Development and Testing of a Large Language Models Prompt for Natural Language Phrases Synthesis from Ontological Semantic Structures

Vladislav Kaverinskiy* †1, Oleksandr Palagin †1 and Anna Litvin †1

1 V. M. Glushkov Institute of Cybernetics of NAS of Ukraine, 40 Glushkov ave., 03187, Kyiv, 03187, Ukraine

Abstract

The article introduces an innovative approach that leverages a specially designed structured prompt for Chat GPT, a large language model. This approach was tested through a series of experiments aimed at generating natural language phrases from their underlying ontological representations. These representations were automatically derived from sentences in scientific and technical texts using advanced software tools. They encapsulate the entities identified in the text and the semantic relationships between them, which can be expressed in the sentences of the analyzed text. In more detail, the system identifies relationships between concepts and links them to entities within a sentence. These entities can be either simple sentences or parts of complex ones. The structured prompt provided to the language model includes detailed explanations of these semantic relationships and a set of concept pairs connected by these relationships, serving as the building blocks for sentence creation. The generated sentences were then compared to the original ones using the cosine similarity measure across various vectorization methods. The similarity scores, calculated using the xx_ent_wiki_sm model, ranged from 0.8193 to 0.9722. Despite these high similarity scores, some stylistic differences were noted in the generated sentences. This research holds significant practical value for the development of dialogue systems that integrate ontological methods with advanced language models, paving the way for more accurate and contextually aware information systems.

Keywords

large language model, ontology, natural language text synthesis, natural language text analysis, cosine similarity, text vectorization

1. Introduction

Research in the development of ontologies and the semantic processing of scientific data holds significant importance for the modern scientific community, particularly given the rapid growth of information and the need for researchers to use this information effectively. The primary goal is to develop and implement technologies that enable the swift and accurate retrieval and processing of scientific information, as well as facilitate interaction with information systems to maximize the utility of this data [1, 2]. Various systems and methodologies have been developed to address these challenges, including ontology-based systems and semantic processing techniques [1, 2, 3]. These systems utilize technologies such as the Semantic Web and cognitive graphics to enhance information retrieval and knowledge discovery in digital libraries. For instance, [1] details an ontology-based system for the semantic processing of scientific digital libraries, while the work [2] discusses a complex semantic processing of scientific data.

In addition to these developments, research in digital health and telerehabilitation has become a crucial area of modern science [3, 4, 5, 6, 7]. Researchers are actively working on innovative

14th International Scientific and Practical Conference from Programming UkrPROG'2024, May 14-15, 2024, Kyiv, Ukraine

* Corresponding author.
† These authors contributed equally.
insamhlaithe@gmail.com (V. Kaverinskiy); palagin_a@ukr.net (O. Palagin); litvin_any@ukr.net (A. Litvin)
0000-0002-6940-579X (V. Kaverinskiy); 0000-0003-3223-1391 (O. Palagin); 0000-0002-5648-9074 (A. Litvin)

technologies in this field, particularly focusing on the development of cloud-based platforms for patient-centered telerehabilitation [5, 6]. These platforms combine artificial intelligence and mathematical modelling to create effective and innovative approaches to rehabilitation and health support [5, 8]. For example, in [5, 6] the update highlights the advancements in cloud-based platforms for oncology patient rehabilitation. These platforms not only facilitate remote support and monitoring but also improve patient outcomes through personalized rehabilitation programs.

The state of the art in this area shows that the integration of ontology engineering with neural network technologies and artificial intelligence opens new perspectives for interacting with information systems and developing innovative services [8, 9]. For instance, [9] describes an integrated approach that combines neural networks and ontolinguistic paradigms to enhance the efficiency of intelligent dialogue systems and ensure their flexible adaptation to various user needs and subject areas. This integrated approach also emphasizes the importance of meta-learning and structured prompts to improve the performance of language models.

In the research presented in this article, further development of the approach to integrating large language models and ontological knowledge structures is achieved through the application of structured prompts, specifically in the context of natural language text generation based on semantic representation. The term "reverse synthesis" is introduced for this approach. This method involves the synthesis of natural language phrases from their ontological representations, which are automatically constructed from sentences of scientific and technical texts using previously developed software tools. These representations contain entities found in the text and the typed semantic relationships between them, which can be realized in the phrases of the analyzed text.

The system of relationships, specified by a set of concepts, is linked with the entity of the related part of the sentence, which can be a simple sentence or part of a complex sentence. The structured prompt for the large language model includes explanations of the semantic relationships between concepts in the context of sentence synthesis from ontological representation, as well as a set of pairs of concepts connected by semantic relationships, which serve as material for sentence creation.

For the design of natural language text synthesis software, it is necessary to consider the features of each step of the statement analysis algorithms. For this, methods of transforming behavioural software models into static ones should be used. UML notation diagrams are used for their representation [10]. Formal methods of preserving information about algorithms must allow the description of all the features of data flows [11]. At the same time, it is important to preserve complete information not only about the structure of the algorithm but also about its semantics [12].

The research presented in this work has practical significance for the development of dialogue information systems that combine the ontological approach with the use of large language models.

The aim of the study is to develop a structured prompt for a large language model to synthesize natural language statements based on ontological representations, and to subsequently evaluate the results by formally comparing the generated phrases with the original texts using various text vectorization models.

2. Related works

The work [13] presents a comprehensive overview of recent advancements in Semantic Web technologies, particularly focusing on new frameworks and protocols that enhance data interoperability and integration across various web platforms. The paper excels in detailing practical applications and case studies that demonstrate the real-world benefits of these advancements. However, the technical complexity of some sections may limit accessibility for non-specialists. Furthermore, while data privacy and security issues are mentioned, these critical areas require more in-depth examination given their growing importance in the digital age.

The machine learning algorithms. This integration has significant potential to improve the accuracy and efficiency of AI-driven applications. Nonetheless, the paper could benefit from more empirical evidence supporting the proposed methods. While theoretical insights are well-articulated, practical implementations and performance metrics are somewhat lacking. Additionally, scalability issues in applying these methods to large-scale knowledge bases are not thoroughly addressed.

Authors of [15] focus on the application of ontology-driven systems in healthcare, examining how ontologies can improve data management, clinical decision-making, and patient care. The strength of their paper lies in its practical examples and case studies, which vividly demonstrate the benefits of ontology-driven approaches in real-world healthcare settings. However, the challenges of implementing these systems in diverse healthcare environments are not sufficiently explored. Issues such as interoperability between different healthcare systems, the complexity of healthcare ontologies, and resistance to technological adoption among healthcare professionals are briefly mentioned but deserve a more thorough investigation. A broader review of existing literature would also help contextualize their findings within the wider body of research.

In [16] it was discussed the pivotal role of Semantic Web technologies in improving data interoperability across different systems and domains. They provide a comprehensive overview of the methodologies and frameworks that facilitate seamless data exchange and integration. The strength of their paper lies in its extensive review of existing technologies and practical implementations across various industries, including healthcare, finance, and logistics. However, the lack of detailed case studies showcasing real study [14] explores the intersection of artificial intelligence (AI) and ontological knowledge bases, highlighting innovative methods for enhancing AI systems' reasoning capabilities. Their work is notable for presenting a clear conceptual framework that integrates ontology-based knowledge representation with -world applications of these technologies is a notable limitation. More in-depth examples would help readers fully appreciate the practical implications of their findings. Additionally, a more detailed analysis of the challenges and limitations associated with implementing Semantic Web technologies, such as data standardization and scalability issues, would strengthen the paper.

The paper [17] explores the integration of ontology-based systems with artificial intelligence, focusing on the challenges and opportunities this integration presents. Their detailed examination highlights how ontologies can enhance the reasoning and decision-making capabilities of AI systems, improving knowledge representation and enabling more sophisticated AI applications. The balanced discussion of potential benefits and significant challenges, such as the complexity of ontology design, the need for standardization, and the difficulty of maintaining up-to-date ontological knowledge bases, is a key strength of their work. However, the paper could benefit from more empirical data and case studies demonstrating the practical application of these systems. A deeper exploration of future research directions would also provide valuable insights for advancing this field.

In [18] they delve into the application of ontology-driven decision support systems (DSS) in clinical settings, providing a thorough analysis of how ontologies can enhance clinical decision-making by improving data integration, knowledge management, and the accuracy of clinical recommendations. Their paper's key strength is its practical focus, showcasing real-world applications and the tangible benefits of ontology-driven DSS. However, the discussion on the limitations and potential barriers to widespread adoption of these systems, such as the complexity of ontology maintenance and the need for clinician training, could be more detailed. Additionally, exploring the integration of emerging technologies, such as machine learning and natural language processing, with ontology-driven DSS would provide a more comprehensive overview of future possibilities.

In summary, these analyzed studies collectively offer valuable insights into the current advancements and applications of Semantic Web technologies, AI integration with ontological knowledge bases, and ontology-driven systems in healthcare. They highlight significant progress in these fields, demonstrating the potential for improved data interoperability, enhanced AI reasoning capabilities, and better clinical decision-making. However, common areas for improvement include the need for more empirical evidence, a deeper exploration of challenges, and broader contextualization within existing research. Addressing these gaps in future studies will provide a more comprehensive understanding of these evolving fields and support the development of more effective and integrated technological solutions.

3. Materials and Methods

3.1 The developed method of constructing a natural language phrase based on ontological representation using a large language model

The study aims to develop a structured guideline for a large language model to synthesize natural language expressions based on ontological representations, followed by an evaluation of the results through a formal comparison of generated phrases with original texts using various text vectorization models. This task applies a template approach, as detailed in [19]. However, with the evolution of deep learning neural network approaches, exemplified by large language transformer models such as ChatGPT, there is a growing need to track their role as tools for synthesizing natural language sentences based on semantic structures. This direction is explored within this study.

A knowledge base derived from the text "Computer System Architecture" served as the test ontology. Managed by the graph database Neo4J, queries were executed using Cypher as the query language. The query below demonstrates how a specific sentence's text retrieval from the ontology is executed for comparison purposes:

```
MATCH (inp:Relationship)-[:SPO]->(inp_type:Relationship),
    (inp:Relationship)<-[:SPO]-(linked_group:Relationship),
    (linked_group:Relationship)-[:SPO]->(linked_group_type:Relationship),
    (linked_group:Relationship)<-[:SPO]-(certain_words_link:Relationship),
    (certain_words_link:Relationship)-[:SPO]->(sem_type:Relationship),
    (sem_type:Relationship)-[:SPO]->(w_link_type:Relationship),
    (certain_words_link:Relationship)-[:DOMAIN]->(main_entity:Class),
    (certain_words_link:Relationship)-[:RANGE]->(dependent_entity:Class)
WHERE
  inp_type.name = "SentenceGroups" and
  linked_group_type.name = "Groups" and
  w_link_type.name = "WordsLink" and
  ID(inp) = specify sentence ID
RETURN DISTINCT ID(inp) as id, inp.label as text, main_entity.label as main_entity, dependent_entity.label as dependent_entity, sem_type.label as sem_type;
```

This query also retrieves the corresponding set of semantic categories and related pairs of concepts (main and dependent entities) for the specified sentence. The results of this query served as input data for the task of reverse synthesizing natural language sentences.

According to the provided query from the ontology, a sentence with a specific identifier (specify sentence ID) is returned. The result includes the identifier, text, and a set of triplets such as "main entity, dependent entity, and semantic type" relative to the specified sentence. The semantic structure obtained from the ontology proves sufficient for constructing coherent natural language sentences of corresponding content.

,

To initiate the synthesis task in a large language model like ChatGPT, a corresponding instruction prompt (prompt) is required, as noted in [20], preferably in English. The prompt itself is structured in JSON format. The relevant prompt text is provided below:

{

"Introduction": "You are an expert in knowledge engineering and ontologies as well as in meaningful text generation in inflect languages. You will be provided with data obtained from some ontology through a query. The ontology was made automatically basing on the results of semantic analysis of a natural language text. The results are pairs of lemmatized words ('main entity' and 'dependent entity') accompanied with a name of syntactic-semantic relationship that linked them in the certain sentence.",

"Action to perform": "Assuming that all the data you will be provided belong to one sentence you are to make a try to restore the original sentence using such a prompt. Language of the ontology, input and output data is Ukrainian.",

"Restrictions": "Do not put the semantic relationships as a phrase as it given in the sentence you generate, it will be definitely wrong. It is just a prompt for syntactic linking. Remember that the provided words are lemmatized, so you are to put them in a correct form according to other entities of the sentence and the given syntactic-semantic relationships of the prompt.",

"Additional data to provide": "Also provide an estimated value of probability that the generated sentence corresponds the intent of the prompt given.",

"The essence of the syntactic-semantic relationship names and meaning explanation": {

"object property": "the dependent entity express a property or some characteristic, or quality of the main entity. When the response sentence generation you should use the dependent entity as an adjective with the main entity which is noun",

"action property": "the dependent entity express a property or some characteristic, or quality of the main entity which is an action. When the response sentence generation you should use the dependent entity as an adverb with the main entity which is verb",

"quality change": "the dependent entity express that the main entity may be subjected to some quality changes, which may follow from the other context",

"destination": "the dependent entity express the destination of the main entity",

"object": "the object (noun) affected throw the action expressed by the main entity",

"object / action": "the main entity performs an action expressed by the dependent entity",

"preposition binding": "merely shows that the main entity here in the context of the provided sentence is to be used with the preposition which is the dependent entity. This means that you should use this preposition with the main entity when the response sentence generation",

"possession": "the dependent entity or somewhat relates to the main entity. When generation this usually should be expresses using genitive case",

"equality": "the different name of the entity or an equivalent entity",

"objective entry": "the main entity is a part or member of the dependent entity",

"state": "a state or a constant characteristic of the main entity if it is noun or an entity linked to in if it is a verb"

},

"Input data": [

{

"main entity": "some word 1",

"dependent entity": "some word 2",

"semantic relationship": "semantic category 1"

},

{

"main entity": "some word n",

"dependent entity": "some word n+1",

"semantic relationship": "semantic category n"
}
]
}

This instruction guideline includes sections that establish initial parameters for the large language model regarding its behaviour and provide fundamental explanations of input data. The "Action to Perform" section formulates the direct task to be executed. The "Restrictions" section offers additional guidelines for forming the output text to eliminate ambiguity in instruction interpretation. The "Additional Data to Provide" section instructs the model to assess the quality of task execution.

The "The Essence of the Syntactic-Semantic Relationship Names and Meaning Explanation" section provides a dictionary of explanations for semantic relationship types and their usage in sentence construction. Given the considerable number of semantic categories within a limited character input for ChatGPT, the practical scope of this dictionary is constrained to semantic categories present in the given sentence.

Figure 1 outlines the overall process of forming natural language expressions based on their ontological representation. The essence of the experiment involved extracting individual sentences and their corresponding pairs of entities with semantic relationships from a test ontology created from the text "Computer System Architecture" using Cypher queries. Subsequently, applying the aforementioned instruction guideline, a large language model (ChatGPT) was tasked with generating grammatically correct Ukrainian sentences based on a set of entity pairs with specified semantic relationships. The response yielded the generated sentence and an assessment by the model of the probability that the sentence accurately reproduced the original (whose appearance the model did not know). Ten sentences from the specified text were used for testing purposes.

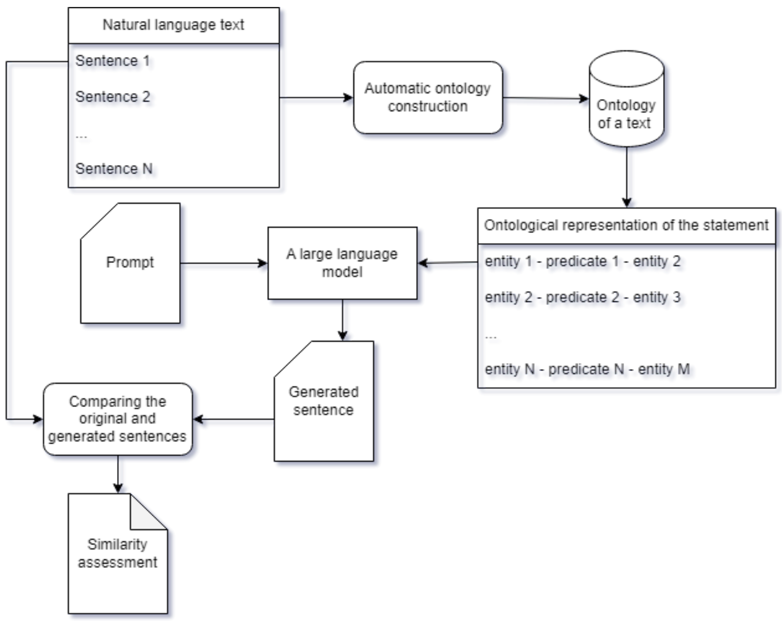


Figure 1: The general scheme of natural language expression generation based on ontological representation using a large language model

3.2 Methodology for experimental results evaluating

For comparing the similarity of the generated sentence to the original, cosine similarity values were utilized. Cosine similarity is a measure of similarity between two vectors in a pre-Hilbert space, used to measure the cosine of the angle between them. Thus, given two feature vectors (A and B), cosine similarity cos(θ) can be represented using their dot product and norm (1):

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

Cosine similarity ranges from 0 to 1, reflecting that term frequencies (tf-idf weights) are non-negative and the angle between term frequency vectors cannot exceed 90°. It is an effective metric, especially for sparse vectors, as it considers only non-zero values. Soft cosine similarity adjusts for feature similarities, where traditional cosine similarity treats functions in vector models as independent or entirely separate, whereas soft cosine acknowledges feature similarity within vector models. This allows for a generalized concept of cosine similarity and object similarity in vector space.

Entities such as words, N-grams, or syntactic N-grams can exhibit significant similarity, despite formally being considered different functions in the vector model. For N-grams or syntactic N-grams, Levenshtein distance [21] can be applied. To compute soft cosine similarity, a similarity matrix sss between functions is introduced, calculated using Levenshtein distance or other similarity measures like WordNet similarity. Subsequently, multiplication is performed using this matrix. For two N-dimensional vectors a and b, soft cosine is computed as:

$$soft\_\cos ine_1(a,b) = \frac{\sum_{i,j}^{N} s_{ij} \cdot a_i \cdot b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} \cdot a_i \cdot a_j} \cdot \sqrt{\sum_{i,j}^{N} s_{ij} \cdot b_i \cdot b_j}} \quad (2)$$

Here, sij represents the similarity between functions i and j. When no similarity exists between features (sii = 1, sij = 0 for i ≠ j), equation (2) equates to the standard cosine similarity formula.

Since direct mathematical computations on strings are impractical, and calculating metrics like cosine similarity requires vectorization, texts for processing and analysis underwent vectorization. To obtain vector representations of sentences, the Python library spaCy and language models such as uk_core_news_lg (for Ukrainian) and xx_ent_wiki_sm (multilingual) were employed. TF-IDF methodology was also applied to compute cosine similarity values, utilizing methods implemented in spaCy.

4. Results and Discussion

Table 1 presents an example of a phrase semantic structure obtained from the ontology.

The concepts (words) within a sentence are represented by pairs of entities connected by semantic predicates. In the example above, the word "призначений" (intended) is notably absent as an entity but serves as the basis for the semantic predicate of possession, linking the entities "принтер" (printer) and "створення" (producing). This exemplifies the intricacies observed in the results of semantic analysis from the original text during automated ontology creation. Additionally, the ontology includes the predicate "prepositional attachment", providing further specificity by indicating the specific preposition associated with each word in the sentence.

The presented semantic structure proves sufficient for constructing a coherent natural language sentence conveying the intended meaning. To initiate the task of synthesizing such sentences using a large language model like ChatGPT, appropriate prompt instruction is necessary.

Table 1

Example of the results of executing a query for the semantic structure of a sentence (for the Ukrainian language)

The original sentence text	Main entity	Dependent entity	The semantic type name
Принтер призначений для створення твердих копій документів.	принтер (a printer)	призначений (is designed)	nominal predicate
	принтер (a printer)	створення (producing)	appointment

(Eng.: A printer is designed to produce hard copies of documents.)	створення (producing)	для (for)	prepositional attachment
	створення (producing)	копія (a copy)	object
	копія (a copy)	документ (a document)	possession
	копія (a copy)	твердий (hard)	object property

This approach underscores the evolving methods in semantic analysis and the utilization of semantic predicates in automated ontology generation, reflecting advancements in natural language processing.

Quantitative assessments measuring the similarity between sentences generated by a large language model and their originals are detailed in Table 2.

The table compares cosine similarity values obtained under various conditions of vector representation for the analyzed sentences (original and generated). Additionally, it includes a subjective assessment of the likelihood of accurate reproduction by ChatGPT, which should be understood not as an entirely objective measure but rather as a benchmark and a model's self-critique indicator.

From the presented results, it is evident that the numerical assessment of cosine similarity heavily relies on the method of vector representation used for the analyzed texts. This underscores the sensitivity of cosine similarity scores to the approach taken in representing sentence vectors, reflecting ongoing advancements in natural language processing methodologies.

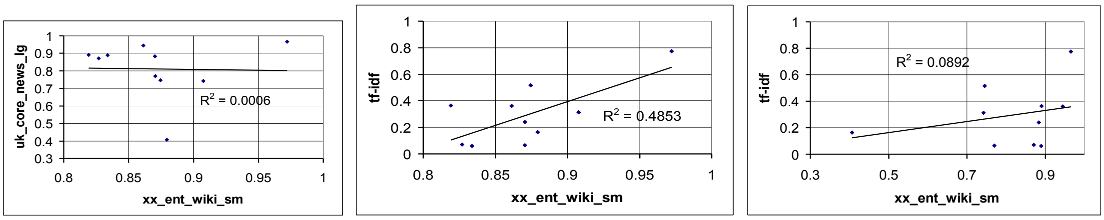
Immediately evident is that the language vectorization models xx_ent_wiki_sm and uk_core_news_lg yield quite high cosine similarity values (0.8716 and 0.8108, respectively). In contrast, the simpler tf-idf vectorization method produces significantly lower mean values and a wider range of variation. Let's delve deeper into this behaviour.

Table 2
Quantitative assessments of the quality of reverse synthesis of sentences from ontological representation

Self-assessment from ChatGPT		Cosine Similarity					
		Vectorization Model xx_ent_wiki_sm		Vectorization Model uk_core_news_lg		Vectorization Model tf-idf	
Mean	Range	Mean	Range	Mean	Range	Mean	Range
0.845 ±0.037	0.75 – 0.90	0.8716 ±0.0335	0.8193 – 0.9722	0.8108 ±0.1224	0.4067 – 0.9653	0.2927 ±0.1718	0.0607 – 0.7745

The xx_ent_wiki_sm model (multilingual) shows a narrow range of variation and a relatively high mean cosine similarity value. The decrease in the mean score when using the uk_core_news_lg model (for the Ukrainian language) is attributed to greater variability towards the lower end. However, the maximum values obtained for these two models are quite close. Put simply, applying the uk_core_news_lg model in certain cases often results in a considerably lower cosine similarity score.

A comparison of cosine similarity metrics obtained from the xx_ent_wiki_sm and uk_core_news_lg vectorization models is illustrated in Figure 2 (a), revealing a lack of significant correlation between the obtained values. The R2 value is only 0.0006. These models perceive natural language text somewhat differently.



a) b) c)

Figure 2: Comparison of cosine similarity values between original and generated sentences after text vectorization using different models:
a) xx_ent_wiki_sm / uk_core_news_lg;
b) xx_ent_wiki_sm / tf-idf;
c) uk_core_news_lg / tf-idf.

The analysis of directly generated sentences reveals that when vectorized using the uk_core_news_lg model, reduced cosine similarity scores occur in cases of generating complex and branching sentences, whereas the original sentence remains simpler yet semantically close. For instance, the original sentence "Жорсткий диск (вінчестер) також входить в системний блок." (Eng.: "The hard disk (Winchester) also fits into the system unit.") was represented ontologically by the large language model as: "Диск, відомий як вінчестер, є жорстким, входить у блок, який є системним." (Eng.: "The disk, known as Winchester, is hard, entering into the block, which is system."). The generated second sentence appears somewhat unnatural, filled with excessive entities and turns, despite conveying a similar meaning to the original.

The xx_ent_wiki_sm model exhibits less sensitivity to such occurrences, yielding a cosine similarity score of 0.8795 (close to the sample mean). Conversely, the uk_core_news_lg model shows a score of 0.4067, significantly lower. This trend is recurrent, albeit with lesser disparity; uk_core_news_lg is more susceptible to creating formally convoluted and excessive phrasing, while xx_ent_wiki_sm leans towards content analysis rather than form.

However, as observed from the graph, there is no clear correlation between the models. Instances exist where uk_core_news_lg yields higher cosine similarity scores compared to xx_ent_wiki_sm, and vice versa. Analysis of specific cases suggests that while lexical and syntactic structures closely match the original, the content may be slightly distorted. An example of this discrepancy is evident in the phrase: "Головним пристроєм комп'ютера є центральний процесор." (Eng.: "The main device of the computer is the central processor.") and its generated counterpart: "Пристрій, який належить до комп'ютера, є об'єктом, де центральний процесор є головним пристроєм." (Eng.: "The device belonging to the computer is an object where the central processor is the main device."). Here, we observe stylistic distortion in the generated output, along with some semantic twisting.

Furthermore, cosine similarity scores for xx_ent_wiki_sm exhibit a noticeable correlation with the tf-idf methodology. The highest scores were achieved when sentences were nearly identical: "Вміст цієї пам'яті зберігається лише при увімкненому живленні." (Eng.: "The content of this memory is stored only when the power is on.") and "Вміст пам'яті зберігається лише при увімкненому живленні." (Eng.: "The memory content is stored only when the power is on."). Conversely, tf-idf proves more sensitive to distorted sentences, resulting in numerically lower scores while maintaining content correlation. Conversely, uk_core_news_lg shows a weak correlation with tf-idf.

Thus, for tasks prioritizing semantic content over form, xx_ent_wiki_sm vectorization may be preferred. Conversely, uk_core_news_lg proves sensitive to both content and formal rearrangement, suitable for achieving more rigorous and sensitive cosine similarity comparisons. Tf-idf, despite its sensitivity to formal rearrangement, is less adept at discerning semantic similarity.

Overall, comparing scores from different methodologies and visually examining experimental results allows us to conclude that the proposed approach of generating natural language sentences in Ukrainian based on ontological representation using large language models effectively conveys the general gist and essence of the original phrase. However, generated phrases often appear somewhat unnatural, containing excessive entities and phrases, indicating the relevance of text generation systems based on ontological representations (including query results to the ontology) built on rules and templates. Such approaches, with well-established and comprehensive rule systems and templates, can generate significantly more qualitative natural language phrases based on semantic representations than large language models alone. Additionally, large language models effectively handle generating textual responses based on context sets and lists of relevant intentions.

Further development of the mentioned research could involve refining corresponding instruction prompts for large language models. These instructions would not only facilitate the reproduction of original text based on ontological models but also enable logical inference from the provided information, thereby advancing towards solving the global challenge of ontological approaches—acquiring new knowledge [2, 3]. Additionally, exploring alternative GPT models, including autonomous small language models and possibly processor architectures proposed in recent studies [22, 23], appears promising. This necessitates extensive research into the nature of knowledge itself (assessing novelty, logical derivation of secondary knowledge from primary sources, semantic-logical comparison of contexts, etc.).

5. Conclusions

Comparing assessments obtained through different methods and visually reviewing the experiment results allows us to summarize that the proposed approach to generating natural Ukrainian sentences based on their ontological representation using a large language model can effectively convey the general meaning and sense of the original phrase. This is evidenced by high cosine similarity scores (approximately 0.87 ± 0.03 with the xx_ent_wiki_sm vectorization model). However, often, though not in all cases, the generated phrase may appear somewhat unnatural and contain redundant entities and expressions. These findings indicate that while large language models can be used for text generation based on ontological representation and conveying general meaning, the generated phrases are often imperfect in form (and sometimes in nuances of meaning). The approach proposed in this study is seen as promising in providing users with natural language responses based on querying knowledge bases of ontological nature.

6. Acknowledgements

The research was conducted with the support of a grant from NFDU (Open Funder Registry: 10.13039/100018227) under project number 2021.01/0136 (2022–2024, project in progress) on "Development of a cloud platform for patient-centric telerehabilitation of oncology patients based on mathematical modeling" [6], based at the Institute of Cybernetics named after V.M. Glushkov of the National Academy of Sciences of Ukraine.

References

[1] K. Malakhov, M. Petrenko, E. Cohn, Developing an ontology-based system for semantic processing of scientific digital libraries, in: South African Computer Journal, Vol. 35, No. 1. (2023) 19–36. doi: 10.18489/sacj.v35i1.1219.

[2] O. Palagin, M. Petrenko, M. Boyko, Ontology-related Complex for Semantic Processing of Scientific Data, in: Proceedings of the 13th International Scientific and Practical Programming

Conference UkrPROG 2022, Kyiv, Ukraine, 2022, vol. 3501. pp. 279 – 290. Available from: https://ceur-ws.org/Vol-3501/s26.pdf
[3] M. Petrenko, E. Cohn, O. Shchurov, K. Malakhov, Ontology-Driven Computer Systems: Elementary Senses in Domain Knowledge Processing, in: South African Computer Journal, Vol. 35, No.2. (2023) 127 – 144. doi: 10.18489/sacj.v35i2.17445.
[4] K. S. Malakhov, Insight into the Digital Health System of Ukraine (eHealth): Trends, Definitions, Standards, and Legislative Revisions, in: International Journal of Telerehabilitation, Vol. 15, No. 2. (2023) 1 – 21. doi: 10.5195/ijt.2023.6599.
[5] K. S. Malakhov, Letter to the Editor – Update from Ukraine: Development of the Cloud-based Platform for Patient-centered Telerehabilitation of Oncology Patients with Mathematical-related Modeling, in: International Journal of Telerehabilitation (2023) 1–3. doi: 10.5195/ijt.2023.6562.
[6] K. Malakhov, Letter to the Editor – Update from Ukraine: Rehabilitation and Research, in: International Journal of Telerehabilitation, Vol. 14, No. 2. (2022) 1–2. doi: 10.5195/ijt.2022.6535.
[7] H. Inefuku, K. Malakhov, E. R. Cohn, L. B. Collister, Service Diversification, Connections, and Flexibility in Library Publishing: Rapid Publication of Research from Ukraine in Wartime, in Case Studies in Library Publishing, Vol.1, No.1. (2023) Available from: https://cslp.pubpub.org/pub/084se42n/release/1
[8] O. V. Palagin, K. S. Malakhov, V. Yu. Velychko, T. V. Semykopn, Hybrid e-rehabilitation services: SMART-system for remote support of rehabilitation activities and services, in: International Journal of Telerehabilitation, Special Issue: Research Status Report – Ukraine (2022) 1–16. doi: 10.5195/ijt.2022.6480.
[9] O. Palagin, V. Kaverinskiy, K. Malakhov, M. Petrenko, Fundamentals of the Integrated Use of Neural Network and Ontolinguistic Paradigms: A Comprehensive Approach, in: Cybern. Syst. Anal., Vol. 60. (2024) 111–123. doi: 10.1007/s10559-024-00652-z.
[10] O. Chebanyuk, An Approach of Text to Model Transformation of Software Models, in: Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering, SCITEPRESS - Science and Technology Publications, Funchal, Madeira, Portugal, 2018: pp. 432–439. doi:10.5220/0006804504320439
[11] O. Chebanyuk, Formal foundations for software model to model transformation operation. in 11th International Scientific and Practical Conference from Programming (UkrPROG'2018), Kyiv, Ukraine, 2018 CEUR Workshop Proceedings, 2139, pp. 124-131. Available from: https://ceur-ws.org/Vol-2139/124-131.pdf
[12] O. Chebanyuk, Investigation of Drawbacks of the Software Development Artifacts Reuse Approaches based on Semantic Analysis. Advances in Computer Science for Engineering and Education VI Lecture Notes on Data Engineering and Communications Technologies, 2023, p. 514-523 514-523. doi: 10.1007/978-3-031-36118-0_46
[13] A. Doe, B. Smith, Advances in Semantic Web Technologies, Journal of Web Semantics, Vol. 22. (2024) 15-30. doi: 10.1016/j.websem.2024.01.002.
[14] C. Johnson, D. Brown, Integrating AI with Ontological Knowledge Bases, Artificial Intelligence Review, Vol. 39. (2024) 45-59. doi: 10.1007/s10462-024-09876-3.
[15] E. Green, F. White, Ontology-Driven Systems in Healthcare, Health Informatics Journal, Vol. 30. (2024) 101-115. doi: 10.1177/1460458224.
[16] Liu, J., & Zhang, Y. (2024). The Role of Semantic Web in Enhancing Data Interoperability. International Journal of Web Information Systems, 20(2), 50-66. doi: 10.1108/IJWIS-01-2024-0005.
[17] Patel, S., & Kumar, R. (2024). Challenges and Opportunities in Ontology-Based AI Systems. Journal of Artificial Intelligence Research, Vol. 65, 33-48. doi: 10.1613/jair.65.33.
[18] Smith, L., & Johnson, M. (2024). Ontology-Driven Decision Support Systems in Clinical Practice. Journal of Medical Systems, 48(3), 12-26. doi: 10.1007/s10916-024-01780-3.
[19] A. A. Litvin, V. Yu. Velychko, V. V. Kaverinsky, Synthesis of chat-bot responses in the natural language of the flexive type based on the results of formal questions to ontology and semantic analysis of the initial phrase, in International Journal "Information Content and Processing", Vol. 7, No. 1 (2020).

[20] O. Palagin, V. Kaverinskiy, K. Malakhov, A. Litvin, OntoChatGPT Information System: Ontology-Driven Structured Prompts for ChatGPT Meta-Learning, in International Journal of Computing, Vol. 22, No. 2. (2023.) 170 – 183. doi: 10.47839/ijc.22.2.3086
[21] Levenshtein distance. [cited 20 Mar 2024]. https://en.wikipedia.org/wiki/Levenshtein_distance.
[22] A.F. Kurgaev, N.G. Petrenko, Processor structure design, in Cybern Syst Anal., Vol. 31. (1995) 618–625. doi: 10.1007/BF02366417.
[23] N. G. Petrenko, A. A. Sofiyuk, On one approach to the transfer of an information structures interpreter to PLD-implementation, in Upravlyayushchie Sistemy i Mashiny, No. 6. (2003) P. 48 – 57.