Perspective Approaches to Semantic Knowledge Representation and their Applications in the Context of the Task of Automatic Identification of the Semantically Equivalent Fragments of the Text Documents

Aliaksandr Hardzei
Minsk State Linguistic University
Minsk, Belarus, 220034
Email: alieks2001@yahoo.com

Yury Krapivin
Brest State Technical University
Brest, Belarus, 224017
Email: ybox@list.ru

***Abstract*—The article presents the solution of the maintenance the trust in science in context of the problem of the automatic identification of the semantically equivalent fragments of the text documents. The solution refers to the implementation of the principles of unification of information representation in the memory of computer systems and semantic knowledge representation according to the Theory for Automatic Generation of Knowledge.**

Key words: **natural language, automatic text processing, research integrity, adopted fragment.**

I. Introduction

As a concept, "Ethics" deals with what is "good" or "bad" and also is concerned with the moral duty and obligations. For generations, every culture and society across the world, has established for their members' an ethical code which is to be adhered and practiced. It is even possible to treat such code as a standard which defines society's (people) values, beliefs and behaviors. The key for applying such standard for different situations of real life is a culture. It plays an important role in establishing the ethical standards in a specific society and to be practiced by all. A quite trivial example that illustrates an obvious fact of existence and ascendancy of ethical questions on the quality of the real life – is research integrity. The consequences of degradation processes that take place might lead to unfortunate results in different spheres from economics up to mental and physical health and the ultimately the failure of the business [1].

II. Reliable science and research integrity

Today high-quality, reliable science and research integrity go hand-in-hand. The scientific research is increasingly interdisciplinary, public-private partnerships are commonplace and there is large-scale global competition. These developments put integrity under pressure [2].

Dr. Robert Merton in his work "A Note on Science and Technology in a Democratic Order" [3] articulated an ethos of science. He argued that, although no formal scientific code exists, the values and norms of modern science can nevertheless be inferred from scientists' common practices and widely held attitudes. Merton discussed four idealized norms: universalism, communality, disinterestedness, and organized skepticism. The first one, universalism, refers to the idea that scientific claims must be held to objective and "preestablished impersonal criteria." This value can be inferred by the scientific method or the requirement of peer review before publication in the vast majority of academic journals. The second one, communality, means that the findings of science are common property to the scientific community and that scientific progress relies on open communication and sharing. According to the norm of disinterestedness the science should limit the influence of bias as much as possible and should be done for the sake of science, rather than self-interest or power. Disinterestedness can often be the most difficult norm to achieve, especially when one's job or academic status relies on publications or citations. Many scientists believe that lack of disinterestedness is a systemic issue – one that funders, publishers, and scientists alike should work to address. Who, for example, gets to prioritize research: those who fund, implement, or publish it? And the last one, organized skepticism, refers to the necessity of proof or verification subjects science to more scrutiny than any other field. This norm points once again to peer review and the value of reproducibility [4].

"We need to take a systematic and institutional approach to overcome challenges to integrity," says project coordinator Hub Zwart of Radboud University Nijmegen in the Netherlands. "Research integrity must be supported by the research culture. This means we don't want a campaign against misconduct. Instead, we want to promote a campaign for good science – research is teamwork and research integrity is teamwork"[2].

In a general way, the term "Education" can be referred to as the transmission of values and knowledge accumulated by a society to its members through the process of socialization or enculturation. Every individual will acquire education that will guide him to learn a culture, wherein he molds his behavior and directs it towards the eventual role in the society. In the context of the research integrity the education can be acquired through activities conducted by the scientific adviser and the entire research environment takes the place of the school.

As shown in [1] institutions of higher education have a major role to play in preparing the younger generation for a propitious future. Apart from imparting quality education, they need to instill high ethical values and practices amongst the young researcher fraternity. The rapid development of information technologies all over the world, which is reflected in the popularization of the electronic form of information storage, accumulation and processing in all areas of human activity, makes this task quite sophisticated. That's why the solving the task at different levels: from college up to research institution can ensure a bright future.

As an example, the wrong understanding of plagiarism and incorrect definition could cause academic society with of numerous cases of plagiarism. Students often unclear understood plagiarism and how to prepare properly a written work. Teachers also could interpret plagiarism differently. Plagiarism often becomes as a moral maze for students. They deal with a moral and ethical dilemma having unclear understanding what behavior is appropriate. Therefore the institutions and academic communities are suggested to discuss and formally provide a detailed and clear definition of plagiarism in process of implementation of plagiarism prevention systems of higher education [5]. In recommendations for plagiarism prevention policy formation is suggested creating a definition of plagiarism not only on institutional but also at national level. This definition should be appropriate to different fields of study and daily use, specifying what behavior and actions are acceptable [6].

The research community promotes integrity to maintain trust in science in different ways. According to [7] Japan Agency for Medical Research and Development (AMED) was launched in April 2015 to promote integrated medical research and development (R&D) ranging from basic research to practical applications, in order to smoothly achieve the nationwide application of research outcomes, and to establish an environment therefor. AMED consolidates budgets for R&D expenses, which had previously been allocated from different sources, such as the Ministry of Education, Culture, Sports, Science and Technology, the Ministry of Health, Labour and Welfare, and the Ministry of Economy, Trade and Industry. It provides funds strategically to universities, research institutions, etc. By promoting medical R&D, AMED aims to achieve the world's highest level of medical care/services to contribute to a society in which people live long and healthy lives. To achieve this mission, it is imperative that R&D funded by AMED is widely understood and supported. Maintaining and improving research integrity is a prerequisite to this end. AMED is taking various measures to ensure fair and appropriate R&D. It is asking researchers to participate in its responsible conduct in research (RCR) education program and to comply with its rules for managing conflicts of interest. In addition, AMED also conducts a grant program to create and distribute a variety of educational materials on RCR and other matters. Further, AMED is establishing a platform that allows researchers to exchange information about research integrity, and it is undertaking additional measures, such as holding meetings and international symposia on research integrity. This example demonstrates that research integrity is not only a located in the minds of individual researchers. Rather it is also a characteristic of organizations, such as research departments, universities, or networks [8]. Furthermore, it is not a stable trait of organizations but represents ongoing processes in organizations. This means that (organizational) integrity can be developed and nurtured – by managers as well as by the members of the organization [9]. The usage of the systems of adopted text fragments recognition [10] -– is one of such traits. The work [11] demonstrates a general algorithm of solving the task of automatic recognition of semantically adopted fragments of the text documents, which is based on the knowledge system, oriented to the recognition of the concepts, actions and their attributes in the text documents.

III. Perspective approaches to semantic knowledge representation for Artificial Intelligence and their applications

Quite bright example of trends in natural language processing – ACL 2019 – the 57th annual meeting of the Association for Computational Linguistics that took place in Italy and enrolled more than 3000 registered attendees (2900 submissions, 660 accepted papers). The application of knowledge graphs in solving natural language processing tasks was researched about 30 papers out of 660 which is a good 5% share for such a huge event [12].

Another one novel and perspective approach to semantic knowledge representation for Artificial Intelligence was formulated by prof. A. Hardzei, which proposed the theory having in the foundation semantic formalism for knowledge representation and knowledge inference.

The theory is based on Universal Semantic Code (USC) developed by prof. V.V. Martynov where there were proposed semantic primitives, i.e. semantically irreducible kernel words, and the rules of their combinations were defined. In general, semantic coding is a process of converting natural language phrases to a chain of semantic primitives or semantic formulas and back. This is the crucial difference with Semantic Web where there are no semantic formulas but semantic tags represented not formally but by means of natural language.

The essential difference of Theory for Automatic Generation of Knowledge Architecture (TAPAZ) vs USC is in the method of defining a structure of the semantic formula and operations of transformation of semantic formulas to each other.

USC operates with complex formulas consisting of two parts reflecting a consequence "if ...than..." or "stimulus → reaction". TAPAZ proposes an alternative way of representing semantic formulas as extended formulas generated by adding parenthesis on the right margin according to determined rules. Each formula has a "semantic counterpart", coupled with a mathematical formalism a highlighted fragment of the Model of the World, representing an interpretation of the formula in natural language. Each formula has one, and only one semantic meaning. The meaning is not assigned but inferred from the structure of the formula.

Besides, the theory is supported by geometric model of the World demonstrating how one "individ", a kind of a pattern as a separate entity in the selected fragment of the Model of the World, which consists of the core, the shell and the surroundings, of the model transfers an impulse to another individ through an environment. The impulse direction depends on the role of the individ. There are only four roles: subject, object, instrument, and mediator. The roles indicate members of some action and may be strictly specified according to its purpose, for example, the role of the instrument may be specialized only as: activator, suppressor, enhancer or converter. The semantic formulas in TAPAZ represent actions surrounded by members of the action.

Accumulating all together a semantic classifier of actions has been proposed. The classifier has 112 semantic classes. Each name of the class represents a highest abstract level of the action and supported with a list of actions giving concrete implementation. For example, the class action "connect" may be implemented by: gluing, nailing soldering etc. Such a structure has an ontological nature and has a practical application for calculation of subject domains [13].

Thus, the system core (World Model) based on TAPAZ is a combination of 112 macro-processes (actions) of the TAPAZ Semantic Classifier with a series of specialized processes of the selected subject domain, for example, Remote Sensing of the Earth (ERS)[1]. Each process has 18 semantic fields in accordance with the TAPAZ Role List of Individs. The semantic weight of ERS-process in a synonymous series is determined by the completeness of the fields, the frequency index and the height of the Oriented Graph of the Semantic Classifier as its vertices are filled.

The construction of the system core is carried out in manual, semi-automatic and automatic modes. At the first stage, the formation of the Intellectual Knowledge Base by prescribing ontological relationships between the independent taxonomy of ERS-individs and the dependent taxonomy of ERS-processes is allowed.

The Specialized Intellectual Knowledge Base (SIKB) of actual space research and technology in the field of ERS combines the TAPAZ-based system core with the periphery generated by the Oriented Graph of the Semantic Classifier, complemented by stepwise recursion and expanded by the TAPAZ Role List of Individs.

The rules for constructing, restricting, reducing and transforming algebraic formulas, the rules for semantic reading of algebraic formulas and interpreting typical combinations of individs, the rules for constructing the Oriented Graph of the Semantic Classifier, the procedure for semantic coding and decoding of its vertices, the groups and rows of the Semantic Classifier, the TAPAZ Role List of Individs and step recursion form the Universal Problem Solver directly interacts with Web Interface.

The algorithm for extracting specialized terminology from ERS-content by an expert to fill the Knowledge Base involves the software ExpertTool version 2.3.4.7, developed by A. A. Matsko. The algorithm for extracting specialized terminology from ERS-content by an expert and filling in the TAPAZ semantic fields (the role list of individs in the ERS-domain) is similar to the TAPAZ Role List of Individs:
subject (initiator→spreader→inspirer→creator)
→instrument(activator→suppressor→enhancer
→converter)→mediator (landmark→locus
→carrier→adapter→material→prototype
→resource→source→chronotope→fund)
→object →product.

Interactive filling the semantic fields for each specified process of the ERS-domain requires answers to typical questions:

Who? With which tool? In relation to whom / what? In what place? Arriving on what? Adjusting with what?

[1] Currently, TAPAZ technology is used as linguistic support for the event's Internet monitoring subsystem 5[1] "Development of an Intelligent Web-Based System for Searching and Processing Content in English and Russian Using the Semantic Coding Method for Information Support for Solving Managerial, Design-and-Search and Expert Tasks in the Field of Remote Sensing of the Earth", section 1 "Development of the Belarusian Space System for Remote Sensing of the Earth", sub-programme 7 "Research and Peaceful Use of Outer Space" of the State program "High Technology and Technics" for 2016–2020.

Making of what? Following what example? Spending what? Knowing what? In what period? Due to whose prompt? Affecting who / what? Produces whom / what?

All search engines currently operating in the world search either by keywords or by tuples of keywords (keywords in nominal and / or verb groups) using standard software tools for content preprocessing and automatic lexical (tagger) and syntax (parser) markup.

The main drawback of both types of search is its inaccuracy and, as a consequence, the immensity for the user of the huge number of URL (Uniform Resource Locator) found by the search engine, forcing search engine developers to limit the search area on the Internet to the region of the query place, which leads to the knowingly incomplete search results. Attempts to supplement the search for keywords and / or their tuples with contextual synonyms based on empirical ontologies that are incorrectly called "semantic"[2] only increase the inaccuracy and incompleteness of the search, overloading the search page with information noise and causing the user to feel unreliable of the received sample.

The TAPAZ technology offers a search by event fragments or technological cycles, which are described by special units that are macro-processes[3] in the assembly, when specialized ERS-processes are put in accordance with a certain algorithm in accordance with TAPAZ macro-processes and the roles of all participants in the events[4] are calculated.

This approach provides maximum accuracy and speed of search, relevance of search results. In addition, it allows you to find similar technological cycles in close (adjacent) and distant subject domains, thereby providing support to the user in analytical activities, which greatly expands the functionality of the search engine, shifting it to the side of inventing.

Within the framework of the "Fractal" Development Work, the software ExpertTool 2.3.4.2 was created, which allows manual semantic markup of content using the TAPAZ technology. A manual of the semantic preprocessing of ERS-content has been prepared for experts in order to unify manual conjugation of specialized ERS-processes with TAPAZ macro-processes. The following were developed: 1) an algorithm for the formation of TAPAZ-units; 2) an algorithm for an expert to extract specialized terminology from thematic ERS-content to fill the Knowledge Base; 3) the algorithm for updating the user request to the system at the request of the system. The properties of the TAPAZ-2 Semantic Classifier Graph as a constructor of the Knowledge Base Architecture of an Intelligent WEB-system and ways to reduce the number of its vertices for sequential generation, processing and storage of the Graph taking into account the capabilities of modern computing technology are determined. A prototype of the TAPAZ-2 dictionary was prepared in the form of an Excel table, including 416 of the most frequency specialized predicates of Remote Sensing and 516 of their uses, coupled with 112 macro-processes of the TAPAZ-2 Semantic Classifier. The semantic structure and semantic functionality of the System are proposed and justified, 18 typical roles of event participants were decoded.

The ultimate goal is automatic semantic markup of ERS-content in the TAPAZ technology, which allows to achieve maximum efficiency (accuracy, speed and completeness) of the search engine, as well as the automatic assembly of TAPAZ-units in the Knowledge Base for the analytical support of management, design-and-search and expert solutions remote sensing tasks.

One more trendy approach to semantic knowledge representation for Artificial Intelligence was formulated by prof. V. Golenkov. He proposed the concept of the Open Semantic Technology for Intelligent Systems Design (OSTIS Technology) [14, 15], focused on the development of computer systems of the new generation (first of all – hybrid intelligent systems [16]) which will be called semantic computer systems or ostis-systems, if it is necessary to emphasize their compliance with the standards of OSTIS Technology. The model of hybrid knowledge bases of ostis-systems and models for representing various types of knowledge within the framework of such a knowledge base [17], as well as model of a hybrid problem solver, which allows to integrate various problem solving models [18], were also proposed.

The main requirement for OSTIS Technology is to ensure the possibility of joint use within the ostis-systems of various types of knowledge and various problems solving models with the possibility of unlimited expansion of the list of knowledge used in ostis-system and problem solving models without significant labor costs. The consequence of this requirement is the need to implement the component approach at all levels, from simple components of knowledge bases and problem solvers to whole ostis-systems. To meet these requirements, the most important task is not only the development of appropriate ostis-systems models and their components, but also the development of an integrated methodology and appropriate tools for automating the construction and modification of ostis-systems [19].

Prof. Golenkov stated also the principle of unification

[2] "Functional concepts such as "Subject", "Predicate" must be carefully distinguished from categorical concepts, such as "Noun component", "Verb", and the difference between them should not be masked by the fact that sometimes the same term is used for concepts both types". Chomsky N. (1972) Aspects of the theory of syntax, Moscow, Moscow State University, p. 65 (in Russian).

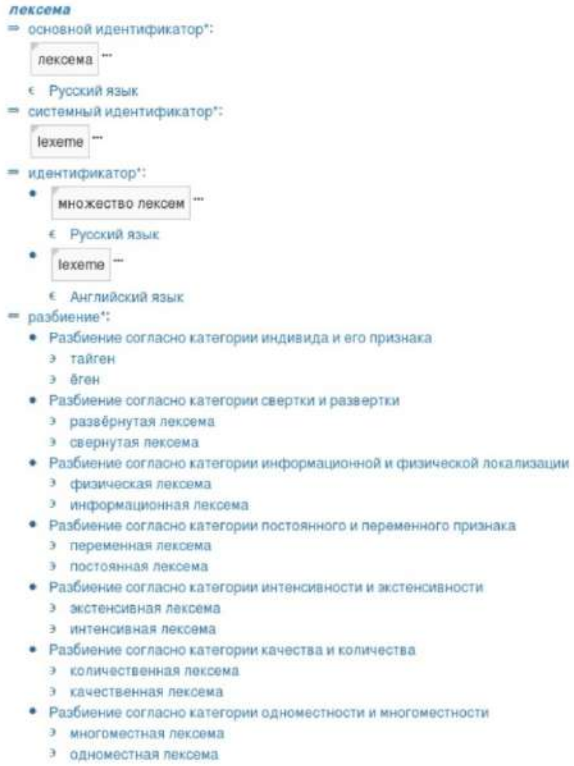
[3] Macro-process is one of 112 extremely abstract processes that are isomorphic to any subject domain and are calculated and encoded by the TAPAZ-algebra.

[4] "There are such concepts as "culprit", "tool", "product of labor" <...> We are here in the field of various categories, apparently ontological, but essentially semantic". Kotarbinski T. (1975) Treatise on the Good Work. Moscow, Economics, p.31 (in Russian).

of information representation in the memory of computer systems and developing the standard that should not limit the creative freedom of the developer, but guarantee the compatibility of its results. As the standard of the universal sense representation of information in the memory of computer systems he proposed SC-code (Semantic Computer Code). Unlike USC of V.V. Martynov, who made an important step in creating a universal formal method of sense coding of knowledge, it, firstly, is nonlinear in nature and, secondly, is specifically focused on coding information in the memory of computers of a new generation, focused on the development of semantically compatible intelligent systems and called semantic associative computers. Thus, the main leitmotif of the proposed sense presentation of information is the orientation to the formal memory model of a non-Von-Neumann computer designed for the implementation of intelligent systems using the sense representation of information. The features of this representation are as follows:

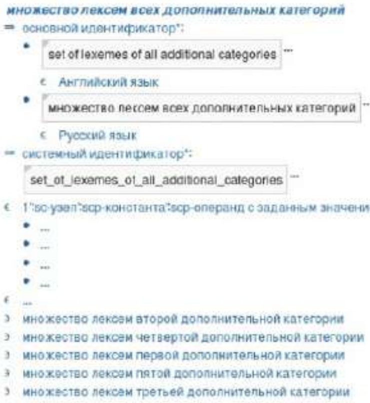
- associativity;
- all information is enclosed in a connections configuration, i.e. processing information is reduced to the reconfiguration of connections (to graph-dynamic processes);
- transparent semantic interpretability and, as a result, semantic compatibility [19].

Taking into account the trends of the AI approaches development, it is proposed to strengthen the linguistic database [20] with the help of approaches mentioned above in the context of solving the task of automatic recognition of semantically adopted fragments of the text documents. The examples of the ostis-knowledge base that describe the knowledge domain of grammatical categories in terms of TAPAZ, available to use in implementation of algorithms of the natural language processing at different levels of the deep of the analysis, are shown in Figures 1 – 3.



лексема
⇒ основной идентификатор*:
лексема
∈ Русский язык
⇒ системный идентификатор*:
lexeme
⇒ идентификатор*:
• множество лексем
∈ Русский язык
• lexeme
∈ Английский язык
⇒ разбиение*:
• Разбиение согласно категории индивида и его признака
∋ тайген
∋ ёген
• Разбиение согласно категории свертки и развертки
∋ развёрнутая лексема
∋ свернутая лексема
• Разбиение согласно категории информационной и физической локализации
∋ физическая лексема
∋ информационная лексема
• Разбиение согласно категории постоянного и переменного признака
∋ переменная лексема
∋ постоянная лексема
• Разбиение согласно категории интенсивности и экстенсивности
∋ экстенсивная лексема
∋ интенсивная лексема
• Разбиение согласно категории качества и количества
∋ количественная лексема
∋ качественная лексема
• Разбиение согласно категории одноместности и многоместности
∋ многоместная лексема
∋ одноместная лексема

Figure 1. The class of base semantic categories



множество лексем всех дополнительных категорий
⇒ основной идентификатор*:
• set of lexemes of all additional categories
∈ Английский язык
• множество пексем всех дополнительных категорий
∈ Русский язык
⇒ системный идентификатор*:
set_of_lexemes_of_all_additional_categories
∈ 1"sc-узел"scp-константа"scp-операнд с заданным значением":
• ...
• ...
• ...
• ...
∈ ...
∋ множество лексем второй дополнительной категории
∋ множество лексем четвертой дополнительной категории
∋ множество лексем первой дополнительной категории
∋ множество лексем пятой дополнительной категории
∋ множество лексем третьей дополнительной категории

Figure 2. The class of supplementary semantic categories.

IV. Conclusion

An involvement of the well-developed linguistic text analysis that is based on the knowledge of natural language, together with the implementation of the principles of unification of information representation in the memory of computer systems and semantic knowledge representation according to the Theory for Automatic Generation of Knowledge Architecture might made a contribution to maintenance the trust in science and the development of national science in the sphere of the Artificial Intelligence.

References

[1] Ethics in Higher Education. Available at: https://www.igi-global.com/gateway/chapter/114343. (accessed 2019, Dec).

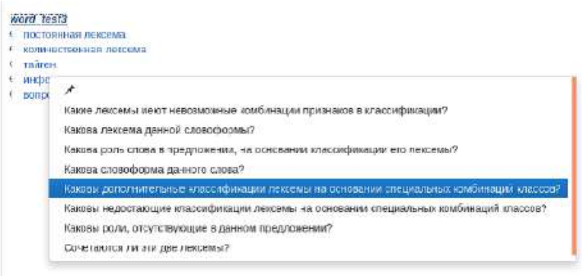
[2] Promoting research integrity for a ... - Information Centre - Research & Innovation - European Commission. Available at: https://ec.europa.eu/research/infocentre/article-en.cfm?artid=49608. (accessed 2019, Dec).

[3] Robert K. Merton, The Normative Structure of Science (1942). Available at: https://panarchy.org/merton/science.html. (accessed 2019, Dec).

[4] Merton's norms and the Scientific Ethos. Available at: https://www.futurelearn.com/courses/open-social-science-research/0/steps/31422. (accessed 2019, Nov).

[5] Carroll, J., & Appleton, J. Plagiarism: A good practice guide. Available at: http://www.jisc.ac.uk/uploaded-documents/brookes.pdf. (accessed 2019, Dec).

[6] Sarlauskiene, L., Stabingis, L. Understanding of plagiarism by the students in HEIs of Lithuania / *Procedia - Social and Behavioral*



word test3
- постоянная лексема
- количественная лексема
- тайген
- инфо
- вопр

Какие лексемы иеют невозможные комбинации признаков в классификации?

Какова лексема данной словофоомы?

Какова роль слова в предложении, на основании классификации его лексемы?

Какова словоформа данного слова?

Каковы дополнительные классификации лексемы на основании специальных комбинаций классов?

Каковы недостающие классификации лексемы на основании специальных комбинаций классов?

Каковы роли, отсутствующие в данном предложении?

Сочетаются ли эти две лексемы?

Figure 3. The way of communication with the help of proper intellectual agents of natural language text processing.

Sciences, 2014, pp.638 – 646.

[7] Yakugaku Zasshi.Promoting Research Integrity. Available at: doi: 10.1248/yakushi.17-00181-3. (accessed 2019, Dec).

[8] Palazzo, G. Organizational integrity—understanding the dimensions of ethical and unethical behavior in corporations, *Corporate ethics and corporate governance*, 2007, pp. 113-128.

[9] Paine, L. Managing for organizational integrity, *Harvard business review*, 1994. vol 72, no. 2, pp. 106-117.

[10] Krapivin, Y. System «PlagiarismControl» as the Tool for the Expertise of the Text Documents, *Informatics*, 2018, vol. 15, no. 1, pp. 103–109 (in Russian).

[11] Krapivin, Y.B. Metody i algoritmy avtomaticheskogo raspoznavanija vosproizvedennyh fragmentov tekstovyh dokumentov: *avtoref. dis. kand. teh. nauk: 05.13.17*, Bel. gos. un-t., - Minsk, 2019. - 24 p.

[12] ACL 2019 . Available at: http://www.acl2019.org/EN/index.xhtml. (accessed 2019, Dec).

[13] Hardzei, A. Theory for Automatic Generation of Knowledge Architecture: TAPAZ-2, Minsk, 2017, 50p.

[14] Golenkov, V. V., Guljakina, N. A."Proekt otkrytoj semanticheskoj tehnologii komponentnogo proektirovanija intellektual'nyh sistem. chast' 1: Principy sozdanija [project of], open semantic technology for component design of intelligent systems. part 1: Creation principles, *Ontologija proektirovanija [Ontology of design]*,2014, no. 1, pp. 42–64.

[15] Golenkov, V. [et al.] "From training intelligent systems to training their development tools," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, Ed., BSUIR. Minsk , BSUIR, 2018, pp. 81–98

[16] Kolesnikov, A. Gibridnye intellektual'nye sistemy: Teoriya I tekhnologiya razrabotki [Hybrid intelligent systems: theory and technology of development], A. M. Yashin, Ed. SPb.: Izd-vo SPbGTU, 2001.

[17] Davydenko, I. "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2018, pp.99–118.

[18] Shunkevich, D. "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2018, pp.119–132.

[19] Golenkov, V. [et al.] "Methods and tools for ensuring compatibility of computer systems," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2019, pp. 53–90.

[20] Krapivin, Y. Funktsional'nost' cross-language v zadache avtomaticheskogo raspoznavaniya semanticheski ekvivalentnykh fragmentov tekstovykh dokumentov [Cross-language Functionality in the Problem of the Automatic Identification of the Semantically Equivalent Fragments of the Text Documents], *Iskusstvennyi intellect [Artificial Intelligence]*, 2013, vol. 62, no. 4, pp. 187–194.

Перспективные подходы к семантическому представлению знаний и их приложения в контексте задачи автоматического распознавания семантически эквивалентных фрагментов текстовых документов

Гордей А.Н., Крапивин Ю.Б.

В статье представлено решение проблемы поддержания доверия к науке в контексте задачи автоматического распознавания семантически эквивалентных фрагментов текстовых документов. Решение основывается на реализации принципов унификации представления информации в памяти компьютерных систем и представления семантических знаний в соответствии с теорией автоматического порождения архитектуры знаний (ТАПАЗ).

Ключевые слова: естественный язык, автоматическая обработка текста, достоверность научного исследования, заимствованный фрагмент.

Received 06.02.2020