c5efa775151bfa.pdf, last accessed 2018/10/04.
6. Elango, P.: Coreference Resolution: A Survey: Technical Report. UW-Madison (2006), https://ccc.inaoep.mx/~villasen/index_archivos/cursoTATII/Entidades Nombradas/Elango-SurveyCoreferenceResolution.pdf, last accessed 2018/04/01.
7. Prokofyev, R., Tonon, A., Luggen, M., Vouilloz, L., Difallah, D.E., Cudr´e-Mauroux, P.: SANAPHOR: Ontology-Based Coreference Resolution. In: 14th International Semantic Web Conference, part I, LNCS, vol. 9366, pp. 458-473. Springer, Cham (2015).
8. Lee, H., Recasens, M., Chang, A., Surdeanu, M., Jurafsky D.: Joint Entity and Event Coreference Resolution across Documents. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language, EMNLP-CoNLL 2012, pp. 489–500 (2012).
9. Cybulska, A., Vossen, P.: "Bag of Events" Approach to Event Coreference Resolution. Supervised Classification of Event Templates. International Journal of Computational Linguistics and Applications 6(2), 11-27 (2015).
10. Borgo, S., Bozzato, L., Aprosio, A.P., Rospocher, M., Serafini L.: On Coreferring Text-extracted Event Descriptions with the aid of Ontological Reasoning. Technical Report (2016), https://arxiv.org/pdf/1612.00227.pdf, last accessed 2018/10/04.

11. Bejan, C.A., Harabagiu, S.: Unsupervised event coreference resolution with rich linguistic features. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp.1412-1422 (2010).
12. Araki, J., Liu, Z., Hovy, E., Mitamura, T.: Detecting Subevent Structure for Event Coreference Resolution. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), pp. 4553–4558 (2014).
13. Mayfield, J., Alexander, D., Dorr, B.J., Eisner, J., Elsayed, T., Finin, T., Fink, C., Freedman, M., Garera, N., McNamee, P., Mohammad, S., Oard, D., Piatko, C., Sayeed, A.B., Syed, Z., Weischedel, R.M., Xu, T., Yarowsky, D.: Cross-Document Coreference Resolution: A Key Technology for Learning by Reading Association for the Advancement of Artificial Intelligence. In: AAAI Spring Symposium: Learning by Reading and Learning to Read, pp.65-70 (2009).
14. Yatskevich M., Welty C., Murdock J.W. Coreference resolution on RDF Graphs generated from Information Extraction: first results. ISWC'06 Workshop on Web Content Mining with Human Language Technologies (2006).
15. Hladky, D., Ehrlich, C., Efimenko, I., Vorobyov V.: Discover Shadow Groups from the Dark Web. In: Web Intelligence and Security: Advances in Data and Text Mining Techniques for Detecting and Preventing Terrorist Activities on the Web, pp. 67-81 (2010).
16. Suleymanova, E., Trofimov, I.: A method for coreference resolution within information extraction. In: Program Systems: Theory and Applications 1(15), 15–30 (2013). (in Russian)
17. Giv`on T.: Coherence in text, coherence in mind. Pragmatics and cognition 1(2), 171–227 (1993).
18. Garanina, N., Sidorova, E.: Ontology Population as Algebraic Information System Processing Based on Multi-agent Natural Language Text Analysis Algorithms. Programming and Computer Software 41(3), 140–148 (2015).
19. Garanina, N, Sidorova, E., Seryi, A.: Multiagent Approach to Coreference Resolution Based on the Multifactor Similarity in Ontology Population. Programming and Computer Software 44(1), 23–34 (2018).
20. Garanina, N., Sidorova, E., Anureev, I.: Conflict resolution in multi-agent systems with typed relations for ontology population. Programming and Computer Software 42(4), 31–45 (2016).
21. Toldova, S., Roytberg, A., Nedoluzhko, A., Kurzukov, M., Ladygina, A., Vasilyeva, M., Azerkovich, I., Grishina, Y., Sim, G., Ivanova, A., Gorshkov, D.: Evaluating Anaphora and Coreference Resolution for Russian. In: Computational Linguistics and Intellectual Technologies, Proceedings of the International Conference "Dialog 2013", pp. 681–695. Publishing House of the RSUH, Moscow (2013).