ПРИКЛАДНА ЛІНГВІСТИКА

UDC 811.101
DOI https://doi.org/10.52726/as.humanities/2024.3.6

M. V. NADUTENKO
Candidate of Technical Sciences in Structural, Applied, and Mathematical Linguistics, Head of Informatics Department, Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine, Secretary of Scientific Council "Information. Language. Intellect" of the National Academy of Sciences of Ukraine, Kyiv, Ukraine
E-mail: hostmaster@ulif.org.ua
https://orcid.org/0000-0001-6732-8455

M. V. NADUTENKO
Candidate of Philological Sciences, Senior Researcher,
Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine, Kyiv, Ukraine
E-mail: margo.nadutenko@gmail.com
https://orcid.org/0000-0003-3215-0970

O. L. FAST
Candidate of Sciences (Pedagogics), Associate Professor,
Vice-Rector for Research, Teaching and International Relations,
Municipal Higher Educational Institution "Lutsk Pedagogical College" of the Volyn Regional Council, Lutsk, Ukraine
E-mail: ofast@lpc.ukr.education
https://orcid.org/0000-0001-7216-0044

INNOVATIVE LANGUAGE TECHNOLOGIES AND ARTIFICIAL INTELLIGENCE: DEVELOPMENT AND APPLICATIONS BY THE POLYHEDRON PLATFORM BASED ON UKRAINIAN LEXICOGRAPHIC THEORIES

The article presents a comprehensive approach to the creation and implementation of intelligent language technologies developed by the authors of POLYHEDRON, which are based on the fundamental lexicographic theories devised by the Ukrainian Lingua-Information Foundation of the National Academy of Sciences of Ukraine. The authors focus on the POLYHEDRON technology family, which encompasses tools for lexicographic and corpus-based text processing, systems for parsing files in various formats, modules for morphological and semantic analysis, as well as innovative means for constructing dynamic ontologies and decision support systems. This enables a multilayered processing of natural language (Ukrainian, English, Russian, French, German, and Italian) and the creation of scalable information resources geared toward both scientific and practical applications.

The central element of the study is a hybrid architecture that combines statistical methods (deep neural networks of the transformer type) with lexicographic-ontological models. This approach allows for the effective simultaneous analysis of both the syntactic and semantic structure of sentences, the revelation of latent relationships between terms and concepts, and the formation of dynamic ontologies that are continuously updated based on new textual data. The work especially emphasizes the role of dynamic knowledge compression technologies, which ensure optimized storage and processing of information, thereby enabling the use of models with a smaller footprint without a loss in analytical accuracy.

One of the key application areas for the described technologies is the automated monitoring and analysis of large volumes of text, including legal documents, scientific and technical publications, media materials, etc. For this purpose, dedicated parsing subsystems (MxParse, MxDocArch, OCR modules) have been developed that support DOC, PDF, TXT, HTML, and other formats, as well as recognize scanned images and audio/video files. The "AVALANCHE" and "INVISIBLE" technologies, developed by the authors, play an important role by providing unique capabilities for the rapid indexing and retrieval of data in multilingual corpora: the former is responsible for the persistent storage of billions of objects on disk, while the latter efficiently manages large structures in volatile memory.

The article also introduces the INTELLIGENCE-ANALYTICS platform, which integrates three key components: a neural network, an ontological module, and multi-criteria decision analysis (MCDA) mechanisms. This integration makes it possible to identify non-obvious relationships between documents, prioritize among various alternatives, and generate flexible analytical reports to support real-time decision-making. Promising application areas include national security, education, legal expertise, scientific research, and information management in large organizations.

The authors emphasize the critical importance of developing proprietary national language models, particularly for the Ukrainian language, that can be competitive with foreign counterparts. The proposed concepts demonstrate that the integration of a lexicographic-ontological approach with modern neural algorithms maintains the quality of processing while reducing model size and computational resource requirements. This is particularly relevant given the limitations of targeted funding and infrastructural challenges.

Thanks to the conducted research and collaboration with partners from scientific and educational institutions, unique developments have been achieved that are capable of accelerating the digitization of documents, supporting the development of high-tech products in Ukraine, and strengthening information security. The article underlines the need for consolidating scientific and technological potential, establishing public–private projects, and expanding cooperation between NAS Ukraine institutions, universities, and the private sector. The authors see this as the main impetus for creating an integrated language-information ecosystem capable of addressing the intellectual challenges of modernity and stimulating scientific and technological progress in the country.

Key words: artificial intelligence methods; natural language processing; lexicographic technologies; dynamic ontologies; morphological and semantic analysis; neural networks; multi-criteria decision analysis; POLYHEDRON.

Problem statement. The issue of developing and implementing artificial intelligence methods in linguistically oriented areas of information technologies has traditionally been the focus of attention of the Ukrainian Language and Information Foundation of the National Academy of Sciences of Ukraine (ULIF). Among the linguistic scientific institutions and centers of Ukraine, UMIF thoroughly embodies its achievements in computer linguistic technologies and systems, ensuring the functioning and access of Internet users to more than 60 electronic dictionary and other language and information systems, as well as objects of the National Dictionary Database (1, 2, 3, 4, 5).

Intellectual properties of language and development of corresponding models and technological tools in this field continue to be among the main directions of scientific activity of ULIF. As the further course of events showed, the above-mentioned scientific approaches and results obtained at ULIF turned out to be extremely effective in the field of artificial intelligence and recently demonstrate great relevance and prospects. The mentioned issues, additionally actualized by the situation around the national security and defense of the country, were heard at the meeting of the Presidium of the NAS of Ukraine when considering the issue "Linguistic dimensions of the problems of national security and defense of Ukraine".

The development and implementation of the national methods, means and technologies of intellectual operation of language and information processes is vital for Ukraine becoming a challenge to its national security.

Particular attention should be paid to the need for the national information and software solutions in this area, for which we propose to use the family of information and linguistic software platforms "POLYHEDRON"[1] based on the principles of natural language and created at the National Academy of Sciences of Ukraine. These tools are capable of providing a modern, highly efficient and high-tech basis for linguistic support of the legal and regulatory process and helping to overcome numerous systemic shortcomings of the domestic legal and regulatory base under the following conditions: comprehensive application based on intellectualization, virtualization and Big Data; the intellectual substrate of which should be Large Language Models; created on the instrumental basis of artificial neural networks.

Research objectives: to highlight the main methods of artificial intelligence, which have been developed and practically implemented by the Ukrainian Language and Information Fund of the NAS of Ukraine; to characterize the technological groups of the POLYHEDRON technol-

[1] O. E. Stryzhak, V. Yu. Velychko, V. V. Prykhodniuk, M. V. Nadutenko, V. V. Gorborukov, O. V. Franchuk, S. O. Dovgiy, O. V. Lisovyi, Palagin O. V., Sergeyko I. V., Shirokov V. A., Romanenkov Yu. O., Shostak I. V., Potapov H. M., Filisteev D. A., Chepkov I. B., Gupalo A. .Y., Globa L. S., Gladun O. V., Popova M. A., Shevchenko L. L., Malakhov K. S. Computer program "**Cognitive IT platform POLYHEDRON (CIT POLYHEDRON)**". Certificate of copyright registration No. 96078 (2020).

O. E. Stryzhak, V. Yu. Velychko, V. V. Prykhodnyuk, **M. V. Nadutenko**, *V. V. Gorborukov, O. V. Franchuk, O. V. Palagin, V. A. Shirokov, Potapov H.M., Malakhov K.S., Kaverynskyi V.V., Dovgiy S.O., Lisovyi O.V., Chepkov I.B.* **Computer program "POLYHEDRON – Transdisciplinary Analysis" ("POLYHEDRON-TA"). Certificate of copyright registration** *No. 113978 (2022).*

ogy family and their capabilities in the formation of multi-layer linguistic corpora, which technologically facilitates decision-making processes.

Results and Discussion.

FAMILY OF TECHNOLOGIES "POLYHEDRON"[2]

Some of the work in this area is carried out in cooperation with the V. M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine (O. V. Palagin, National Academy of Sciences of Ukraine) and the Scientific Center "Junior Academy of Sciences" of the Ministry of Education and Science of Ukraine and the National Academy of Sciences of Ukraine (O. E. Stryzhak, Doctor of Technical Sciences). The vast majority of this work is concentrated around the POLYHEDRON family of information and linguistic technologies created in the process of this cooperation, which was discussed above.

1. Lexicographic technologies

Search technologies. Technologies for working with web protocols and parsing web resources. System architecture: based on the technology of lexicographic environments and agents (L-agents). It is: a system of L-agents: a leader of the crawler working group, a set of crawler templates, a system for managing crawler working environments, a set of L-agents of format processors (HTML, text, xml, json, SiteMap, documents in various formats via an interface to "MxDocArch" and "MxParse" (mht, rtf, pdf, doc, docx, odt, xls, xlsx, csv, DjVu, epub, fb2, fb3, txt, msg, zip, rar). Additional technologies used: technologies of universal document architecture "MxDocArch" (see below); technologies of parsing file formats "MxParse" (see below); technologies of determining the language of the text (based on dictionaries of N-grams in ontological form (dictionaries of n-garam of 82 languages from WikiPedia); technologies of determining the encoding of the text (based on dictionaries of encoding samples (encoding sample dic); technologies of optical recognition of documents OCR.

Assignment: automatic collection of documents in various formats from various remote sources, Internet sites, etc. Finding the necessary data sources on the Internet (based on solving a multi-criteria optimization problem) and building optimal ways to bypass these sources by a crawler. The possible multi-criteria optimization for crawler path selection is defined as:

$$\min_{P\in\mathcal{P}} \sum_{i=1}^{n} c_i\,(P), \quad \text{subject to constraints } G_j(P) \le 0,\ \forall j,$$

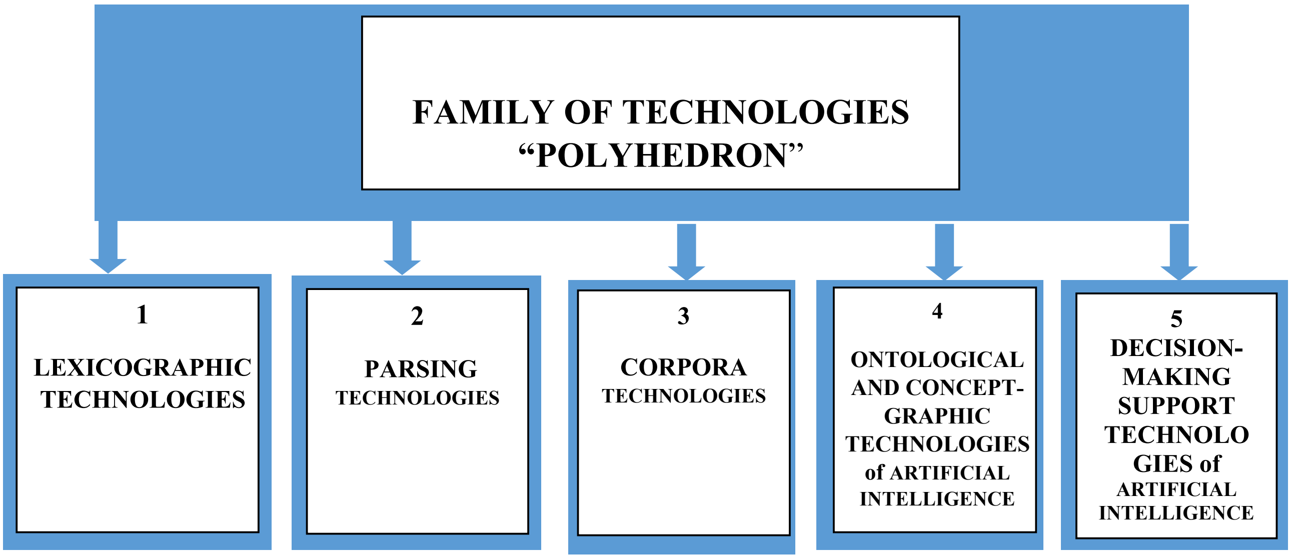
FAMILY OF TECHNOLOGIES "POLYHEDRON"
1 LEXICOGRAPHIC TECHNOLOGIES
2 PARSING TECHNOLOGIES
3 CORPORA TECHNOLOGIES
4 ONTOLOGICAL AND CONCEPT-GRAPHIC TECHNOLOGIES of ARTIFICIAL INTELLIGENCE
5 DECISION-MAKING SUPPORT TECHNOLO GIES of ARTIFICIAL INTELLIGENCE

Fig. 1. Technological groups in the structure of the POLYHEDRON technology family

where:

P is a path or a set of paths for the crawler,

$c_i(P)$ is a cost function (e.g., time or resource consumption) for path P,

[2] O. E. Stryzhak, V. Yu. Velychko, V. V. Prykhodniuk, M. V. Nadutenko, V. V. Gorborukov, O. V. Franchuk, S. O. Dovgiy, O. V. Lisovyi, Palagin O. V., Sergeyko I. V., Shirokov V. A., Romanenkov Yu. O., Shostak I. V., Potapov H. M., Filisteev D. A., Chepkov I. B., Gupalo A. Y., Globa L. S., Gladun O. V., Popova M. A., Shevchenko L. L., Malakhov K. S. Computer program "**Cognitive IT platform POLYHEDRON (CIT POLYHEDRON)**". Certificate of copyright registration No. 96078 (2020).

$G_j(P)$ are constraint functions (e.g., bandwidth limits).

Potential fundamental and applied thematic areas: Research based on the theory of lexicographic systems of general methods for parsing lexicographic systems and cognitive generalization of their structures.

2. Parsing technologies

2.1. File format parsing technologies "MxParse" and universal document architecture technologies "MxDocArch"

System architecture: based on a modular approach. It is: a universal document architecture interface, procedures for working with document architecture and a set of document processing modules in various formats (mht, rtf, pdf, doc, docx, odt, xls, xlsx, csv, DjVu, epub, fb2, fb3, txt, msg, zip, rar). It is part of the subsystem: POLYHEDRON-Crawler.

Assignment: automatic selection of text and other structural data (sections, pages, headings, subtitles, tables, lists, figures, styles, footnotes, etc.) from documents in various formats for presentation in a single document architecture for the purpose of further processing.

Possible fundamental and applied thematic areas: Research based on the theory of lexicographic systems of general methods for parsing lexicographic systems and cognitive generalization of their structures.

2.2. Optical Document Recognition (OCR) Technologies

System architecture: based on a modular approach. It consists of: a universal document architecture interface and a set of modules for processing scanned documents in PNG, TIFF, JPEG, PDF and other formats (using the Tesseract Open Source OCR Engine library) in combination with original development of integration modules and original recognition dictionaries in the Ukrainian language.

Included in the subsystem: POLYHEDRON-Crawler.

Additional technologies used: technologies for determining text encoding (based on dictionaries of encoding samples (encoding sample dic), technologies for determining text language (based on dictionaries of N-grams in ontological form (dictionaries of n-garam of 82 languages from WikiPedia).

Assignment: automatic selection of text and other structural data (pages, headings, subtitles, tables, lists, figures, styles, footnotes, etc.) from scanned documents in PNG, TIFF, JPEG, PDF and other formats for presentation in a single document architecture for further processing.

Possible fundamental and applied thematic areas: Research based on the theory of lexicographic systems of general methods for parsing lexicographic systems and cognitive generalization of their structures.

2.3. Speech Recognition Technologies (under development).

System architecture: based on a modular approach

Represented by: a set of modules for processing sound and video files in WAV, MP3, MP4, WMA, OGG and other formats (using the OpenSpeech library) in combination with the originally developed integration modules and original dictionaries and recognition models in Ukrainian language. Included in the subsystem: POLYHEDRON-Crawler.

Additional technologies used: technologies for building language models using accent dictionaries from individual languages (based on dictionaries in ontological form).

Assignment: automatic extraction of text and other structural data from sound and video files for presentation in a single document architecture for the purpose of further processing.

The probabilistic approach to OCR or speech recognition is defined as:

$$P(\text{recognized} \mid \text{input}) =$$
$$= \arg\max_{w \in W} \prod_{i=1}^{m} p_i\,(w_i \mid \text{input}),$$

where:

input is the image or audio signal,

w is a sequence of words (or characters),

p_i is the probability model for each segment i.

Possible fundamental and applied thematic areas: Research based on the theory of lexicographic systems of intelligent speech-information interfaces.

3. Linguistic Corpora Technologies

3. 1. Technologies of linguistic-semantic Analysis and Indexing

System architecture: based on the technology of lexicographic agents (L-agents) and environments).

Represented by: a system of L-agents: a leader of the indexer working group, a set of indexer templates, a system for managing indexer working environments, a set of L-agents of processors of structural analysis, morphological analysis, lexical analysis, semantic analysis, a multi-index database management agent, virtual working environments, virtual storage (using unique Persistent Storage Technology, called «AVALANCHE»), virtual memory addressing (using unique In-Memory Storage Technology, called «INVISIBLE»). It is part of the subsystem: POLYHEDRON-Indexer.

Software libraries: .NET Framework 4.7–4.8, Net Core 5.0–6.0, ICU 70 Unicode 14 graphematic analysis modules, tokenization modules (author's development), pre-morphanalysis modules (author's development), syntactic analysis modules (author's development), concept analysis modules (author's development).

Dictionary systems: morphological dictionaries of Ukrainian, Russian, English languages (in Hunspell format and binary format) – author's development, C# (Ukrainian Language and Information Fund (ULIF) of the National Academy of Sciences of Ukraine[3]. ULIF of the NAS of Ukraine is one of the developers of dictionary systems for the "Ruta-Play" system.

Additional technologies used: technologies for building language models using accent dictionaries of every single language (based on dictionaries in ontological form); technologies for automatic analysis of synonym equivalents of language constructions in Ukrainian, English, Russian, French, German and Italian languages, unique Persistent Storage Technology "AVALANCHE", unique In-Memory Storage Technology "INVISIBLE".

Assignment: automatic morphological and semantic analysis and indexing of texts in natural language (Ukrainian, English, Russian, French, German and Italian) with full-text search capabilities in the specified languages for the purpose of further processing; presentation of the syntactic and semantic structure of sentences; automatic selection of contexts in which multi-word terms are used; selection of given semantic relations; formation and display of dynamic thematic classifiers, catalogs and registers of information resources; (indexing of information arrays; Thematic classification of information arrays and documents; automatic creation of dynamic thematic classifiers, catalogs and registers of information resources; aggregated representation of information resources); automatic definition of contextual connections between terms from different documents. This allows to reveal hidden connections between objects and events that are described in different documents of the POLYHEDRON database.

The morphological tagging approach is expressed as:

$$\mathrm{Tag}(w) = \underset{t\in T}{\mathrm{argmax}} P(t \mid w,\mathrm{context}),$$

where:

w is a token,

T is the set of possible tags,

$P(\cdot)$ is the probability derived from a morphological or semantic model.

Possible fundamental and applied thematic areas: integration of corpora L-systems based on the theory of lexicographic environments; creation of virtual terminology systems and dictionaries by fields of knowledge (creation of the necessary terminological multilingual dictionaries and dictionaries of named entities in ontological format); development of an integrated dictionary base for intellectual cognitive analysis (creation of the necessary integrated semantic-grammatical multilingual dictionaries in ontological format).

Unique Persistent Storage Technology «AVALANCHE»

Represented by: developed by the authors algorithm for omitting data from sheets and custom caching, along with variable-length data encoding (speed is 100 times faster than other NOSQL solutions and regular B+-trees).

Assignment: to permanently store billions of objects (with complex structure) on disk. It is part of the subsystem: POLYHEDRON-Indexer.

3.3. Unique In-Memory Storage Technology «INVISIBLE»

Represented by: a progressive variation of the lz4 algorithm in combination with variable-length data encoding, plus invisible to the system garbage collector (access speed is higher than other solutions (such as MemCache, Redis, etc.).

Assignment: To permanently store billions of objects (with a complex structure) in memory

[3] https://lcorp.ulif.org.ua/dictua/ , https://lcorp.ulif.org.ua/.

for a long time). It is part of the subsystem: POLYHEDRON-Indexer.

3.4. Automatic Analysis of Synonym Equivalents of Language Constructions Technologies in Ukrainian, English, Russian, French, German and Italian Languages

System architecture: based on a modular approach. Presented by: a set of processing modules. Included in the subsystem: POLYHEDRON-Indexer.

Dictionary systems: dictionaries of synonymous equivalents of Ukrainian, English, Russian, French, German and Italian languages – author's development, C# (Ukrainian Language and Information Fund of the National Academy of Sciences of Ukraine).

Assignment: for automatic synonymous analysis of equivalents of language structures in Ukrainian, English, Russian, French, German and Italian. Used for semantic translation of search queries and when comparing multilingual lexical structures.

The simplified synonym matching criterion is defined as:

$$\mathrm{SynMatch}(u,v)=\begin{cases}1, & \text{if } \mathrm{SemDist}(u,v)\leq\delta,\\ 0, & \text{otherwise,}\end{cases}$$

where:

SemDist(u, v) is a semantic distance function between two lexical items u and v,

δ is a matching threshold.

Possible fundamental and applied thematic areas: integration of corpus L-systems based on the theory of lexicographic environments; creation of virtual terminology systems and dictionaries by branches of knowledge (creation of the necessary terminology multilingual dictionaries and dictionaries of named entities in ontological format); development of an integrated dictionary database for intellectual cognitive analysis (creation of the necessary integrated semantic-grammatical multilingual dictionaries in ontological format).

4. Ontological and Conceptographic Technologies of Artificial Intelligence

4.1. Conceptographic cognitive technologies. Knowledge Understanding and Dynamic Ontologies Generation technology.

System architecture: based on a modular approach.

Represented by: a set of processing modules. Included in the subsystem: Conceptographic analysis system. Additional technologies used: GT-DMX-PQC Generative Transformer – Dynamic MiXed Precision&Quantization&Compression.

Dictionary systems: dictionaries of synonymous equivalents of Ukrainian, English, Russian, French, German and Italian languages – authors' development, C# (ULIF); dictionaries of named entities (authors' development), C#.

Assignment: automatic conceptographic analysis of natural language text documents in Ukrainian, Russian and English, which includes: means of automatic detection of named entities and semantic relations and means of automatic establishment of key semantic concepts of the text of the subject area using dictionary, corpora and machine learning technologies, generation of dynamic ontologies.

Possible fundamental and applied thematic areas: integration of corpus L-systems based on the theory of lexicographic environments; creation of virtual terminology systems and dictionaries by branches of knowledge (creation of the necessary terminological multilingual dictionaries and dictionaries of named entities in ontological format).

Development of an integrated dictionary base of intellectual cognitive analysis (creation of the necessary integrated semantic-grammatical multilingual dictionaries in ontological format).

4.2. GT-DMX-PQC Generative Transformer – Dynamic MiXed Precision & Quantization & Compression

System architecture: based on lexicographic agent (L-agent) and environment technology. Included in the subsystem: Conceptographic analysis system.

Software libraries: .NET Framework 4.7–4.8, .Net Core 5.0–6.0.

Represented by: Generative Transformer – Dynamic MiXed Precision&Quantization&Compression (GT-DMX-PQC) is a deep neural network based on the architecture of a transformer decoder (and encoder, if necessary), which supports dynamic tensor operations using MKL and CUDA, dynamic precision mixing for operations with FP32, FP16 and Int8 and dynamic quantization for tensor manipulation, dynamic knowledge compression and outlier elimination technology. GT-DMX-PQC also uses SentencePiece or ICU tokenization, Adam optimization, attention equalization and a gradient distribution and synchronization mechanism between multiple GPUs and CPUs.

Dynamic knowledge compression is a technology that allows you to extract the structure of concepts from any text and compare it with the concepts of the base and dynamically extended ontology, as a person does with a dictionary. Existing static base knowledge will be compressed as much as possible. Uncertainties will be interpreted with a lower level of compression depending on the required level of detail.

Assignment: Support for the processes of understanding and knowledge generation for the generation of dynamic ontologies (automatic detection of named entities and semantic relations and means of automatically establishing key semantic concepts of the text of the subject area).

4.3. Dynamic Ontologies

We use a transformer generating network to automatically generate a structure over a text (or context) based on our own, developed at the ULIF system-linguistic approach, according to which:

- The text is interpreted as a Substance.
- The formed Structure corresponds to the Subject's point of view, and expands and complements it. The set of Subject's points of view is given in the form of basic ontologies.
- The basic ontology is linguistic and is created using lexicographic technologies. The main sources of basic ontologies are explanatory dictionaries and terminological dictionaries. Using the general laws of constructing ontologies based on dictionaries, it is possible to automatically build dynamic ontologies from any texts. Thus, from texts (represented in the form of ontologies) it is possible to automatically build a dynamic picture of the world using dynamic compression of knowledge.

Consequently, dynamic knowledge compression is a linguistic and information technology that allows you to expound the structure of concepts from any text and compare it with the concepts of the basic and dynamically extended ontology, like a person does with an explanatory dictionary.

The concept-scoring approach is defined as:

$$\text{ConceptScore}(c) = \sum_{i=1}^{k} \alpha_i \cdot \text{freq}(c, D_i),$$

where:

c is a concept,

freq(c, D_i) is its frequency (or relevance) in document D_i,

α_i are weighting coefficients.

Existing static knowledge base will be compressed as much as possible. Uncertainties will be interpreted with a lower level of compression depending on the required level of detail.

Dynamic ontologies and dynamic knowledge compression are key features of our hybrid lexicographic approach to artificial intelligence. In this way, it is possible to achieve AI parameters that are no worse than those known (GPT class, etc.) using an order of magnitude smaller neural networks and a set of dynamic ontology constructors and dynamic knowledge compressors.

4.4. Control Ontology Technology

System architecture: based on a modular approach.

Represented by: control ontology is a technological solution for formalizing a certain process. Designed for formalizing a subject area by an expert (or, if possible, automatically), focused on specific specialized software tools: structuring subsystem; display subsystem; the ontological process mapping component; display component of transdisciplinary integrated arrays; display component of GIS applications. Included in the subsystem: POLYHEDRON-Ontology.

Assignment: culturing weak and unstructured information using specialized rules specified in the format of λ-expressions; taxonomic (hierarchical) representation of the structure of information arrays presented in Ukrainian, Russian and English.

The schematic ontology set definition is expressed as:

$$\text{Ontology}(C) = \bigcup_{i=1}^{n} \{O_i \mid \phi(O_i, C) = 1\},$$

where:

C is a set of concepts,

O_i are candidate ontology elements,

ϕ is a truth function indicating which elements belong in the ontology for C.

Possible fundamental and applied thematic areas: theory of lexical-concept graphic systems and its application; creation of digital virtual terminology systems and dictionaries by fields of knowledge (creation of the necessary terminological multilingual dictionaries and dictionaries of named entities in ontological formats); development of a digital integrated virtualized dictionary base for intellectual cognitive analysis (creation of the necessary integrated semantic-grammatical multilingual dictionaries in ontological formats); research on formal principles of building dynamic ontologies based on explanatory dictionaries (hyper chains, hyper cycles, etc.).

Decision-Making Support Technologies of Artificial Intelligence

5.1. Technology of Multi-Criteria Ranking of Alternatives Represented by Ontology Objects

System architecture: Three-Tier Architecture. Represented by: ontology-oriented MCDA-system for solving decision-making problems. It is part of the subsystem: POLYHEDRON-Alternative.

Assignment: the implemented services provide the decision-maker (DPM) with a wide range of opportunities for formulating and solving decision-making problems. The information environment of these problems is formed on the basis of an ontological description of the subject area. This allows the DPM to automatically receive a list of indicators for the selected problems that characterize the given alternatives and their values. Further, the decision-making process depends on the specific goal of the DPM and involves solving the following types of problems:

- ranking of alternatives (objects, strategies, development paths, etc.); the available initial set of alternatives is arranged in such a way that the decision-maker has the opportunity to assess the impact of each option (alternative) and on this basis make the optimal decision;
- ranking of alternatives – calculation of rating scores based on the user-selected system of preferences for a set of criterion functions;
- rational choice – establishment of the best (worst) alternative taking into account the user-selected subsets of criteria;
- multi-criteria comparative analysis of alternatives based on a visual representation of solutions to decision-making problems for different subsets of criteria depending on the intentions of the decision-maker.

The multi-criteria ranking is defined as:

$$\text{Rank}(a_j) = \sum_{k=1}^{m} w_k \cdot s_{jk},$$

where:

α_j is an alternative,

s_{jk} is its score under criterion k,

w_k is the weight for criterion k.

Possible fundamental and applied thematic areas: development of models, mathematical and software for solving direct and inverse multi-criteria ranking problems.

Artificial Intelligence Platform INTELLIGENCE-ANALYTICS

INTELLIGENCE-ANALYTICS (IA) is an intellectual platform developed on the basis of the POLYHEDRON family of technologies, with a component architecture of cognitive services that implement semantic-linguistic and conceptual analysis of large volumes of information, namely: narratives, documents and tabular data, identification of logistical relationships between them, evaluation criteria and decision-making support, etc. The services of the AI platform implement cognitive functions from categorization and structuring of narratives and data to identification of criterion indicators and selection of methods to support decision-making processes. These processes are represented by the following technological chain: documents → linguistic corpus → neural network formation → semantic analysis → ontology generation → identification of evaluation criteria → generation of analytical platforms for evaluation and decision-making.

The platform functionality has a hybrid format, consisting of three circuits: neural network – ontologies – multi-criteria analysis.

Technologically, the AI platform includes a set of cognitive services capable of automatically, using machine learning, generating ontologies for each document being analyzed, as well as for groups of a large number of documents. When generating document ontologies, the AI platform detects attribute data that characterizes all objects that make up the content of documents both quantitatively and qualitatively.

Neural networks provide the detection of deep inter-contextual connections between different documents with different semantic distances. This will ensure the fixation of hierarchies between objects that make up the content of documents.

The hybrid architecture of the AI platform provides the formation of multi-layered linguistic corpora, which technologically facilitates decision-making processes.

Currently, we have developed the following groups of cognitive services:

– formation of neural networks on the subject of the content of these documents;

– linguistic-semantic analysis of documents;

– conceptographic analysis of documents;

– identification of attributes of objects that make up the content of documents;

– selection of criteria for evaluating objects that make up the content of documents from attributes;

– generation of ontologies for each document and for groups of documents;

– generation of analytical platforms for supporting decision-making processes based on the identified criteria;

– selection of decision-making methods within the conditions of the posed problem/task.

Services implemented on the AI platform provide the ODA with a wide range of opportunities for formulating and solving decision-making tasks. This allows the ODA to automatically receive a list of indicators for the selected tasks that characterize the given alternatives and their values.

The schematic representation of the hybrid AI platform is:

HybridAI(documents) = (NeuralNetwork + OntologicalModule) + MultiCriteriaAnalysis.

Conclusions. Intellectualization and virtualization of linguistic research is as important and relevant as the creation of intellectual information tools based on natural language. It is worth constantly remembering that "…intelligence is a form of individualization of systems, which is inherent in language status". And the creation of the Ukrainian version of the Large Language Model should be perceived as one of the most important tasks of the Ukrainian state in the scientific and technological sphere.

We define a new research paradigm of linguistics adequate to modern challenges as informational-evolutionary-phenomenological, referring the reader to the works for details of its content[4].

From the above, it follows that there is a need to have our own, national tools for conducting linguistic research and creating relevant intellectual technologies. As far as we know, only the POLYHEDRON family of information and linguistic technologies can serve as a prototype of such tools today.

We emphasize that all the examples of intellectual linguistic and information technologies given here were created practically without targeted funding, as part of the creative initiative of the authors of the specified developments. It should be noted, however, that the relevant theoretical and model basis for their creation was developed in accordance with the scientific plans of ULIF, and the National Center "Junior Academy of Sciences" of the Ministry of Education and Science of Ukraine.

Despite this, a number of practical tasks in the field of intellectual information processing are undoubtedly successfully solved by existing AI tools. Given the critical importance of creating and implementing AI tools in scientific research and information practice, as well as taking into account the limited state funds for conducting relevant research and development, from the above we conclude that it is necessary to consolidate the existing scientific and technological potential of Ukraine in this area.

The corresponding proposal was formulated in clause 3.2. of the resolution of the Presidium of the NAS of Ukraine dated 15.11.2023 [6], as the need to "Unify the scientific, technological, human resources and material and technical potential of the institutions of the National Academy of Sciences of Ukraine, universities and private structures in order to create an interdepartmental association (in the form of a scientific and technological corporation) on the basis of public-private partnership to provide the sphere of national security and defense of Ukraine with modern intellectual language and information means and technologies and effective interaction with international structures in this field."

[4] Software products: URL: https://central.ulif.org.ua/

BIBLIOFRAPHY

1. УМІФ: Український мовно-інформаційний фонд НАН України. URL: https://www.ulif.org.ua/about.

2. УМІФ. Проекти. URL: https://ulif.org.ua/projects

3. Програмні продукти. URL: https://central.ulif.org.ua/

4. УМІФ. Ресурси. URL: https://lcorp.ulif.org.ua/LSlist

5. Nadutenko M., Prykhodniuk V., Shyrokov V., Stryzhak O. Ontology-Driven Lexicographic Systems. Advances in Information and Communication. FICC 2022. Lecture Notes in Networks and Systems. Cham : Springer. 2022. С. 204–215. DOI: 10.1007/978-3-030-98012-2_16

6. Широков В.А. Лінгвістичні виміри проблем національної безпеки та оборони України. *Вісник* Національної академії наук України. 2024. № 1. С. 56–71. https://doi.org/10.15407/visn2024.01.056

REFERENCES

1. ULIF: Ukrainian Lingua-Information Foundation of the National Academy of Sciences of Ukraine URL: https://www.ulif.org.ua/about.

2. ULIF Projects: URL: https://ulif.org.ua/projects

3. Software Products: URL: https://central.ulif.org.ua/

4. ULIF Resources: URL: https://lcorp.ulif.org.ua/LSlist

5. Nadutenko M., Prykhodniuk V., Shyrokov V., Stryzhak O. (2022) Ontology-Driven Lexicographic Systems. Advances in Information and Communication. FICC 2022. Lecture Notes in Networks and Systems. Cham : Springer. pp. 204–215. DOI: 10.1007/978-3-030-98012-2_16

6. Shyrokov V.A. (2024) Linguistic dimensions of problems of national security and defense of Ukraine. Visn. Nac. Akad. Nauk Ukr. (1): 56–71. https://doi.org/10.15407/visn2024.01.056

М. В. НАДУТЕНКО
кандидат технічних наук зі структурної, прикладної та математичної лінгвістики, завідувач відділу інформатики, Український мовно-інформаційний фонд Національної академії наук України, секретар наукової ради «Інформація. Мова. Інтелект» Національної академії наук України, м. Київ, Україна
Електронна пошта: hostmaster@ulif.org.ua
https://orcid.org/0000-0001-6732-8455

М. В. НАДУТЕНКО
кандидат філологічних наук, старший науковий співробітник,
Український мовно-інформаційний фонд Національної академії наук України, м. Київ, Україна
Електронна пошта: margo.nadutenko@gmail.com
http://orcid.org/0000-0003-3215-0970

О. Л. ФАСТ
кандидат педагогічних наук, доцент, проректор з науково-педагогічної роботи та міжнародної співпраці, Комунальний заклад вищої освіти «Луцький педагогічний коледж» Волинської обласної ради, м. Луцьк, Україна
Електронна пошта: olhafast.ua@gmail.com
http://orcid.org/0000-0001-7216-0044

ІННОВАЦІЙНІ МОВНІ ТЕХНОЛОГІЇ ТА ШТУЧНИЙ ІНТЕЛЕКТ: РОЗРОБКА ТА ЗАСТОСУВАННЯ ПЛАТФОРМИ POLYHEDRON НА ОСНОВІ УКРАЇНСЬКИХ ЛЕКСИКОГРАФІЧНИХ ТЕОРІЙ

У статті представлено комплексний підхід до створення та впровадження інтелектуальних мовних технологій, розроблених авторами POLYHEDRON, що ґрунтуються на фундаментальних лексикографічних теоріях, розроблених Українським мовно-інформаційним фондом НАН України. Автори акцентують увагу на сімействі технологій POLYHEDRON, яке включає інструменти для лексикографічного та корпусного опрацювання текстів, системи парсингу файлів різних форматів, модулі морфолого-семантичного аналізу, а також інноваційні засоби побудови динамічних онтологій і підтримки прийняття рішень. Завдяки цьому забезпечується багаторівнева обробка природної мови (української, англійської, російської, французької, німецької та італійської) та створення масштабованих інформаційних ресурсів, орієнтованих як на науковий, так і на прикладний використок.

Центральний елемент дослідження – гібридна архітектура, що об'єднує статистичні методи (глибокі нейронні мережі типу трансформерів) із лексикографічно-онтологічними моделями. Такий підхід дозволяє водночас ефективно аналізувати синтаксичну й семантичну структуру речень, виявляти латентні зв'язки між термінами й концептами та формувати динамічні онтології, які постійно оновлюються на основі нових текстових даних. У роботі особливо підкреслено роль технологій динамічного стискання знань, які забезпечують оптимізоване зберігання та обробку інформації, даючи змогу використовувати менші за обсягом нейронні моделі без погіршення точності аналізу.

Одним із ключових напрямів застосування описаних технологій є автоматизований моніторинг та аналіз великих масивів тексту, включно з документами правового характеру, науково-технічними публікаціями, медіа-матеріалами тощо. Для цього було розроблено окремі підсистеми парсингу (MxParse, MxDocArch, OCR-модулі), що підтримують формати DOC, PDF, TXT, HTML та інші, а також розпізнають відскановані зображення й аудіо/відео-файли. Важливу роль відіграють технології «AVALANCHE» та «INVISIBLE», розроблені авторами, які забезпечують унікальні можливості для швидкої індексації та пошуку даних у багатомовних корпусах: перша відповідає за постійне зберігання мільярдів об'єктів на диску, а друга – за ефективну роботу з великими структурами в оперативній пам'яті.

У статті також представлено платформу INTELLIGENCE-ANALYTICS, що поєднує три важливі складові: нейронну мережу, онтологічний модуль і механізми мультикритеріального аналізу (MCDA). Така інтеграція дає змогу знаходити неочевидні зв'язки між документами, визначати пріоритети серед множини альтернатив, а також формувати гнучкі аналітичні звіти для прийняття рішень у реальному часі. Серед перспективних напрямів застосування – галузі національної безпеки, освіти, правової експертизи, наукових досліджень, а також інформаційного менеджменту у великих організаціях.

Автори наголошують на критичній важливості розвитку власних національних мовних моделей, зокрема україномовних, які можуть бути конкурентоспроможними з зарубіжними аналогами. Запропоновані концепції підтверджують, що укорінення лексикографічно-онтологічного підходу разом із сучасними нейронними алгоритмами дозволяє зберегти якість обробки при зменшенні розміру моделей та обчислювальних ресурсів. Це особливо актуально з огляду на обмеження цільового фінансування та інфраструктурні виклики.

Завдяки проведеним дослідженням та співпраці з партнерами з наукових і освітніх установ створено унікальні напрацювання, які здатні прискорити темпи оцифрування документів, підтримати розвиток високотехнологічних продуктів в Україні та зміцнити інформаційну безпеку. У статті підкреслено потребу консолідації науково-технічного потенціалу, формування державно-приватних проєктів і розширення партнерства між установами НАН України, університетами й приватним сектором. Автори вбачають у цьому головний імпульс до створення інтегрованої мовно-інформаційної екосистеми, здатної вирішувати інтелектуальні виклики сучасності та стимулювати науково-технологічний прогрес у країні.

Ключові слова: методи штучного інтелекту; обробка природної мови; лексикографічні технології; динамічні онтології; морфолого-семантичний аналіз; нейронні мережі; мультикритеріальний аналіз; POLYHEDRON.