Russian Lexeme Processing and Generating a Tag-Semantic Dictionary for a Selected Domain

A. Hardzei
Minsk State Linguistic University
Minsk, Belarus
alieks2001@yahoo.com

R. Panashchik
The United Institute of Informatics Problems of the NAS
Minsk, Belarus
roman00201p@gmail.com

M. Svyatoshchik
The United Institute of Informatics Problems of the NAS
Minsk, Belarus
svyatoshchikm@mail.ru

O. Stralchonak
The United Institute of Informatics Problems of the NAS
Minsk, Belarus
oleg@stralchonak.com

V. Tkachenko
The United Institute of Informatics Problems of the NAS
Minsk, Belarus
tkach@newman.bas-net.by

A. Shumilin
Academy of Sciences of Belarus, Department of Physics, Mathematics and Informatics
Minsk, Belarus
shumilin@presidium.bas-net.by

***Abstract*—The methodology and algorithm for marking lexical units of Russian-language texts in Combinatory Semantics technology are presented. Fuzzy logic and fuzzy neural network rules have been proposed to automatically convert Parts of Speech to Parts of Language.**

***Keywords*—lexical analysis, linguistic semantic category, tagging, morphology, fuzzy rules, neural network.**

I. Introduction

Syntactic analysis (defining the role structure of a sentence) and formalization of semantics (determining the role structure of an event described by a sentence) are key stages in processing and understanding natural language texts, and while syntactic analysis (parsing) is actively used in computer systems to formulate rules or to detect contradictions in the description of any language grammar, then the capabilities of hardware-software complexes in explaining semantic relations, formation of semantic text representation, coding and decoding of the World Model are still in their development. The rapidly expanding number of scientific articles and publications has long ago exceeded the limit of their manual processing by specialists. Without automatic analysis of texts, their abstracting and summarizing, creation of reference lists and electronic academic books 'instantly' at the user's request, automatic semantic diagnostics and expertise – in other words, without artificial intelligent knowledge management systems – not every scientist or specialist can successfully navigate in their own and related scientific domains [1]. Active development of computer hardware requires new approaches to Natural Language Processing, new theories and algorithms that take into account the semantic meanings of language units, their contextual and meta-contextual dependence. The principles of building a lexical analyzer as the main tool for an intelligent system of semantic text processing based on the Theory for Automatic Generation of Knowledge Architecture (TAPAZ-2) are presented below [2].

II. Artificial Intelligence in the Internet Content Processing

Today, the development of Artificial Intelligence in the sphere of the Internet content and Natural Language Processing follows three directions:

- the Internet and Web services intellectualization (Semantic Web Project of W3C Consortium, Scheme.org International Public Resource Project of the Google, Microsoft, Yahoo and Yandex Developer Community), transformation of the Internet into a 'Giant Global Graph' (GGG) with simultaneous presentation of the possibility to connect automatic control systems to it ('smart home', 'smart city' projects, etc.);
- using machine learning, neural networks and other algorithms for statistical processing of content, including Big Data, in the creation of local and global information resources as a basis for the digitalization of the economy and its transition to the 7th Technological Paradigm; the most famous example of this approach is the development of chatbots with Generative Artificial Intelligence (ChatGPT from OpenAI, DeepSeek, etc.) as services for checking and acquisition of information to replace 'traditional' search engines;
- increasing the efficiency of Large Language Models (LLMs) by integrating them with the RAG (Retrieval-Augmented Generation) system to connect with relevant external knowledge sources (articles, technical descriptions, etc.) and to search information.

The rapid development of 'activity-based technology' (since 2011) and 'Activity Vocabulary technology' (since 2017) has shown that in the nearest 10–15 years the main efforts of international scientific and financial centers will be focused on the construction of knowledge graphs for automatic extraction of semantically meaningful information from the Internet search pages, i.e. on the gradual development of a formalized meta-language, correlating in semantic power with natural language and capable of representing it in machine-readable form [3].

The main achievement of the Minsk School of Combinatory Semantics is the Theory for Automatic Generation of Knowledge Architecture (TAPAZ-2, http://tapaz.by) for computer modelling of human intellectual activity systems and inventive problem solving. The formal apparatus of TAPAZ-2, being a finalization of V. V. Martynov's Universal Semantic Code (USC), is intended for algebraic coding of semantics (meaning of signs, sense of sentences, content of patterns in the World Model and connections in the knowledge architecture), i. e. for calculating the semantics of subject domains, as well as automatic semantic marking up of any texts [4, p. 5–18].

TAPAZ code is a unconventional correlation of an algebraic expression with the semantic counterpart (an individ or its attribute in the role structure of an event) and has strict rules for interpreting regular superposition of individs with the transition from morphological categories of Parts of Speech (PS) to semantic categories of Parts of Language (PL) [5]–[7]. Thus, TAPAZ codes codify all macroprocesses and processes of subject domains, while simultaneously calculating them. Through the reverse process of decoding the patterns of the World Model rather than the categories of Linguistic Image of the World, TAPAZ code decodes the meaning of words and word-combinations, the sense of sentences and texts, i.e. natural-language semantics. This makes it unique and gives it an advantage over the mentioned international achievements in this field [8, p. 5–26].

III. The Architecture of the Universal Semantic Analyzer

The ideology of TAPAZ is the creation of a flexible and open architecture of a multilevel universal semantic analyzer for automatic markup of Internet content or text corpora from different subject domains (Fig. 1). After developing this system and training it on the texts from different subject domains marked up in the PL paradigm, it will be possible to widely implement and use this system for intelligent knowledge management. The kernel of the whole system is a lexical analyzer, with the help of which a tag-semantic dictionary of the selected subject domain is created according to a certain algorithm, where each lexical unit (word) is tagged with a certain semantic code-category from the Special Paradigm of PL [5, p. 173–179]. A tag is a code marker of the corresponding

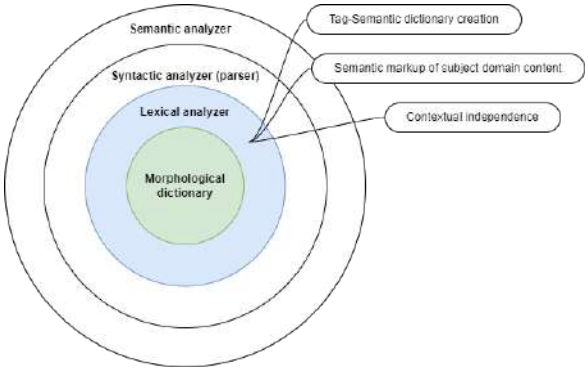

Figure 1. The architecture of the Universal Semantic Analyzer

category in the PL paradigm – taigens (substantives) and yogens (predicatives), facilitating the process of semantic markup of sentences and search for necessary information. For example, the word *lecturer* is assigned a tag: *taigen, reduced, contracted, physical, constant, common, qualitative, single-place*. One sign can also belong to different tags: *canteen* (dinnerware), *canteen* (room for eating), etc. The tag-semantic dictionary constructed in this way allows us to solve the problem of homonymy [9, p. 29–36] and will be used later at the parsing stage for syntactic and semantic analysis of texts.

IV. The Algorithm of the Lexical Analyzer

After its launch lexical analyzer connects to the available dictionaries. Actual project is functioning on the source morphological dictionary containing over 300,000 words, pre-marked with the following morphological characteristics: Part of Speech, Animate – Inanimate, Gender, Number, Case, Abbreviation, Name, Toponym, Organization, Possible substance, Trademark, Qualitative, Superlative, Ordinal, Possessive, Comparative degree, Category of aspect, Perfect – Imperfect, Category of superlative, Multiple, Reflexive, Person categories, Categories of Time, Mood category, Collateral category, Interrogative, Demonstrative, Introductory word, Adverbial participle, Imperfective verb, Verbal noun, Predicative, Collective, Preposition variant, Adjective form.

Upon selection source morphological dictionary the user should load text corpora (*.txt or *.docx formats) into the database. Then starts text pre-processing step, consisting in anomalies detection and cleaning. Such lexical "noise" may contain any abnormal symbols, which should not be treated as lexical unit.

The algorithm of the Lexical Analyzer consists of the following sequence of steps (Fig. 2):

1) Extracting the morphological characteristics of PS (gender, case, person, number, declension, etc.) of each word of the text corpora from publicly available dictionaries.

2) Constructing fuzzy logic rules for analyzing the morphological characteristics of PS and classifying them as PL.
3) Tabular single-valued determination of semantic tags of words (lemmas and their word forms) according to the rules of step 2.
4) Manual expert inspection of the tabular translation and markup of certain corpora of texts (not classified in the steps 2 and 3) for further training of the neural network.
5) Rule-based learning and training of fuzzy sets to recognize Parts of Language in order to generalize the fuzzy rules (step 2) to large volumes of texts and to solve lack of morphological characteristics and corresponding rules problems.
6) The final step of the algorithm is the creation of a tag-semantic dictionary (assigning a semantic tag to each lexical unit of the subject domain).

As the architecture, a fuzzy neural network NEFClass M (Modified Neuro Fuzzy Classifier) based on the generalized fuzzy perceptron architecture using the numerical optimization algorithm of the Gaussian belonging function and adapted to the existing task of semantic markup was chosen (Fig. 3). Fuzziness arises from incomplete measurements of object properties (absence or incompleteness of morphological characteristics of words in publicly available dictionaries).

The fuzzy rules describing the data are of the following form: The input layer of the network is a matrix of PS and morphological characteristics obtained from existing dictionaries. The hidden layer is a set of fuzzy rules for converting PS and morphological characteristics into PL. Finally, the output layer of the network is the resulting class of semantic tag according to the PL paradigm.

if *x1* is $\mu 1$ **and** *x2* is $\mu 2$ **and** ... **and** *xn* is μn,

then the sample *(x1, x2, ... , xn)* belongs to type *i*,

where μi, ... , μn are fuzzy sets.

For example,

if the word = *NOBODY* as SPEECH PART = *pronoun* AND morphological characteristic = *negative* **AND** morphological characteristic = *singular*

then the example (*NOBODY*) has tag = *taigen contracted physical variable qualitative arbitrary narrative non-personal multi-place intensive*
(code 1 0 0 0 0 0 0 0 -1 -1 0 1 -1 0 -1 0 -1 0 0 0 0)

V. Realization and Results

The WEBSEMDICT program is designed to work with the morphological and semantic dictionaries of Russian words and to search for these words in selected text corpora.

The functions consist of selecting articles, dissertation abstracts or small books and further analyzing their contents using selected dictionaries by the user. The downloaded text is placed in the text area on the site, possible noises are indicated, the name of the document and the text is converted into a list of unique, non-repeating sorted words. When saving the obtained results, the words have been searched in the database and marking is made about the found words considering their occurrence.

The resulting list of words with their morphological characteristics can be viewed in the corresponding table, and these same characteristics can be modified. The tag-semantic dictionary table directly indicates the semantic characteristics of these words, which can be edited by the expert if necessary. The user's personal account is represented by 5 data blocks that contain required personal information (name, role, email, phone, etc.). The site has a feedback mechanism with the developer, where the user can ask a question and get a prompt answer for support.

The registered users and administrators have access to the following functionality:

1) Working with text corpora;
2) Access to the morphological dictionary, displayed as a table;
3) Access to the tag-semantic dictionary, displayed as a table;
4) The ability to contact the site administrator with a question by clicking on the "Support" button;
5) Access to your personal account, with the ability to install a two-factor protection system, correct the data entered during registration, as well as change the email and password. Access to deleting an account is disabled.
6) Access to unrecognized words that are not in the database. You can also do personal work with them by eliminating errors in the words themselves, indicating their belonging to the basic form (lemma) or to the word form, as well as in dicating their morphological characteristics.

VI. Realization and Results

The WEBSEMDICT software implementing the described method is developed as a Web-application based on microservice client-server architecture using ASP.Net Core MVC tools. The server part of the software is written in the high-level compiled software language C# under .NET Core Version 6.0. The client part is written in the interpreted software language JavaScript, markup language HTML-5, and using CSS-3, Bootstrap, Razor-pages. The database is built on MS SQL [10].

At this stage of the project, the whole database contains 312,546 lemmas (base words) and 2,469,893 word forms. For this purpose, collections of texts in the subject domain 'optoelectronics' consisting of 113 text files in *.txt and *.docx format with a total volume of extracted 34,605 words were used. As a result of the software's functioning, the texts were cleaned from 'noise' with

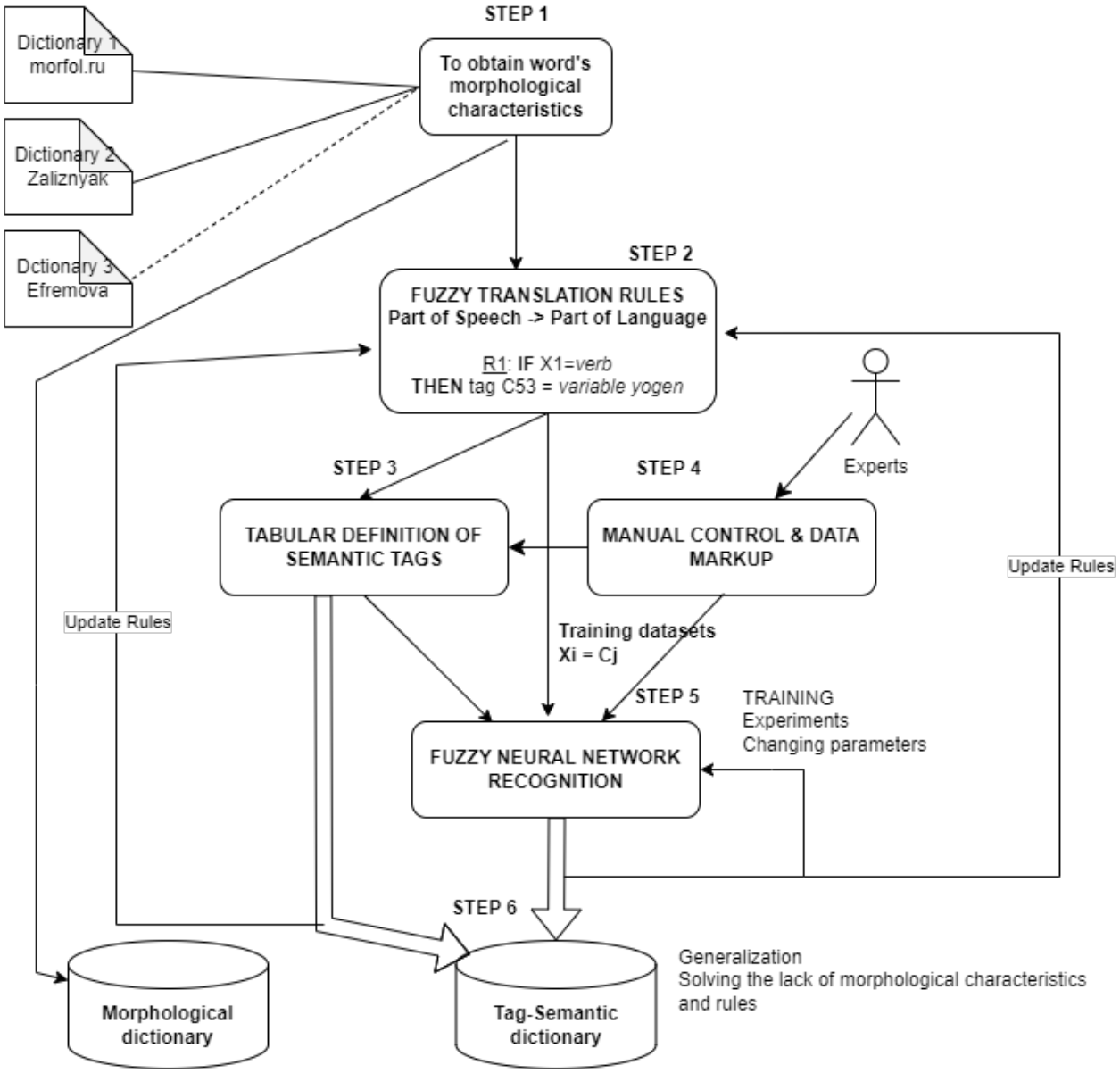

Figure 2. The algorithm for semantic markup and tag-semantic dictionary creation

the subsequent creation of a morphological dictionary containing *8,948* unique words and 25,655 word forms having morphological characteristics. The morphological dictionary contains 57 unique characteristics. On the basis of TAPAZ technology, 98 fuzzy rules for converting Parts of Speech into Parts of Language have been developed with the ability to assign a compound semantic tag to each word in the dictionary (360 unique tags in total). After applying these fuzzy rules, a tag-semantic dictionary consisting of 5,224 lemmas and 20,152 word forms with semantic characteristics was formed.

In the future, it is planned to increase the Tag-Semantic Dictionary loading large texts of different subject domains with its further automatic processing by neural network combined with markup on the basis of PS-PL rules.

VII. Conclusion

A new approach to the construction of TAPAZ-based intelligent systems for the tasks of semantic processing of natural language texts is based on the classification of lexical units not on their frequency-statistical characteristics, as in algorithms of generative artificial intelligence, but on the basis of their semantic relations as PL and syntactic roles in a sentence. A hybrid method of tagging sentences using PL with automated tagging and construction of the Tag-Semantic Dictionary is proposed and described, which combines the application of

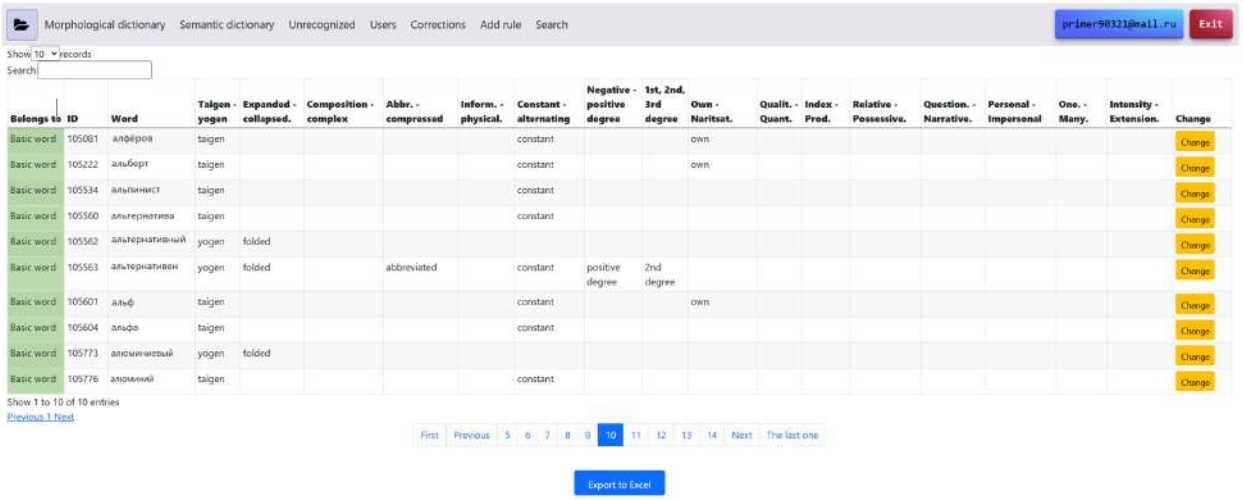

Figure 3. The Tag-Semantic Dictionary

tabular, unambiguous tagging of PS-PL by a fuzzy neural network and expert verification.

In terms of scaling our research results, determining the relevance of the training domain, and selecting adaptive machine learning algorithms, we are very much looking forward to comprehensive cooperation with Professor V. A. Golovko, a leading Belarusian expert in the field of artificial neural networks [11, p. 81–101]. We are also actively building up cooperation with Professor V. V. Golenkov within the framework of Open Semantic Technologies for Intelligent Systems (OSTIS Technology) [12, p. 81–98], which is focused on the development of hybrid intelligent systems. There are already good examples of OSTIS knowledge bases describing lexical, syntactic and semantic categories in terms of TAPAZ and available for implementation in Natural Language Processing and Understanding algorithms [13, p. 183–188], [14, p. 123–140], [15, p. 192–221]. OSTIS technology, with its original object-oriented high-level software language, is, in our opinion, the most suitable for immersing TAPAZ algorithms in a software environment.

This work is a significant step towards the promotion of a hardware-software complex for semantic processing of natural language texts, understanding of natural language and solving inventive problems in the forward and backward direction from the top-level ontology defined structurally to the semantic, syntactic and lexical analysis of any technical and technological descriptions in natural language [16, p. 176–197]. The research contributes to the creation of national artificial intelligent systems based on combinatory or knowledge-based methods.

References

[1] Journal Rankings. SJR. Scimago Journal and Country Rank. Available at: https://www.scimagojr.com/journalrank.php?type=j (accessed 21.12.2022).

[2] Hardzei A. Theory for Automatic Generation of Knowledge Architecture: TAPAZ-2. Transl. from Rus. I. M. Boyko. Rev. English edn. Minsk, The Republican Institute of Higher School Publ., 2017, 50 p. Available at: http://tapaz.by (accessed 18.03.2025).

[3] Activity Vocabulary. W3C Recommendation. Available at: https://www.w3.org/TR/activitystreams-vocabulary/#dfn-activity (accessed 20.07.2023).

[4] Hardzei A. More About the Semantic Approach to NLP Problem Solving. Inostrannye yazyki v vysshej shkole [Foreign Languages in Tertiary Education]. 2023; 2(65):5–18. (In Russian) DOI: 10.37724/RSU.2023.66.3.001.

[5] Hardzei A. The Paradigm of Parts of the Language. Materially VIII Mezhdunarodnoy konferentsii “Slovoobrazovaniye i nominativenaya derivatsiya v slavyanskikh yazykakh” [Proceedings of VIII International Research Conference “Word-Building and Nominative Derivation in Slavik Languages”]. Grodno, Grodno State University Publ., 2003, pp. 173–179. (In Russian).

[6] Hardzei A. Parts of Language and the Procedures of Its Delineation. Puti Podnebesnoi [The Paths of the Middle Kingdom]. Minsk, Belarusian State University Publ., 2006, iss. 1, pt. 1, pp. 69–75. (In Russian).

[7] Hardzei A. Metasemantics of Language Categories. Vtorye chneniya, posvyshchennye pamyati professor V. I. Karpova [Second Readings Dedicated to the Memory of Professor V. A. Karpov]. Minsk, Belarussian State University Publ., 2008, pp. 19–24. (In Russian).

[8] Hardzei A. Semantic Markup of the Event and its Display by Means of the Chinese and Russian Languages. Inostrannye yazyki v vysshej shkole [Foreign Languages in Tertiary Education]. 2021;2(57):5–26 (In Russian). DOI: 10.37724/RSU.2021.57.2.001.

[9] Svyatoshchik, M. The Conversion of Parts of Speech into Parts of Language. Vesnik MSLU [Vestnik MSLU. Philology]. Minsk : Minsk State Linguistic University Publ., vol. 1(122), 2023, pp. 29–36. (In Russian).

[10] Hardzei A., Panashchik R., Svyatoshchik M., Strelchenok O., Tkachenko V. Algorithm of Automated Semantic Markup of Texts. Development of Informatization and State System of Scientific and Technical Information (RINTI-2024): Reports of the XXIII International Conf., Minsk, 21 November 2024. Minsk : OIPI NAS of Belarus, 2024, pp. 400–404. (In Russian).

[11] Golovko, V. Neuro-Symbolic Artificial Intelligence: Application for Control the Quality of Product Labeling / V. Golovko, A. Kroshchanka, M. Kovalev, V. Taberko, D. Ivaniuk. Communications in Computer and Information Science (CCIS). Switzerland : Springer Nature Switzerland AG, 2020, vol. 1282, pp. 81–101.

[12] Golenkov V., Gulyakina N., Grakova N., Davydenko I., Nikulenka V., Eremeev A., Tarasov V. From Training Intelligent Systems to

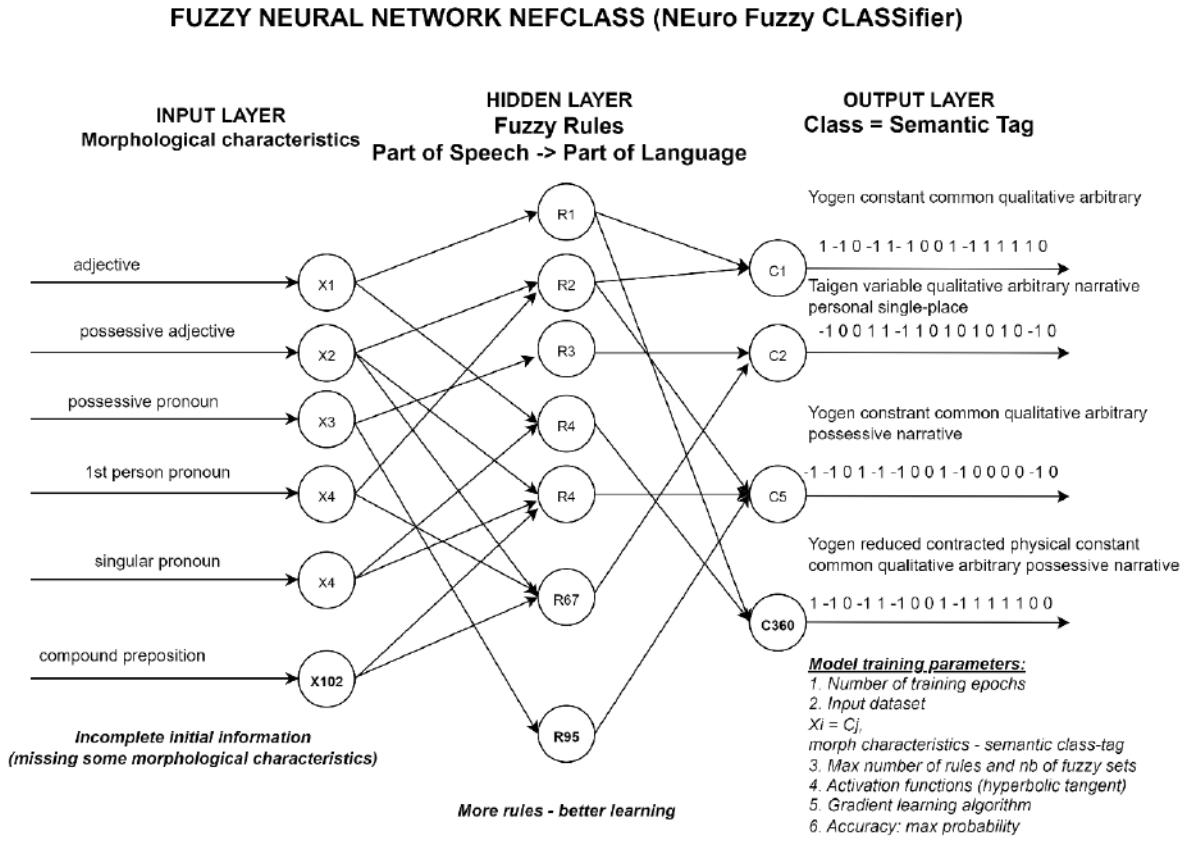

Figure 4. Fuzzy neural network NEFClassy

Training Their Development Tools. Open Semantic Technologies for Intelligent Systems (OSTIS) : Conference. Minsk : Belarussian State University of Informatics and Radioelectronics Publ., 2018, iss. 2, pp. 81–98.

[13] Hardzei A., Krapivin Y. 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. Golenkov V. V. et al. (eds.). Open Semantic Technologies for Intelligent Systems (OSTIS). Minsk, Belarussian State University of Informatics and Radioelectronics Publ., 2020, iss. 4, pp. 183–188.

[14] Hardzei A., Svyatoshchik M., Bobyor L., Nikiforov S. Processing and Understanding of the Natural Language by an Intelligent System. Open Semantic Technologies for Intelligent Systems. Minsk, BSUIR Publ., 2021, iss. 5, pp. 123–140.

[15] Hardzei A., Svyatoshchik M., Bobyor L., Nikiforov S. Universal Semantic Markup and Top-level Ontology Representation. Communications in Computer and Information Science (CCIS). Switzerland, Springer Nature Switzerland AG, 2022. Vol. 1625. Pp. 192–221. Available at: https://link.springer.com/chapter/10.1007/978-3-031-15882-7_11#citeas (accessed 18.03.2025).

[16] Hardzei A. Plagiarism Problem Solving Based on Combinatory Semantics. Communications in Computer and Information Science (CCIS). Switzerland, Springer Nature Switzerland AG, 2020, vol. 1282, pp. 176–197. Available at: https://link.springer.com/book/10.1007%2F978-3-030-60447-9 (accessed 18.03.2025).

АВТОМАТИЧЕСКАЯ ОБРАБОТКА РУССКИХ ЛЕКСИЧЕСКИХ ЕДИНИЦ И СОЗДАНИЕ ТЕГО-СЕМАНТИЧЕСКОГО СЛОВАРЯ ВЫДЕЛЕННОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

Гордей А. Н., Панащик Р. С.,
Святощик М. И., Стрельченок О. А.,
Ткаченко В. В. , Шумилин А. Г.

В статье представлены методология и алгоритм разметки лексических единиц русскоязычных текстов в технологии комбина́торной семантики. л автоматиаии еревода асте реи в асти ка редлоенравила неетко лоики и неетко неронной сети.

Received 23.03.2025