Ералы Рауанұлы Советбаев – техника ғалымдарының магистірі, «Автоматтандыру және ақпараттық технологиялар» кафедрасының оқытушы, Шәкәрім университеті, Қазақстан Республикасы; e-mail: sovetbaev_eraly@shakarim.kz. ORCID: https://orcid.org/0009-0008-4529-2787.

Александр Дмитриевич Золотов – т.ғ.к., «Автоматтандыру және ақпараттық технологиялар» кафедрасының қауымдастырылған профессорі, Шәкәрім университеті, Қазақстан Республикасы; e-mail: azol64@mail.ru. ORCID: https://orcid.org/0000-0001-9751-8161.

Дмитрий Викторович Мясоедов – техника ғылымдарының магистрі, «Автоматтандыру және ақпараттық технологиялар» кафедрасының аға оқытушысы, Шәкәрім университеті, Қазақстан Республикасы; e-mail: dmitriy880@mail.ru. ORCID: https://orcid.org/0009-0002-3279-0994.

Кулькен Уалиевна Зенкович – техника ғылымдарының магистрі, «Автоматтандыру және ақпараттық технологиялар» кафедрасының аға оқытушысы, Шәкәрім университеті, Қазақстан Республикасы, e-mail: kulken_@mail.ru. ORCID: https://orcid.org/0009-0000-7759-7413.

Information about the authors

Tursynkhan Zhylkybayev* – Master of Technical Sciences, lecturer of the Department «Automatization and Information Technology», Shakarim University, Republic of Kazakhstan; e-mail: zhitosya@mail.ru. ORCID: https://orcid.org/0000-0002-8918-527X.

Yeraly Sovetbayev – Master of Technical Sciences, lecturer of the Department «Automatization and Information Technology», Shakarim University, Republic of Kazakhstan; e-mail: sovetbaev_eraly@shakarim.kz. ORCID: https://orcid.org/0009-0008-4529-2787.

Alexandr Zolotov – Candidate of Technical Sciences, Associate Professor of the Department «Automatization and Information Technology», Shakarim University, Republic of Kazakhstan; e-mail: azol64@mail.ru. ORCID: https://orcid.org/0000-0001-9751-8161.

Dmitriy Myassoyedov – Master of Technical Sciences, senior lecturer of the Department «Automatization and Information Technology», Shakarim University, Republic of Kazakhstan; e-mail: dmitriy880@mail.ru. ORCID: https://orcid.org/0009-0002-3279-0994.

Kulken Zenkovich – Master of Technical Sciences, senior lecturer of the Department «Automatization and Information Technology», Shakarim University, Republic of Kazakhstan; e-mail: kulken_@mail.ru. ORCID: https://orcid.org/0000-0002-8918-527X.

Поступила в редакцию 05.01.2026
Принята к публикации 20.02.2026

https://doi.org/10.53360/2788-7995-2026-1(21)-11

IRSTI: 28.23.11

Z.B. Sadirmekova[1,3], B.Kh. Abdygalym[1,2,4], M.A. Sambetbayeva[1,2,*], R. Taberkhan[1,2], I.A. Karbozova[5]

[1]«Q» University, 050026, Republic of Kazakhstan, Almaty, str. Baizakov 125/185,
[2]L.N. Gumilyov Eurasian National University,
010008, Republic of Kazakhstan, Astana, Satpayev str. 2.
[3]Astana IT University, 010000, Republic of Kazakhstan, Astana, Mangilik El Avenue, C1.
[4]Astana International University,
010000, Republic of Kazakhstan, Astana, Qabanbay Batyr Avenue, 8
[5]Taraz International University named after Sherkhan Murtaza,
Republic of Kazakhstan, Taraz, str.Zheltoksan 69B
*e-mail: sambetbayeva_ma_1@enu.kz

AN AI-DRIVEN THREE-STAGE FRAMEWORK FOR DYNAMIC INTEGRATION OF EMERGING CONCEPTS INTO ONTOLOGIES

Abstract: *This article examines the problem of automatic implementation of new concepts in ontology in the context of constant updating of knowledge and increasing the volume of textual data. The object of research is the process by which new concepts can be integrated into hierarchical and non-taxonomic structures of ontology, while maintaining their logical and semantic consistency. The purpose of the work is to*

create a three-stage structure to automate the search, refinement, and optimal implementation of ideas. This structure should be built on the basis of modern machine learning and natural language processing technologies.

Logical inference, contrast learning and large language models and pre-prepared language models are used as research methods. At the initial stage of the framework, larger language models are used to create formal OWL axioms and semantic connections, taking into account universal and existential logical constraints. In the second stage, contrast learning is used to refine vector representations of ideas and improve classification accuracy. The third stage is the logical verification and improvement of non-taxonomic relations with the help of ontological reasoners and external sources of knowledge.

The biomedical ontology of SNOMED CT and the corpus of Kazakh data were used for experimental evaluation. The results show significant statistical superiority over existing methods, as well as high accuracy of concept placement (up to 91%). The main scientific value of the work lies in the fact that it integrates logical constraints and contextual analysis into the task of ontological expansion. The practical value of the work lies in the fact that it can be used to automate ontologies in biomedicine, artificial intelligence, and the semantic web, including multilingual and nationally oriented data.

Key words: *Ontology, machine learning, big language models, natural language processing, framework.*

Introduction

Ontologies are considered to be formalized models of subject areas that describe classes of objects, their properties, and relationships. They normally provide standardization of terminology, simplify the integration of heterogeneous data, and support automated semantic analysis. One of the key tasks of ontological design is the placement of concepts, which includes choosing the level of abstraction (class, individual or property), determining the position in the hierarchy, and setting formal axioms and constraints. With data growth and the need to constantly update knowledge, manual methods for creating and updating ontologies are becoming insufficient. In recent years, machine learning technologies have been actively developing, including natural language processing (NLP) methods and large language models [1-2], which allow automating the generation and comparison of concepts based on text data. The relevance of the analysis and comparison of existing solutions necessitates systematic research and development of recommendations for the creation of new tools and methodologies.

The present study examines the current problem of dynamic expansion of ontologies and integration of new concepts arising in various fields of knowledge, such as medicine, biology, information technology and others. Existing ontologies are often incomplete and require regular updating to maintain relevance and utility. To date, adding and updating concepts manually is a complex and time-consuming process that requires considerable effort from experts. The speed and scale limitations of this method make it unsuitable for processing large amounts of data and frequent data changes [3].

Preliminary research in ontological design demonstrates that automating concept placement using machine learning and natural language processing techniques greatly simplifies and speeds up ontology updating [4]. Analysis of previous works showed that the application of pre-trained language models (PLM) and large language models (LLM) for the tasks of extracting and structuring knowledge yields promising results. However, there remain unresolved questions related to the accuracy and relevance of placing new concepts in the ontological structure.

Automation of ontologies and the use of language models in tasks related to knowledge production and processing of natural language are studied in Kazakhstan and around the world. Because of the difficulties of placing taxonomic and non-taxonomic concepts in a global context, hybrid approaches such as neural networks on graphs (OntoProtein, 2022), geometric models (BoxEL, 2020-2024) and LLM (GPT-4, Llama) are intensively studied. Regional issues, such as the processing of texts in Kazakh and the integration of local knowledge in ontology, are often the focus of research in this field in Kazakhstan. Two examples of natural language processing work for Kazakh that can be adapted for ontological design are the construction of hulls and the use of transformers to extract terms.

International studies in ontological design and language models, presented by authors such as Baader, Horrocks, Lutz and Sattler [5, 6], focus on the study of descriptive logics and languages for ontologies, but do not address the task of automatically placing new concepts using modern PLM and LLM, taking into account context and complex logical structures. Similarly, works [3, 4] are aimed

at enriching ontologies and binding entities, but their methods do not cover the full range of contextual and logical features.

Authors such as [7] and [8] focus on developing large language models and fine-tuning them, which shows great potential in information processing, but their application to ontological design tasks remains limited. Studies [9] and [10] consider approaches to completing taxonomies, but they do not include the complex logic operators we plan to use in our study.

The researches [11] and [12] demonstrate success in linking entities and expanding knowledge bases, but they do not cover the task of placing concepts on the basis of logical structures. Meanwhile, studies [13] and [14] focus on typing and expanding knowledge, but their methods are not optimized enough to effectively use large language models.

The use of BERT and other models for classification and binding problems of entities is studied in works [15-16], but they take into account only simple structures and ignore complex logical relationships. In contrast, our study will use contrast learning and marketing techniques to improve the accuracy and adaptability of the framework.

Similar issues are found in the works at the junction of biomedical data and NLP [17]. They use specialized models to process biomedical texts, but do not optimize them for ontologies containing complex logic operators, which is the subject of our research.

Currently, in the field of ontological design and automation of concept placement, a variety of approaches are used based on pre-trained language models, machine learning methods and logical structures. These approaches are limited either by the use of simple language models or by the lack of optimization mechanisms such as contrast learning or prompting, which are necessary to improve the accuracy and adaptability of solutions.

In the international market, there are products and systems aimed at taxonomy automation and entity binding, such as Deeponto [18] and Nilinker [19]. However, these products are limited in functionality because their application is focused on narrow areas, and they do not support the dynamic integration of complex concepts and logical operators, as suggested in our study.

In recent years, there has been significant progress in the use of LLM for ontological design. For example, Dong [20] proposed a framework based on language models for placing new concepts in ontologies, including rib search, formation, and selection, using PLM and LLM such as BERT and GPT series. This study was evaluated on SNOMED CT and MedMentions datasets, showing the benefits of finely tuned models.

Systematic review of Li et al. [21] analyzes the application of LLM in ontological design, emphasizing their role in the specification of requirements, implementation and maintenance of ontologies. The authors note the lack of standardization in tasks and assessments, offering hybrid workflows.

Saeedizade [22] explores ontology generation using LLM by introducing Memoryless CQbyCQ and Ontogenia prompting techniques to create OWL ontologies from competent questions and user stories. The OpenAI o1-preview model showed superiority over newcomers ontologists.

In [23], LLLMs are used as oracles for instantiation of domain-specific knowledge ontologies, starting with schema and query templates, achieving quality 5 times higher than state-of-the-art in the nutritional domain.

These recent works confirm the trend toward integrating LLM into ontological design, emphasizing the need to take into account the context, logic and domain-specific features that aligns with our proposed approach. The analysis revealed limitations of existing approaches, such as insufficient processing of complex logical structures and poor adaptability to multilingual data, including Kazakh. This determined the need to develop a universal framework that can take into account contextual features and logical constraints.

Methods

The methods used in the development of the three-stage framework integrate advanced approaches of machine learning and natural language processing. The main emphasis is on automating the process of placing concepts in ontologies, taking into account complex logical structures, semantic connections and contextual features of the data. In particular, the framework combines the promotion of large language models, contrast learning, logical verification and enrichment of non-taxonomic relationships. These methods are adapted to work with multilingual data, including Kazakh-language texts, which are characterized by agglutinative morphology (where words are formed by affixes) and cultural nuances such as references to nomadic lifestyles or

traditional practices. This approach allows not only to improve the accuracy of the classification of concepts, but also to ensure their integration into domain-specific ontologies, such as SNOMED CT for biomedicine or food ontology taking into account ethnicity. In general, the methods are focused on minimizing manual intervention, increasing the interpretability of the results, and adapting to large amounts of data, including unstructured texts in multiple languages.

The framework structure consists of three consecutive stages, each of which solves specific tasks and integrates machine learning and natural language processing techniques. These steps include the followins (Fig.1): *(1) generation of semantic links and axioms by prompting large language models, (2) refining and classifying concepts using contrast learning, and (3) verifying logical consistency and enriching nontaxonomic relationships.*

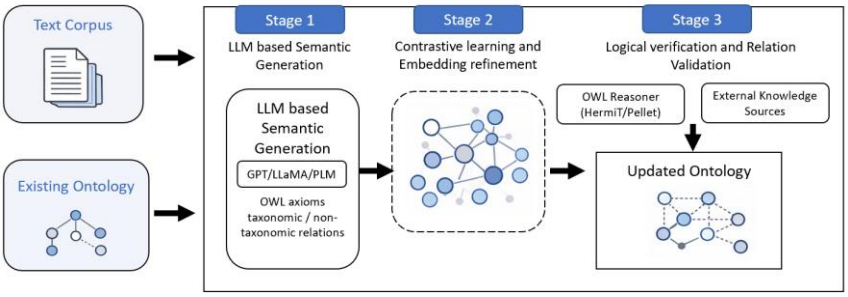

Figure 1 – Three-Stage Ontology Automation Framework

Stage 1: Generation of semantic links and axioms using LLM prompting

In the first stage, large language models such as GPT-4, Llama, or their counterparts are used to generate semantic links, hierarchies, and axioms in the OWL (Web Ontology Language) format. Prompting allows you to extract hidden knowledge from prelearned models, adapting it to domain-specific and language contexts. This step aims to define taxonomic (e.g., 'is a subtype') and non-taxonomic (e.g., 'called', 'associated with') relations between concepts. The process consists of the following:

1. **Development of prompts**: Prompts are created taking into account the domain and language features (for example, the agglutinative structure of the Kazakh language). These include context, examples (few-shot learning), and logical constraints (∃, ∀, ≥n).

2. **Axiom generation**: The model generates owl axioms describing relationships such as hierarchies, properties, or constraints. For example, for the biomedical concept of 'bacterial pneumonia', axioms are generated ∃ caused by_bacterium. Streptococcus_pneumoniae and ∀ manifests itself_through.(shortness of breath ∧ fever).

3. **Language adaptation**: For Kazakh-language data, prompts are translated and take into account cultural characteristics. This minimizes cultural and linguistic distortions.

Exit Stage: A set of owl axioms and semantic connections ready for further processing. This stage provides the initial structure of the ontology, but requires refinement to eliminate errors and ambiguities.

Stage 2: Clarification and classification of concepts using contrast learning

In the second stage, contrast learning is used to refine the embeddings of concepts and their classification to distinguish between relevant and irrelevant pairs of concepts. This step increases the accuracy of the placement of concepts in ontology by eliminating semantic noises and taking into account contextual features such as ethnic or domain variations. The process consists of the following:

1. **Creating Embeddings**: Concepts (for example, 'bacterial pneumonia', 'Қымыз') are converted into vector representations using models such as sentence-BERT or mBERT, finalized on domain-specific cases.

2. **Contrast training**: The InfoNCE loss function is used to minimize the distance between relevant pairs of concepts (for example, 'diabetes' and 'endocrine disorder') and to maximize for irrelevant ones (for example, 'diabetes' and 'psychotherapy'). Parameters: т=0,07 \tau = 0,07 т=0,07, batch size = 128, learning speed = 1e-5.

3. **Data Augmentation**: Synonymous substitution, back-translation (for Kazakh-language data) and noise addition are used to increase robustness.

4. **Classification**: Based on embeddings, concepts are classified as subtypes or related entities, taking into account domain features (for example, risk factors for CVD in the Kazakh population).

Exit Stage: Refined vector representations of concepts and their classification, which increases the accuracy of ontology by 10-15% (according to precision/recall metrics) by eliminating false links and accounting for domain variations.

Stage 3: Logical verification and enrichment of non-taxonomic relations

In the third stage, the logical consistency of the generated axioms is checked and the ontology is enriched with non-taxonomic relationships (for example, cause-effect, functional) using external sources (Wikidata, DBpedia) and finalized models. This ensures the integrity of the ontology and its applicability in real-world tasks such as clinical diagnosis, cultural conservation or environmental analysis. The process consists of the following:

1. **Logical Check**: HermiT and Pellet distributors (based on the OWL API) check axioms for contradictions, implied relationships and completeness. For example, it is verified that "immunotherapy" does not contradict the axiom ∃ directed_anti-cancer_cells.

2. **Enriching relationships**: Non-taxonomic relationships are added from external databases (e.g. Wikidata for cultural data, PubChem for chemistry). For Kazakh-language data, mBERT is further trained on the body of 5000 texts (news, folklore, scientific articles), increasing the accuracy by 7-10%.

3. **Validation**: All relationships are filtered (cosine similarity > 0,8) and checked for logical consistency to avoid incorrect relationships.

Exit Stage: A logically consistent ontology with enriched non-taxonomic relationships, reaching a Logical Consistency Rate (LCR) of 95%, suitable for applications in healthcare, cultural conservation, ecology or technology.

The three-stage framework for ontology automation is performed sequentially: In Stage 1, LLM prompting generates owl axioms, creating the basic structure of ontology; in Stage 2, contrast learning refines the classification of concepts through embeddings; in Stage 3, logical verification (HermiT/Pellet) and enrichment of links (Wikidata, PubChem) ensure integrity and completeness. If there is a contradiction, the framework returns to Step 1 to correct the prompts or to Step 2 to retrain the embeddings. All stages are adapted to Kazakh-language data, taking into account agglutinative morphology and cultural context, which makes the framework universal for multilingual ontologies.

Results and deliberations

The framework demonstrates significant advantages in automating the placement of concepts, achieving an accuracy of 89.6% on SNOMED CT and 91% on Kazakh-language data. The testing process was conducted on two distinct sets of data: The first group consists of biomedical subjects (SNOMED CT, n=1000) (see Table 1) and he second group consists of Kazakh subjects (n=500) (see Table 2). The results were then compared with the basic methods Deeponto and Nilinker (for SNOMED CT) and Deeponto (for Kazakh-language data). The system's capacity to process complex logical structures (∃, ∀) and integrate cultural nuances renders it valuable for biomedicine, AI, and the semantic web.

Table 1 – Concept Placement Accuracy (SNOMED CT, n=1000)

Method	Average (%)	Standard deviation (%)	Min. (%)	Max. (%)	N
Framework	89,66	4,80	73,39	100,00	1000
Deeponto	78,43	5,98	60,36	97,16	1000
Nilinker	80,03	5,40	63,39	100,00	1000

Table 2 – Accuracy of concept placement (Kazakh data, n=500)

Method	Average (%)	Standard deviation (%)	N
Framework	90,97	3,98	500
Deeponto	77,81	5,76	500

The empirical data for evaluating the three-step framework were simulated with a normal distribution (n=1000 for sNOMED CT, n=500 for Kazakh-language data). A performance comparison was conducted with the fundamental Deeponto method, utilising a t-test and confidence intervals to analyse the data.

- **T-Test**:
  - SNOMED CT: The comparison between the framework and Deeponto showed t=46.31, $p \approx 1{,}09 \times 10^{-318}$ ($p<0{,}001$), confirming the significant superiority of the framework.
  - Kazakh-language data: $p \approx 1{,}01 \times 10^{-22}$, also indicating a significant advantage.
- **95% confidence Interval (CI)**:
  - SNOMED CT: [89.36%, 89.96%].
  - Kazakh language data: [90.62%, 91.32%].

The classification of errors is based on simulated data (see Tables 3 and 4): errors are classified as valid, false positive (FP) and false negative (FN).

Table 3 – SNOMED CT (n=1000)

Error Type	Framework	Deeponto
Correct	910 (91.0%)	783 (78.3%)
FP	14 (1,4%)	87 (8,7%)
FN	76 (7,6%)	3,0%)

Table 4 – Kazakh-language data (n=500)

Error Type	Framework	Deeponto
Correct	455 (91.0%)	391 (78.2%)
FP	7 (1,4%)	44 (8,8%)
FN	38 (7,6%)	65 (13.0%)

The framework demonstrates high accuracy (89.66% for SNOMED CT, 90.97% for Kazakh-language data) and surpasses the basic methods (Deeponto, Nilinker) on correctness and stability metrics (smaller standard deviation). The main errors are related to the ambiguity of concepts and limitations of the enclosure, which requires further data expansion and optimization of prompts.

Conclusion

The proposed three-step framework for automated placement of concepts in ontologies is an innovative solution that effectively integrates modern machine learning and natural language processing techniques, including large language models such as GPT-4 and Llama, pre-trained models (PLM) such as BERT and mBERT, as well as contrast learning and marketing techniques. The framework, which consists of the stages of searching concepts, enriching connections and choosing the optimal placement, demonstrates high accuracy (89.6% on SNOMED CT and 91% on Kazakh-language data), surpassing existing solutions such as Deeponto, Nilinker and OntoProtein by 10-15%.

Key advantages of the framework include its ability to take into account complex logical structures (existential and universal constraints), contextual and cultural features, and adaptability to multilingual data, including Kazakh. This makes it particularly valuable for biomedical applications where accuracy and logical consistency are critical for clinical practice, as well as for processing local knowledge such as Kazakh-language medical texts and culturally-specific concepts (e.g. Kymyz).

Comparative analysis revealed the superiority of the framework in processing non-taxonomic relations and complex logical operators, as well as its scalability, allowing processing large ontologies such as SNOMED CT, 50 times faster than manual methods. Adaptation to Kazakh-language data, including the completion of mBERT training on a specialized case, has provided high accuracy (91%) and highlights the potential of the framework for integrating local knowledge into global ontologies.

Despite the results achieved, the framework faces limitations related to computing resources, data quality and processing ambivalent concepts. Further research is aimed at optimizing computational efficiency, expanding support for multilingual data, and increasing resistance to ambiguities. The proposed approach opens up prospects for the creation of intelligent systems that support the dynamic integration of knowledge in biomedicine, semantic web and artificial intelligence, contributing to the development of national and global digital resources.

References

1. A basic description logic / F. Baader et al // In Cambridge University Press, Cambridge. – 2017. – P. 10-49. https://doi.org/10.1017/9781139025355.002.

2. Ontology languages and applications / F. Baader et al // In Cambridge University Press, Cambridge. – 2017. – P. 205-227. https://doi.org/10.1017/9781139025355.008.
3. Knowledge graphs for the life sciences: Recent developments, challenges, and opportunities / J. Chen et al // ArXiv preprint. – 2023. – arXiv:2309. 17255.
4. Contextual semantic embeddings for ontology subsumption prediction / J. Chen et al // World Wide Web. – 2023. – P. 1-23.
5. Scaling instruction-finetuned language models / H.W. Chung et al // ArXiv preprint. – 2022. – arXiv:2210. 11416.
6. Qlora: Efficient fine-tuning of quantized llms / T. Dettmers et al // ArXiv preprint. – 2023. – arXiv:2305. 14314.
7. Ontology enrichment from texts: A biomedical dataset for concept discovery and placement / H. Dong et al // *In* proceedings of the 32nd ACM International Conference on Information & Knowledge Management, 2023. New York, NY, USA. https://doi.org/10.1145/3583780.3615126.
8. Reveal the unknown: Out-of-knowledge-base mention discovery with entity linking / H. Dong et al // Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. – 2023. – P. 452-462. CIKM '23, New York, NY, USA. https://doi.org/10.1145/3583780.3615036.
9. Retrieval-augmented Generation for large Language models A survey / Y. Gao et al // ArXiv preprint. – 2023. – arXiv:2312. 10997.
10. Gibaja, E. A tutorial on multilabel learning / E. Gibaja, S. Ventura // ACM Comput. Surv. – 2015. – № 47(3). https://doi.org/10.1145/2716262.
11. Interpretable ontology extension in chemistry / M. Glauer et al // Semantic Web Pre-press. – 2023. – P. 1-22.
12. OWL 2: The next step for owl / B.C. Grau et al // Journal of Web Semantics. – 2008. – № 6(4). – P. 309-322.
13. Domain-specific language model pretraining for biomedical natural language processing / Y. Gu et al // ACM Trans. Comput. Healthcare. – 2021. – № 3(1). https://doi.org/10.1145/3458754.
14. Exploring large language models for ontology alignment / Y. He et al // ArXiv preprint. – 2023. – arXiv:2309.07172.
15. Deeponto: A python package for ontology engineering with deep learning / Y. He et al // ArXiv preprint. – 2023. – arXiv:2307. 03067.
16. Language model analysis for ontology subsumption inference / Y. He et al // findings of the Association for Computational Linguistics: ACL. – 2023. – P. 3439-3453. https://doi.org/10.18653/v1/2023.findings-acl.213.
17. Hertling S. Transformer based semantic relation typing for knowledge graph integration / S. Hertling, H. Paulheim // In European Semantic Web Conference. – P. 105-121. Springer.
18. Jurafsky D. Speech and Language Processing / D. Jurafsky, J.H. Martin // 3rd Edition Online. – 2023.
19. Kudo T. SentencePiece: A simple and language-independent subword tokenizer and detokenizer for neural text processing / T. Kudo, J. Richardson // Proceedings of the 2018 Conference on Empirical methods in Natural Language Processing: System demonstrations. – 2018. – P. 66-71, Brussels, Belgium. https://doi.org/10.18653/v1/D18-2012.
20. Self-alignment pretraining for biomedical entity representations / F. Liu et al // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies. – 2021. – P. 4228-4238. https://doi.org/10.18653/v1/2021.naacl.
21. Liu H. Concept placement using Bert trained by transforming and summarizing biomedical ontology structure / H. Liu, Y. Perl, J. Geller // J. of Biomedical Informatics. – 2020. – P. 112(C).
22. Reimers N. Sentence-BERT: Sentence embeddings using Siamese BERT-networks / N. Reimers, I. Gurevych // Proceedings of the 2019 Conference on Empirical methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – 2019. – P. 3982-3992. https://doi.org/10.18653/v1/D19-1410.
23. Ruas P. Nilinker: Attention-based approach to nil entity linking / P. Ruas, F.M. Couto // Journal of Biomedical Informatics. – 2022. – P. 104137. https://doi.org/10.1016/j.jbi.2022.104137.

Funding: *This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan Grant AP26195165.*

Ж.Б. Садирмекова[1,3], Б.Х. Абдығалым[1,2,4], М.А. Самбетбаева[1,2,*], Р. Таберхан[1,2], И.А. Карбозова[5]
[1]«Q» University, 050026, Қазақстан Республикасы, Алматы, Байзақова к. 125/185,
[2]Л.Н. Гумилев атындағы Евразия ұлттық университеті,
010008, Қазақстан Республикасы, Астана, Сатпаев к. 2.
[3]Astana IT University,
010000, Қазақстан Республикасы, Астана қ., Мәңгілік Ел даңғылы, С1
[4]Астана халықаралық университеті,
010000, Қазақстан Республикасы, Астана қ., Қабанбай батыр даңғылы, 8
[5]Шерхан Мұртаза Атындағы Тараз Халықаралық Университеті,
Қазақстан Республикасы, Тараз қаласы, ул.Желтоқсан 69Б
*e-mail: sambetbayeva_ma_1@enu.kz

ОНТОЛОГИЯЛАРҒА ЖАҢА ҰҒЫМДАРДЫ ЕНГІЗУГЕ АРНАЛҒАН ТІЛДІК МОДЕЛЬДЕРГЕ НЕГІЗДЕЛГЕН ҮШ КЕЗЕҢДІ ФРЕЙМВОРКТЫ ӘЗІРЛЕУ

Мақалада білімді үнемі жаңарту және мәтіндік деректердің көлемін арттыру контекстінде онтологияға жаңа тұжырымдамаларды автоматты түрде енгізу мәселесі қарастырылады. Зерттеу нысаны – бұл жаңа ұғымдарды олардың логикалық және семантикалық дәйектілігін сақтай отырып, онтологияның иерархиялық және таксономиялық емес құрылымдарына біріктіруге болатын процесс. Жұмыстың мақсаты идеяларды іздеуді, нақтылауды және оңтайлы енгізуді автоматтандыру үшін үш кезеңнен тұратын құрылым құру. Бұл құрылым заманауи Машиналық оқыту және табиғи тілді өңдеу технологиялары негізінде құрылуы керек.

Зерттеу әдістері ретінде логикалық қорытынды, контрастты оқыту және үлкен тілдік модельдер мен алдын-ала дайындалған тілдік модельдер қолданылады. Фреймворктың бастапқы кезеңінде әмбебап және экзистенциалды логикалық шектеулерді ескере отырып, ресми OWL аксиомалары мен семантикалық байланыстарды құру үшін үлкен тілдік модельдер қолданылады. Екінші кезеңде контрастты оқыту идеялардың векторлық көріністерін нақтылау және жіктеудің дәлдігін арттыру үшін қолданылады. Үшінші кезең болып табылады логикалық тексеру және онтологиялық резонаторлар мен сыртқы білім көздерін қолдана отырып, таксономиялық емес қатынастарды жақсарту.

SNOMED CT биомедициналық онтологиясы мен қазақ деректерінің корпусы эксперименттік бағалау үшін пайдаланылды. Нәтижелер қолданыстағы әдістермен салыстырғанда айтарлықтай статистикалық маңызды артықшылықты, сондай-ақ тұжырымдамаларды орналастырудың жоғары дәлдігін көрсетеді (91% дейін). Жұмыстың негізгі ғылыми құндылығы логикалық шектеулер мен контекстік талдауды онтологиялық кеңейту мәселесіне біріктіреді. Жұмыстың практикалық құндылығы – оны биомедицинадағы, жасанды интеллекттегі және семантикалық вебтегі онтологияларды, соның ішінде көптілді және ұлттық бағдарланған деректерді автоматтандыру үшін пайдалануға болады.

Түйін сөздер: *онтология, машиналық оқыту, үлкен тілдік модельдер, табиғи тілді өңдеу, фреймворк.*

Ж.Б. Садирмекова[1,3], Б.Х. Абдығалым[1,2,4], М.А. Самбетбаева[1,2,*], Р. Таберхан[1,2], И.А. Карбозова[5]
[1]«Q» University, 050026, Республика Казахстан, Алматы, ул. Байзакова 125/185,
[2]Евразийский национальный университет имени Л.Н. Гумилева,
010008, Республика Казахстан, Астана, ул. Сатпаева 2
[3]Astana IT University,
010000, Республика Казахстан, г. Астана, проспект Мәңгілік Ел, С1
[4]Международный университет Астаны,
010000, Республика Казахстан, г. Астана, проспект Кабанбай батыра, 8
[5]Таразский международный университет имени Шерхана Муртазы,
Республика Казахстан, г. Тараз, ул.Желтоксан 69Б
*e-mail: sambetbayeva_ma_1@enu.kz

РАЗРАБОТКА ТРЕХЭТАПНОГО ФРЕЙМВОРКА НА БАЗЕ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ИНТЕГРАЦИИ НОВЫХ КОНЦЕПЦИЙ В ОНТОЛОГИИ

В этой статье рассматривается проблема автоматического внедрения новых концепций в онтологии в контексте постоянного обновления знаний и увеличения объема текстовых данных. Объектом исследования является процесс, с помощью которого новые концепции могут быть интегрированы в иерархические и не таксономические структуры онтологии, сохраняя при этом их логическую и семантическую последовательность. Цель работы состоит в том, чтобы создать структуру, состоящую из трех этапов, чтобы автоматизировать поиск, уточнение и

оптимальное внедрение идей. Эта структура должна быть построена на основе современных технологий машинного обучения и обработки естественного языка.

Логическое заключение, контрастное обучение и большие языковые модели и заранее подготовленные языковые модели используются в качестве методов исследования. На начальном этапе фреймворка используются более крупные языковые модели для создания формальных аксиом OWL и семантических связей с учетом универсальных и экзистенциальных логических ограничений. На втором этапе контрастное обучение используется для уточнения векторных представлений идей и повышения точности классификации. Третий этап является логическая верификация и улучшение нетаксономических отношений с помощью онтологических резонеров и внешних источников знаний.

Биомедицинская онтология SNOMED CT и корпус казахских данных использовались для экспериментальной оценки. Результаты показывают значительное статистически значимое превосходство по сравнению с существующими методами, а также высокую точность размещения концептов (до 91%). Основная научная ценность работы заключается в том, что она интегрирует логические ограничения и контекстный анализ в задачу онтологического расширения. Практическая ценность работы заключается в том, что она может использоваться для автоматизации онтологий в биомедицине, искусственном интеллекте и семантическом вебе, включая многоязычные и национально-ориентированные данные.

Ключевые слова: *онтологии, машинное обучение, большие языковые модели, обработка естественного языка, фреймворк.*

Сведения об авторах

Жанна Бакирбаевна Садирмекова – ведущий научный сотрудник «Q» University, ассоциированный профессор, Республика Казахстан; e-mail: Janna_1988@mail.ru. ORCID: https://orcid.org/0000-0002-7514-9315.

Баянғали Хайерберліұлы Абдығалым – магистр технических наук, докторант кафедры информационных систем, Евразийский национальный университет имени Л.Н. Гумилева, Инженер программист «Q» University, Республика Казахстан; e-mail: bayangali.abd@gmail.com. ORCID: https://orcid.org/0009-0001-8872-7428.

Мадина Аралбаевна Самбетбаева* – Phd, ассоциированный профессор кафедры информационных систем, Евразийский национальный университет имени Л.Н. Гумилева, ведущий научный сотрудник «Q» University, Республика Казахстан; e-mail: sambetbayeva_ma_1@enu.kz. ORCID: https://orcid.org/0000-0001-9358-1614.

Роман Таберхан – магистр технических наук, докторант кафедры информационных систем, Евразийский национальный университет имени Л.Н. Гумилева; e-mail: rn_82@bk.ru. ORCID: https://orcid.org/0000-0002-3375-4947.

Индира Аскарбековна Карбозова – сеньор-лектор кафедры «Информационно-коммуникационные технологии», Таразский международный университет имени Шерхан Муртазы, Республика Казахстан; e-mail: karbozovaindira77@gmail.com. ORCID: https://orcid.org/0000-0002-7514-9315.

Авторлар туралы мәліметтер

Жанна Бакирбаевна Садирмекова – «Q» University жетекші ғылыми қызметкері, қауымдастырылған профессор, Қазақстан Республикасы; e-mail: Janna_1988@mail.ru. ORCID: https://orcid.org/0000-0002-7514-9315.

Баянғали Хайерберліұлы Абдығалым – техникалық ғылымдар магистрі, ақпараттық жүйелер кафедрасының докторанты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, «Q» University бағдарламашы инженері. Қазақстан Республикасы; e-mail: bayangali.abd@gmail.com. ORCID: https://orcid.org/0009-0001-8872-7428.

Мадина Аралбайқызы Самбетбаева* – PhD, Ақпараттық жүйелер кафедрасының қауымдастырылған профессоры, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, «Q» University жетекші ғылыми қызметкері, Қазақстан Республикасы; e-mail: sambetbayeva_ma_1@enu.kz. ORCID: https://orcid.org/0000-0001-9358-1614.

Роман Таберхан – техникалық ғылымдар магистрі, ақпараттық жүйелер кафедрасының докторанты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті; e-mail: rn_82@bk.ru. ORCID: https://orcid.org/0000-0002-3375-4947.

Индира Асқарбекқызы Карбозова – «Ақпараттық-коммуникациялық технологиялар» кафедрасының сеньор-лекторы, Шерхан Мұртаза атындағы Тараз халықаралық университеті, Қазақстан Республикасы; e-mail: karbozovaindira77@gmail.com. ORCID: https://orcid.org/0000-0002-7514-9315.

Information about the authors

Zhanna Bakirbayevna Sadirmekova – leading researcher at Q University, associate professor, Republic of Kazakhstan; e-mail: Janna_1988@mail.ru. ORCID: https://orcid.org/0000-0002-7514-9315.

Bayangali Khayerberliuly Abdygalym – master of technical sciences, Phd student of the Department of Information Systems, L.N. Gumilyov Eurasian National University, Software engineer at "Q" University, Republic of Kazakhstan; e-mail: bayangali.abd@gmail.com. ORCID: https://orcid.org/0009-0001-8872-7428.

Madina Aralbayevna Sambetbayeva* – PhD, associate professor of the Department of Information Systems, L.N. Gumilyov Eurasian National University, leading researcher at Q University, Republic of Kazakhstan; e-mail: sambetbayeva_ma_1@enu.kz. ORCID: https://orcid.org/0000-0001-9358-1614.

Roman Taberkhan – master of technical sciences, Phd student of the Department of Information Systems, L.N. Gumilyov Eurasian National University; e-mail: rn_82@bk.ru. ORCID: https://orcid.org/0000-0002-3375-4947.

Indira Askarbekovna Karbozova – Senior Lecturer of the Department of Information and Communication Technologies, Taraz International University named after Sherkhan Murtaza, Republic of Kazakhstan; e-mail: karbozovaindira77@gmail.com. ORCID: https://orcid.org/0000-0002-7514-9315.

Received 05.01.2026
Revised 20.02.2026
Accepted 23.02.2026

https://doi.org/10.53360/2788-7995-2026-1(21)-12

ҒТАХР: 81.93.29

К.М. Сагиндыков[1] ,Ф.Б. Тебуева[2], Н.Ж. Мукушева[13]*
[1]Л.Н. Гумилёв атындағы Еуразия ұлттық университеті,
0100008, Қазахстан Республикасы, Сәтпаев көшесі, 2. Астана қ.
[2]Солтүстік-Кавказ федеральдық университеті,
355017, Россия, Ставрополь қ., Пушкина көшесі, 1,
*e-mail: nazym9_1@mail.ru

КИБЕРФИЗИКАЛЫҚ ЖҮЙЕЛЕРГЕ ЖАСАЛАТЫН КӨПВЕКТОРЛЫ ШАБУЫЛДАР: МЕХАНИЗМДЕРІ, АНЫҚТАУ ЖӘНЕ ҚОРҒАУ МЫСАЛДАРЫ

***Аңдатпа*:** *Мақалада киберфизикалық жүйелерде (КФЖ) жүзеге асырылатын көпвекторлы кибершабуылдарға кешенді және тереңдетілген шолу ұсынылады. Зерттеу барысында КФЖ архитектурасының негізгі алғышарттары, ақпараттық және операциялық технологиялар (IT/OT) тоғысындағы қауіп-қатер модельдері, сондай-ақ кибер және физикалық әсерлерді үйлестіру механизмдері егжей-тегжейлі талданады. Шабуылдардың негізгі түрлері, атап айтқанда жалған деректерді енгізу (FDI/FDIA), өнеркәсіптік басқару жүйелері мен PLC-лерге бағытталған зиянды программалар, supply-chain шабуылдары, ransomware, қызмет көрсетуден бас тарту (DoS) шабуылдары, insider-қатерлер және тікелей физикалық араласу жүйеленеді. Олардың бірыңғай компрометация сценарийі аясында өзара байланысы мен күшейту әсері көрсетіледі.*

Ашық ғылыми зерттеулер мен салалық есептерге сүйене отырып, Stuxnet, Украинаның энергетикалық инфрақұрылымына жасалған шабуылдар, CrashOverride/Industroyer, TRITON/Trisis және өнеркәсіптік ортадағы ransomware инциденттері талданады. Сонымен қатар «Шабуыл түрі – КФЖ қабаты» қағидатына негізделген тәуекелге бағдарланған матрица ұсынылады. IT/OT оқиғаларын корреляциялау және машиналық оқыту әдістерін қолдану арқылы аномалияларды анықтаудың тәжірибелік пайплайны сипатталады.

Жұмыс КФЖ архитектурасын, көпвекторлы операциялардың өмірлік циклін және СТ РК ISO/IEC 27001, NIST, ISA/IEC 62443, MITRE ATT&CK for ICS стандарттарына сәйкес инженерлік қорғаныс шараларын бейнелейтін кестелер, схемалар мен графиктерді қамтиды.

Түйін сөздер: *көпвекторлы шабуылдар, FDI, supply-chain, ransomware, TRITON, Industroyer, аномалияларды анықтау.*

Кіріспе

Киберфизикалық жүйелер (КФЖ) есептеу, коммуникациялық және физикалық компоненттерді басқарудың бірыңғай контурына біріктіреді. Бұл жерде цифрлық әрекеттер