

ANALYSIS AND PROCESSING OF THE TEXT INFORMATION AIMED AT EXTRACTING BASIC KNOWLEDGE

Kryvyi Sergii, Bibikov Dmitriy

Abstract: *The problems extraction knowledge from natural language text is considered. An automatization approach to extraction knowledge is proposed.*

Keywords: *automatically analysis of natural languages text, extraction of knowledge.*

ACM Classification Keywords: *1.2 ARTIFICIAL INTELLIGENCE – 1.2.4 Knowledge Representation Formalisms and Methods.*

Introduction

The rapid development of science and technology during the last decades of the 20th century and early 21st century led to enormous information growth that one person (even highly skilled in science and technology) is unable to learn, understand and use to conduct research. Due to such a situation there is a need to automate the search and processing the necessary information for its subsequent efficient usage. To do this a few problems should be solved [1].

The first and one of the main problems is the analysis of natural language text information (morphological, syntactic, semantic and logical analysis) to extract knowledge.

The second problem is the issue of search engine design and extraction of knowledge, construction of its architecture and development of tools to help the user.

Third issue - is the integration of knowledge from multiple (or two) domains to ensure the effectiveness of interdisciplinary research, using existing algorithms, facts, theoretical principles and practical solutions.

The third problem is closely linked with such problems as the effective usage of automatic search of theorem proof in formal logic and problems similar to it. Application of the prover can only be successful when all the necessary information will be at its disposal. For example, if the prover needs to prove the Lagrange theorem which concerns the divisibility of the order of a finite group by the index of its subgroups, it's not enough for the prover to have axiomatically group theory. It also needs the axiomatic of divisibility theory, and perhaps axiomatic of Peano along with some additional facts from other areas. The solution to this problem is easier to find if prover has an integrated system that includes relevant information from other areas of knowledge. Then the prover finds the necessary information and uses it to conduct successful proof.

Certain process formalization is discussed in this paper, like the analysis of natural language texts (NLT), getting primary information from the NLT, finding logical conclusions from these facts and making sure they are correct.

Short overview of research methods of natural language texts

Formalization of natural language to automate the analysis of natural language texts was initiated in the early 30s of the 20th century by A. Tarski and his students, although this need was expressed long before - by Aristotle, Leibniz and Euler. In particular, Aristotle has provided four types of statements:

A - "all X is Y"; E - "no X is not Y"; I - "is Y some X", O - "some X is not Y".

These types of statements were called syllogisms, and the approach of Aristotle - syllogistics. Later Euler outlined his understanding of Aristotle syllogistics using the geometric interpretation of syllogistics as circles (this interpretation was called Euler's circles). Euler's ideas were developed further in the works of French mathematician and astronomer J.D. Zherhon who introduced types of relations and syllogistic interpretation of Aristotle in terms of these relations. The main types of relations imposed by Zherhon are: G1 – "the same or equivalent"; G2 – "left-side inclusion"; G3 – "special case of shrinkage"; G4 – "right-side inclusion"; G5 – "incompatibility" [9].

Zherhon showed that each type of a syllogism of Aristotle can be expressed as a set of some possible options of such relations. In particular A: (G1, G2), E: (G5), I: (G1, G2, G3, G4), O: (G3, G4, G5). For example, the statement I means that some non empty subsets of the set or class X is included in Y. The main difficulty in using Zherhon's relations is that almost all types of relations in a complex sentence require a large number of test options. Therefore, a more appropriate approach was based on the usage of formal mathematical logic.

Much more serious attempts were made by A. Tarski [5, 6], which led to the emergence of the notion of satisfaction of formulas – a more general concept than the notion of truth. This notion Tarski applied to open and closed formulas (in sentences of natural language a closed formula means a phrase), and this helped to formulate the concept of truth of natural language sentences and impose it on every open atomic formula that consists of a primitive predicate (subject constant) and as many variables, as correspond to the predicate arity. As the set of such formulas is finite, this approach is constructive.

The next attempt to improve formalization of A. Tarski was made by D. Davidson [7]. He proposed to add the recursive definition of truth to the concepts of truth and satisfiability. Thus the T theory, which includes a recursive definition of truth, does not explain how the meaning of phrases depends on the meaning of words of these phrases. But since the word "value" is not synonymous with "truth", the definition of truth is not always the definition (meaning). Thus, Davidson's thesis is not quite obvious, but it's easy to justify. Hence, it's necessary to compare the meaning of the narrative sentences with the terms of their correctness.

If we accept the equation "value of a narrative sentence = the condition of truth of this sentence", a condition must be set: if the definition describes how conditions of truth of a compound sentence depend on the conditions of validity of its constituent simple sentences, then the definition should describe how the value of a complex sentence depends on the values of its constituent simple sentences.

Montague also believed that the methods of formal semantics can be applied to the study of natural language semantics. But, unlike Davidson, he rejected the application of first order predicate logic, opting for categorical grammars. These grammars include those categories that specialists in traditional grammars use in definitions of natural language, e.g. such categories as subject or predicate. This is an opposite approach to the Davidson's. This made it possible to replace the notion of Montague's absolute truth with the notion of relative truth in the model, because in one model one sentence can be true, and in the other - false. This expansion helped to define the notion of logical truth and logical consequence of a larger fragment of natural language [8]. So Montague

highlighted two elements: intention (meaning) and extension (denotation) and applied them to subjects, predicates and phrases. There are other approaches to the analysis of natural language texts based on the notions of semantic networks, frames, etc.

If we briefly characterize the ideas that dominated the last decade of the 20th century, they can be reduced to the following.

In the beginning of the 1970s the studies of natural and artificial (formal) languages certain attempts to construct a theory dominated. This theory would consider both natural and artificial languages. Syntax is only considered in terms of semantics. The goal of semantics is to explain concepts of truth and logical imitation. The purpose of syntax is to characterize syntactic categories that form expressions.

This paper considers an approach which combines aspects of algebraic analysis of natural language and also the logical aspects of such analysis. The following text is structured as follows.

Formal statement of the problem of knowledge extraction from NLT is formulated. From this formulation follows concretization of texts of certain restrictions. As an examples of this type of constraints the syllogistics of Aristotle. In particular, for the syllogistics of Aristotle we suggest to build a set-theoretic interpretation of the extended system of rules of inference and analysis of possible situations that may arise in the application of this system of rules.

Later texts of definitions are considered that concern relations of subordination.

A formal statement of the problem of knowledge extraction from NLT

Before we begin the review of the system of processing and extraction of knowledge contained in the NLT, we shall define the notion of knowledge and knowledge extraction from NLT. For this purpose, we use the concepts used in programming with constraints [2].

Let this set D , which is identified in a finite set of n -arity relation R on D , so $R_i \subseteq D^n$, where $R_i \in R \subseteq D$, $i = 1, \dots, k$. The language restrictions L on D we call some non empty set $L \subseteq R \subseteq D$. The problem of satisfiability of constraints is formulated as follows.

Definition 1. For any set D and any constraint language L over D the constraint satisfaction problem (CSP (L)) is the solution of this combinatorial problem:

Instance: A triple $P = (V, D, C)$, where V is a finite set of variables; C is a set of constraints (C_1, \dots, C_q) ; each constraint $C_i \in C$ is a pair (s_i, R_i) , where s_i is a n -element sequence of V , which is called the domain restrictions, $R_i \in L$ is a n -ary relation over D , called the constraint relation.

Question: Does there exist a function $\varphi: V \rightarrow D$ such that for each constraint $(s, R) \in C$, which $s = (v_1, v_2, \dots, v_n)$, the tuple $(\varphi(v_1), \varphi(v_2), \dots, \varphi(v_n)) \in R$?

Set D in this case is called the domain of the problem, the set of all solutions CSP when $P = (V, D, C)$ is indicated by $\text{Sol}(P)$.

In case with NLT analysis, in order to extract knowledge a set D , field of the problem, is interpreted as a set of objects extracted from the input text T , which is factorized according to some equivalence relation R (we call this relation synonymous relation), which has "coded" relations $R_i, i = 1, 2, \dots, k$. Variables of the set of variables $V = \{v_1, v_2, \dots, v_m\}$ take their values in this factorized set of objects that appear in the text T (these can be lexico-grammatical levels, specific objects like people, dates, objects, etc.)

The problem of knowledge extraction from NLT is a problem of search of an interpretation $\varphi: V \rightarrow D$, while building obvious R_i relations from a set of $L \subseteq R$. The ratio $R_i \in L, i = 1, 2, \dots, k$, extracted from the text T, we shall call knowledge. This interpretation we are building in iterative manner.

While analyzing the NLT, our primary task is to build two fundamental relations which are present in virtually every NLT. This equivalence relation and partial order are known as generally valid. The first of these relations has been already discussed and it defines classes of synonymous objects, the second relation explains the hierarchy of equivalence classes. Both of these form the basis for building ontology, while the knowledge gained at this stage - will be called basic. With respect to partial order a different semantic meaning can be included: it may be relevant taxonomy ("belong" to the set, class, group, etc.), attitude of patrimony ("consist of"), related genealogy ("father-son"), cause-related relations ("if - then"), an attribute relation, etc.

The above definition of knowledge extraction from NLT is quite general and needs refinement. Let's consider some of the concrete definitions. Refinement can be performed in different directions depending on subject area and purpose pursued by analysis of NLT. We shall illustrate them with examples:

1) Aristotle's syllogism and its set-theoretic interpretation. If we analyze the types of generally meaningful relations, one can notice that they are associated primarily with the relation of partial order. This type of relation has the distributive grating which has useful properties and these properties can be used to generate effects, e.g. generate new knowledge. Moreover, it is easy to notice that the Aristotle's syllogisms are interpreted in the algebra of sets and relations with such dependencies [3, 10]:

$$A(X, Y) \Leftrightarrow X \subseteq Y, E(X, Y) \Leftrightarrow X \cap Y = \emptyset, I(X, Y) \Leftrightarrow X \cap Y \neq \emptyset, O(X, Y) \Leftrightarrow X \setminus Y \neq \emptyset.$$

And from the algebra of sets and relations it is known that operations of union, intersection and complement, this algebra is a Boolean ring, carriers of which are partially ordered sets against set-theoretic inclusion. Laws of algebra and properties of partially ordered sets can be used as inference rules in such a formal system. More details of these opportunities are considered in [3]. In particular, these are the laws of algebra of sets and relations (commutativity, associativity, distributivity, idempotency, acquisitions and De Morgan's laws), three basic properties of the inclusion relation (transitivity, and antisymmetry contraposition) and the law of double complementation. Thus the first two properties act as inference rules. We shall illustrate them with examples.

Example. The following facts are set (taken from the book by Carol L. "The History of nodules"): "All the little children are foolish"; "Anyone who can tame a snake deserves respect"; "All the stupid people do not deserve respect".

Let us find out what consequences follow from these facts. Note that this type of facts logic sometimes calls polysyllogisms or sorites. And syllogism is called a system which has only two replaces.

Now we define key terms that make up the system of facts, denote them and select the universe U. In this example, basic terms are: "little children" (C), "smart people" (P), "those who tame snakes" (T), "those who deserves respect" (Π). Clearly, these terms represent some sets in the universe "people". Their negations will be under the following terms: "grownups" ($\neg C$), "foolish people" ($\neg P$), "those who cannot tame snakes" ($\neg T$), "those who do not deserve respect" ($\neg \Pi$). Now take our facts look as follows: $C \subseteq \neg P, T \subseteq \Pi, \neg P \subseteq \neg \Pi$.

Thus, the grating is identified (as a universal set), which consists of elements ($\emptyset, U, P, T, \Pi, \neg C, \neg P, \neg T, \neg \Pi$), where U is the universe. Thus, the first effects of these facts are the following based on the rule of

contraposition (rule of contraposition in this interpretation of the form " $A \subseteq B$ follows from $\neg B \subseteq \neg A$, where the sign \neg means set complement):

$$(C^1): P \subseteq \neg C, \quad (C^2): \neg \Pi \subseteq \neg T, \quad (C^3): \Pi \subseteq P.$$

If we transfer the obtained effects in natural language, they respectively mean the following facts: "All smart people are not little children", "those who do not deserve respect, do not tame snakes", "he, who is a clever person, deserves respect". Using transitivity rule, we get the following consequences:

$$(C^4): C \subseteq \neg \Pi, \quad (C^5): T \subseteq P, \quad (C^6): \neg P \subseteq \neg T, \quad (C^7): \Pi \subseteq \neg C.$$

From these effects by the same rule of transitivity we get two more results:

$$(C^8): C \subseteq \neg T, \quad (C^9): T \subseteq \neg C.$$

If we translate the last consequences into natural language, they will sound like this: "All small children cannot tame snakes", "all who tame snakes are not small children».

From this example it follows that the problem of finding the effects is reduced to a problem of constructing a contra positive, transitive and ant-symmetrical closure (CTA circuit (closure)) of a basic set of formulas. When a closure is transitive, such situations may arise:

K1) the following formula is obtained $A \rightarrow \neg A$ or $\neg A \rightarrow A$;

K2) in the process of building a transitive closure, at least one cycle is obtained.

We will now consider what these situations in our case mean. The first formula in the case of K1) corresponds to the situation $A' \subseteq A$. From the properties of intersection and its complement we have $A \cap A' = \emptyset$, so the inclusion is true only if A is an empty set. The second formula in the case of K1) means the complement A' for set A should be a universal set. In terms of algebra of sets such situations cannot be described as contentious, but in our case, this situation means that some object must exist and at the same time does not exist. This situation is interesting for at least two reasons. The first reason is that the situation is catastrophic, and the second reason makes it possible to exclude certain terms that have led to controversy from consideration. We analyze these cases in detail. The first cause of such contention $A \rightarrow \neg A$ is the appearance of formulas like $A \rightarrow B$ and $A \rightarrow \neg B$ in the set of consequences. Another cause of controversy $A \rightarrow \neg A$ is the emergence of two formulas $A \rightarrow B$ and $\neg A \rightarrow B$.

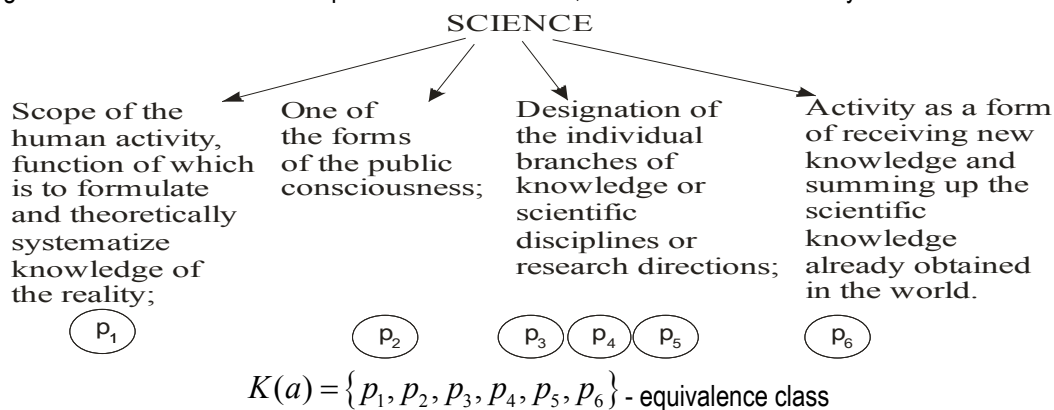
Note that the existence of such contradictions $A \rightarrow \neg A$ do not always lead to catastrophic consequences. Sometimes the appearance of such conflicts makes it possible to recognize the facts that lead to conflicts and delete conflicts.

The situations for K1) and K2) another extremely important point follows. Contradictions such as K1), K2) are formal because they appear only as a result of logical analysis given set of facts, but there is still a kind of contradiction, which differs significantly from the contradictions K1) and K2). Suppose that as a result of construction of ACT-circuit with good sound and proven facts that are not contentious, such as known and grounded theories, received conflicting effects, i.e. effects that are contrary to the facts of the original theories. It is said that controversy exists between the initial theories, and this is a sign of the emergence of new knowledge or at least the impetus for the analysis and search for causes of the appearing controversies. All of these gives us the ground to introduce such a definition.

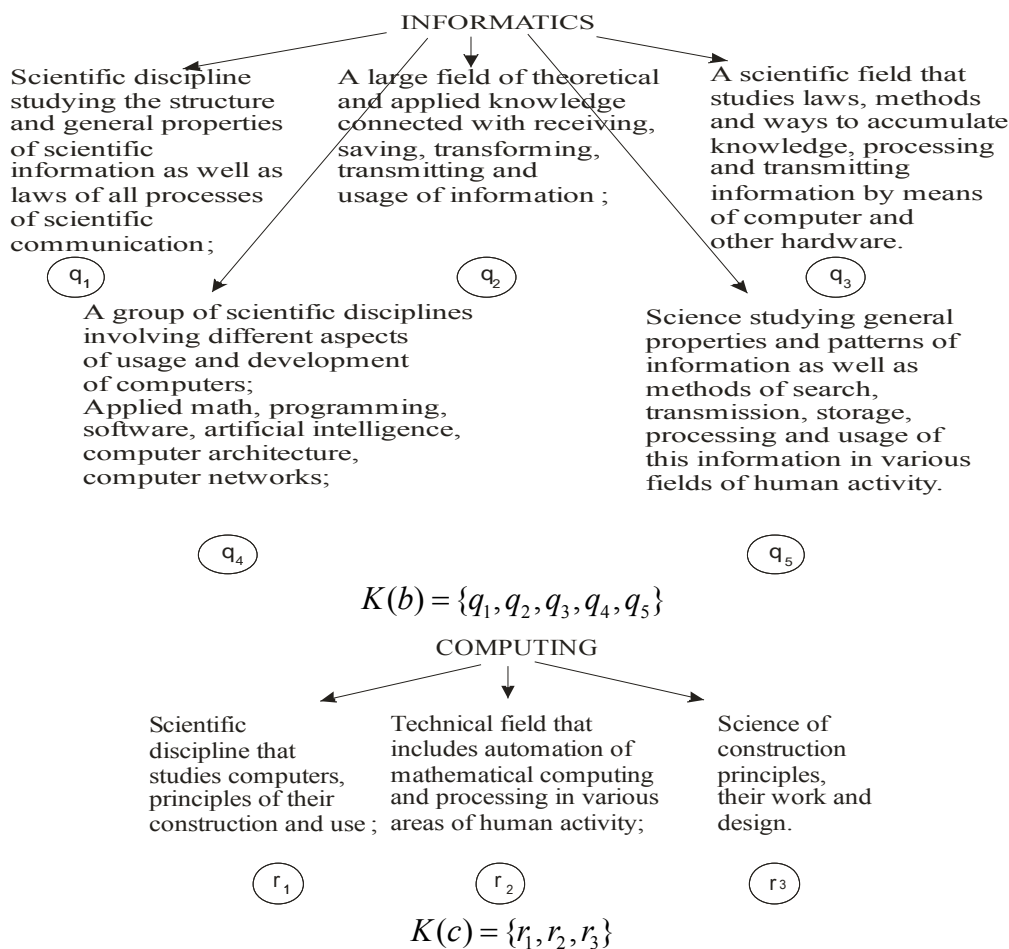
Definition 2. Information system is called correct if contradictions like K1) or K2) don't appear.

It is known that in the process of building ACT-closure, according to different sets of initial facts, you can get one and the same set of consequences. This allows us to write the following equivalence relation on the sets of facts: two sets of facts F and F' are called equivalent if the ACT (F) = ACT (F'). Based on this relation the structure of information system and initial set of facts can be simplified (by elimination).

2) Processing definitions. We consider the text that can be called conventionally structured. These are natural language definitions. Below are examples of such definitions, drawn from the dictionary.



The following definition is also taken from the dictionary.



These examples show that the construction of equivalence classes is not difficult. As a result of construction of equivalence classes of objects such results appear:

$$K(a) = \{p_1, p_2, p_3, p_4, p_5, p_6\}, K(b) = \{q_1, q_2, q_3, q_4, q_5\}, K(c) = \{r_1, r_2, r_3\}.$$

The problem appears when computing the second relation, that determines the ratio of subordination (hierarchy) between the equivalence classes found. But such an abstract representation of classes $K_i(x)$ of this relation cannot be determined (not enough information). It is necessary to know the structural characteristics of the elements of the classes $K_i(x)$. So, naturally the necessity to structure elements of equivalence classes appears. For example, if you return to the previous example, every element of the class $K(a)$ takes form:

$$p_1 = (p_{11}, p_{12}, p_{13}), p_2 = (p_{21}), p_3 = (p_{31}, p_{32}, p_{33}), p_4 = (p_{41}, p_{42}), p_5 = (p_{51}, p_{52}), \\ p_6 = (p_{61}, p_{62}),$$

where p_{11} = "sphere of human activity", p_{12} = "development of knowledge of the objective reality,"

p_{13} = "system of knowledge of the objective reality", p_{21} = "form of social consciousness",

p_{31} = "industry knowledge", p_{32} = "discipline", p_{33} = "scientific direction"

p_{41} = "activities to obtain new knowledge", p_{42} = "summation of knowledge of the SMW".

Similarly, other elements in the equivalence classes are structured.

$$r_1 = (r_{11}, r_{12}, r_{13}, \dots), r_2 = (r_{21}, r_{22}, r_{23}, \dots), r_3 = (r_{31}, r_{32}, r_{33}, \dots), \\ q_1 = (q_{11}, q_{12}, q_{13}, \dots), q_2 = (q_{21}, q_{22}, q_{23}, \dots), q_3 = (q_{31}, q_{32}, q_{33}, \dots), q_4 = (q_{41}, q_{42}, q_{43}, \dots) \\ q_5 = (q_{51}, q_{52}, q_{53}, \dots)$$

From the structure follows this formulation. If the equivalence class belongs to the object a , it looks like a formal definition of disjunction of elements that make up this class. Each item that is part of a class equivalence is described by the corresponding equivalence predicate, so if $K(a) = \{p_1, p_2, p_3, p_4, p_5, p_6\}$, then $p(a) \Leftrightarrow p_1(a) \vee \dots \vee p_6(a)$, where p_i are predicates characterizing element of the class $K(a)$, and their disjunction characterizes a class concept a .

Further, if $q_i \in K(a)$ and $q_i = (q_{i1}, q_{i2}, \dots, q_{ik})$, then element p_i (or object p_i), characterized by attributes p_{ij} , are represented as conjunction $p_i(a) \Leftrightarrow p_{i1}(a) \wedge \dots \wedge p_{ik}(a)$, where $p_{ij}(a)$ - is a predicate that characterizes the concept of separate attribute a , $i = 1, \dots, l; j = 1, \dots, k$.

Thus, each class $K(a)$ is described by disjunctive form, like:

$$p(a) \Leftrightarrow (p_{11}(a) \wedge \dots \wedge p_{1m_1}(a)) \vee \dots \vee (p_{l1}(a) \wedge \dots \wedge p_{lm_l}(a)).$$

Noted formalization defines a partial order relation, which is found the following way:

$$K(a) \leq K(b) \Leftrightarrow (\exists p_i(a))(\exists q_j(b))(q_j(b) \leq p_i(a)),$$

where $q_j(b) \leq p_i(a)$ means that $q_j(b)$ included as a member of the conjunctive in $p_i(a)$.

So related partial order naturally requires a predicate-relational representation of the objects of equivalence classes and most of these classes [4]. To illustrate the information, we go back to the example above. Let's consider, how the fact that class of "SCIENCE" subordinates the class of "INFORMATICS".

Class "SCIENCE", $K(a)$, is described by a formula: $p(a) \Leftrightarrow p_1(a) \vee p_2(a) \vee \dots \vee p_6(a)$, where $p_1(a) \Leftrightarrow$ SCOPE-HUMAN-ACTIVITY(a),

$$p_2(a) \Leftrightarrow \text{FORMS-PUBLIC-CONSCIOUSNESS}(a),$$

$$p_3(a) \Leftrightarrow \text{SCIENCE-DISCIPLINE}(a), p_4(a) \Leftrightarrow \text{SCIENCE-DIRECTION}(a),$$

$$p_5(a) \Leftrightarrow \text{ACTIVITY-RECEIVING-KNOWLEDGE}(a),$$

$$p_6(a) \Leftrightarrow \text{ACTIVITY-SUMMING-SCIENTIFIC-KNOWLEDGE}(a),$$

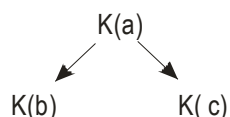
Class "INFORMATICS", noted as $K(b)$, is described by a formula:

$$q(b) \Leftrightarrow q_1(b) \vee q_2(b) \vee q_3(b) \vee \dots \vee q_l(b),$$

where $q_1(b) \Leftrightarrow p_3(a) \wedge \text{GENERAL-PROPERTIES-SCIENCE-INFORM}(b) \wedge \dots$

Using the above definition of partial order relations, we find that $p_3(a)$ belongs to a class $K(a)$ definition, and $q_1(b) \vee p_3(b) = p_3(b)$ (based on the law of absorption), which means that $K(b) \leq K(a)$.

Similarly we define the subordination of class $K(c)$ to the class $K(a)$, resulting in a graph:



So, the received hierarchy can be changed or modified through the dialogue with the user in order to achieve a more correct representation.

To sum up,, we formalize what was discussed in the above article about the processing of text definitions in the following way.

Suppose T is the set of texts definitions. Under this set of T a set of classes is constructed, determined by the equivalence relation R and is factor set D . Obtained in this manner set D of special relations are defined R_2, \dots, R_k , which describe the characteristic properties of elements D , i.e. elements of equivalence classes. These relations, which we will call primary knowledge are presented in the form of predicates that define the partial order relation R_1 . This relation is the second relation on which ontology is based. More precisely, ontology is built on transitive closure R_1^* by relation R_1 , which is correlated with the relation R .

Definition 3. Relations R and R_1^* we will call consistent if $\forall a, b \in D$ there is inclusion $(a, b) \in R^* R_1^*$, where $R^* R_1^*$ - and the superposition relations R and R_1 .

From this definition the primary ontology follows logically: $O = (D = T / R, \mathfrak{R} = \{R_1, R_2, \dots, R_k\}, \varphi, A)$ where $\varphi: D \rightarrow T$ - interpretation - A set of axioms, defined by predicates that describe the characteristic properties of elements from D , where R_2, \dots, R_k - corresponding relations, and R_1 partial order relation.

Conclusion

The proposed ways of automatic processing NLT are the basis of both theoretical and practical process of the analysis of the process of extraction of basic knowledge from NLT and their representation in the form of ontology. Using this framework, and especially its implementation, we will increase its capacity by building new meta relations over the built relations which are separate parts of basic knowledge, that are in the given NLT.

As a basis for building system analysis NLT, algebraic system of listed entities is used with the prospect of further hardware implementation. And the construction of a relevant ontology is performed on the basis of available tools and systems of construction of ontologies.

Bibliography

1. Palagin A.V., Kryvyi S. L., Petrenko N.G. Knowledge-oriented information systems with natural languages processing: foundation of methodologies and architecture-structured organization. – *journal. USiM.* – 2009. - №3. – C.42-55 (in russian)
2. Cohen D. Jeavons P. The Complexity of Constraint Languages. In "Handbook of Constraint Programming - Edited by F. Rossi, P. van Beek and T. Walsh. -2006. – P. 245 - 280.
3. Kulik B.A. Logic of natural reasoning - S.-Petersburg: Newskij dialect.- 2001.- 127 c. (in russian)
4. Rubashkin V.S. Representation and analysis of sense in informational systems. – M.: Nauka.- 1989.-188 c. (in russian)
5. Tarski A. The semantic conception of truth. *Philosophy and phenomenological Research.* – v.4. – 1944. – P. 241-375.
6. *Tarski A. Logique, Semantique and Metamathematique (1923-1944).* Colin. – Paris. – 1972.
7. *Davidson D. Proceedings of Philosophical Logic.* – Reidel. – Dordrecht. – 1969.
8. *Montague R. Universal grammars. Theoria. Formal Phylisophy: Selected Papers of R. Montague.* – Yale University Press. -1974. - vol. 36. – PP222-246.
9. *Thause A, Gribomont P., Hulin G. and other. Logical approach to artificial intellect. From modal logic to logic of data bases.* - M.: Mir.-1998. – 494 c. (translated in russian)
10. Kolmogorov A.N., Dragalin A.G. Introduction to mathematical logic. M.: Publ. MGU,1982, 118 c. (in russian)

Author information

Kryvyi Sergii – professor of Kiev national university; Ukraina, Kiev; str. Vladimirska, 40; e-mail: krivoi@i.com.ua

Area of scientific activity: Discrete mathematics, analysis, verification and program development

Bibikov Dmitriy – post graduate student of Glushkov's institute of cybernetics of NASU; e-mail: bb_coff@mail.ru

Area of scientific activity: Artificial intelligence, automatization of analysis NLT