CYBERNETICS
and COMPUTER
TECHNOLOGIES

Automating the formation of the knowledge base's conceptual structure is crucial for predictive modeling of patient trajectories, which is essential for effective treatment and resource optimization in rehabilitation. This process addresses the multifactorial nature of medical data and integrates information from various sources, overcoming the limitations of traditional databases. This article presents a novel approach to automating the conceptual structure formation of a medical knowledge base using deep learning techniques. By leveraging BioBERT and word embeddings, we aim to enhance the extraction and integration of symptoms from medical notes. BioBERT, a powerful pre-trained language model specifically designed for biomedical text, is employed to identify relevant symptoms within unstructured clinical data. Symptoms are transformed into word embeddings for precise comparisons using cosine similarity, identifying synonyms and updating the knowledge base. This automated methodology ensures comprehensive, current medical information, enhancing clinical decision support and patient outcomes.

Keywords: *Knowledge-Oriented Management Systems, knowledge base, Support Vector Machine, Word2Vec, Skip-Gram, BioBERT.*

UDC 004.8:004.91 DOI:10.34229/2707-451X.24.4.10

D. SYMONOV

AUTOMATING THE FORMATION OF THE CONCEPTUAL STRUCTURE OF THE KNOWLEDGE BASE USING DEEP LEARNING

Introduction. One of the key aspects of modern information technology is the ability to automate processes. Building and using the conceptual structure of the knowledge base is becoming an essential need in the modern world, where the amount of information is growing exponentially. Accordingly, the ability to automate processes, including the construction of ontologies, which requires the extraction of knowledge from full-text sources and their automatic structuring, is important. Knowledge bases are used to manage complex dynamic systems [1] by ensuring the storage, organization, and access to a large amount of information that allows for effective analysis and prediction of the behavior of such systems. Knowledge bases simplify the choice of models for the functioning of multicomponent information systems [2] by providing centralized access to structured information and templates. This allows you to find optimal solutions faster and adapt models to changing conditions, increasing the efficiency and flexibility of systems.

The conceptual structure of a knowledge base is a system of organized concepts and relationships between them, which allows for more efficient storage, retrieval, and use of information. It has several key advantages that make it an essential tool in many industries:

1) Data unification and systematization: Its role in large volumes of data is critical to ensure accuracy and reliability in both research and other areas of activity.

2) Automation support: The conceptual structure of the knowledge base allows automating data processing with machine learning and artificial intelligence algorithms that require structured data, increasing the efficiency of information systems and reducing time and resources.

3) Integration and analysis: The presence of a conceptual structure of the knowledge base facilitates the integration of different data sources into a single system [2], which facilitates the comparison and analysis of information, which is critical for transdisciplinary research.

4) Creation of new knowledge: The conceptual structure allows to identify new connections between concepts, which opens up new opportunities for research and development of new technologies.

Developing a conceptual structure for a knowledge base plays a key role in healthcare data management, especially in the face of the growing complexity of healthcare. It not only organizes information efficiently, but also creates the basis for building predictive models that significantly improve the prediction of patient treatment trajectories. The use of the conceptual structure of the knowledge base allows to improve the quality of variable selection, automate the process of variable selection, improve the interpretability of models and reduce multicollinearity, which ultimately increases the accuracy of models. Thanks to these advantages, analysts can create more efficient and reliable regression models, which significantly improves decision support in various industries.

Problem's Formulation. Rehabilitation of patients after various diseases or injuries requires accurate prediction of the recovery trajectory to ensure effective treatment and optimization of resources. The modeling process must take into account a large number of factors, such as patient complaints, laboratory values, medical history, age, and gender. Accordingly, building such predictive models is challenging due to the multifactorial nature of medical data, variability of clinical pathways, and individual patient characteristics.

Another major problem with modeling and predicting patient trajectories is the lack of structured medical information and the difficulty of integrating data from different sources. Traditional databases used in medicine often fail to adequately reflect the dynamics and contextual information that are critical for forecasting, as these concepts are used as variables in regression models.

The importance of automating the formation of the conceptual structure of the knowledge base lies in the ability to create flexible and scalable systems that can integrate new knowledge and adapt to changes in medical practice. The use of deep learning for this task opens up opportunities for building decision support systems that can quickly respond to changes in the patient's condition and provide more accurate forecasting of their trajectories. This, in turn, helps to improve the quality of medical services and optimize the use of resources in the healthcare system. Thus, automating the formation of the conceptual structure of the knowledge base using deep learning is a key step in creating predictive models of patient trajectories, which allows solving existing problems and ensuring more efficient medical practice.

Solving of the dual problem. Knowledge-Oriented Management Systems [3] play a key role in the development of predictive models due to their ability to integrate, process and analyze large amounts of data. These systems provide a structuring of knowledge that allows for more efficient use of information to create models that can predict various events and trends.

Knowledge-Oriented Management Systems are based on the principles of organizing and managing knowledge, which includes collecting, storing, processing, and transmitting information. Thanks to these principles, they become a source of structured and unstructured data that can be used to train deep models. For example, data collected from various sources, such as scientific articles, technical reports, and other information resources, can be integrated into one common knowledge base, which simplifies access to the necessary information and increases the accuracy of forecasts.

Deep learning, which is used to automate the formation of the conceptual structure of a knowledge base, is able to work effectively with large amounts of data obtained from Knowledge-Oriented Management Systems. By using machine learning algorithms, such systems can identify patterns and relationships between different data elements, which allows them to create more accurate and reliable predictive models.

In the context of deep learning, knowledge bases represented in the form of semantic networks can serve as an effective source of information. Thanks to structured relationships between data, such databases allow the model to recognize patterns faster and more accurately, which is important for building predictive models. A semantic network can include concepts (notions), events, characteristics (properties), and their values, which provides a multidimensional representation of information and improves the quality of forecasting. Semantic networks allow you to form complex hierarchical structures, where concepts can belong to one or more classes, and classes, in turn, can be parts of other classes. Such a multi-level structure facilitates flexible and accurate knowledge representation, making it ideal for use in deep learning, where it is important to take into account different aspects and levels of information. The knowledge base, organized as a semantic network, also allows for various data operations, such as creating class instances, establishing and breaking relationships between classes, selecting instances, etc. This provides high flexibility in data management and facilitates the integration of new knowledge into the system. Using such operations allows the deep learning model to dynamically adapt to changes in the knowledge base and improve its predictive capabilities.

Thus, Knowledge-Oriented Management Systems are an important source for predictive models for determining the trajectory of patient rehabilitation, as they provide the necessary data and facilitate their effective use. The integration of Knowledge-Oriented Management Systems with deep learning methods can significantly improve the quality of forecasts and increase management efficiency through the use of constantly updated knowledge. This, in turn, promotes innovation and improves the quality of patient care.

The patient's rehabilitation trajectory is a complex and multidimensional process that includes a sequence of medical, physical, psychological and social measures aimed at restoring the patient's health and functional capabilities after an illness or injury. Effective rehabilitation depends on an accurate understanding of the patient's individual needs, predicting potential complications, and adapting rehabilitation measures at each stage of treatment. Modeling the patient's rehabilitation trajectory is an important task that can be solved with the help of modern forecasting technologies, including deep learning and artificial intelligence.

One of the main challenges of building predictive models of rehabilitation trajectories is the need to take into account a large number of factors that can affect the recovery process. These factors can be very diverse, including the patient's genetic characteristics, lifestyle, physical activity level, socioeconomic status, and others. In addition, rehabilitation processes are often non-linear and may depend on dynamic changes in the patient's condition.

Another problem is ensuring the accuracy and reliability of the models. Insufficient data, poor quality, or incorrectly specified parameters can lead to inaccurate predictions, which in turn can negatively affect the rehabilitation process. Therefore, it is important to collect and use a large amount of high-quality data, as well as to continuously improve models based on new knowledge and data. The use of knowledge bases can significantly improve modeling results. Knowledge bases that contain information about typical rehabilitation trajectories, clinical guidelines, and the experience of physicians allow for more accurate and adaptable models.

Automating the formation of the conceptual structure of the knowledge base with the help of deep learning allows you to systematize and integrate a large amount of medical information. This creates the basis for effective modeling of the patient's rehabilitation trajectory and clinical decision support. The application of such approaches will help to improve the quality of medical care and improve patient rehabilitation outcomes.

Suppose there is a knowledge base $Z = \bigcup_{q=1}^{Q} Z_q$ consisting of sets of criteria points $X = \bigcup_{l=1}^{L} X_l$, $X_l = \left\{x_i^l\right\}, i = \overline{1,n}$ sets of knowledge base state points $Z_q = \left\{z_i^q\right\}, i = \overline{1,n}$, and values of the objective function at each stage of the rehabilitation process y_s. A generalized system for predicting the patient's rehabilitation trajectory is shown in fig. 1.

Modeling a patient's rehabilitation trajectory includes several steps (fig. 1). The first step is the collection and preliminary processing of data related to the patient's health status, medical history, therapies, test results, and other relevant factors that form a set of variables $\{x_i\}$. This data can be both structured and unstructured, which makes it difficult to process and analyze. This stage involves processing the available knowledge $\{z_i\}$ in the knowledge base and identifying information that will be useful for building a forecast model. The model quality assessment at each stage of the rehabilitation cycle t_s is based on the level of variance $D(y)$, which reflects the stability and accuracy of the model's predictions in patient recovery.

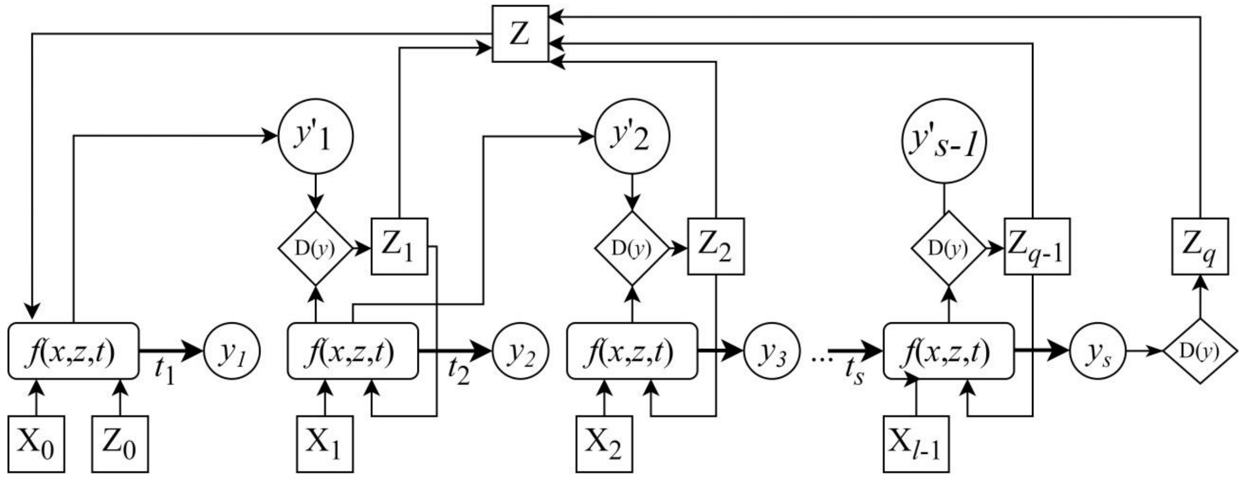

FIG. 1. A scheme for using knowledge to predict a patient's trajectory

The next step is to create a model that can take this data into account and predict the likely outcomes of rehabilitation. As a forecasting method, it is advisable to use a dynamic SVM (Support Vector Machine) regression model that allows you to adapt the forecasting results at each stage of rehabilitation, including by dynamically updating the available knowledge $\{z_i\}$ in the knowledge base.

Patient trajectory prediction using SVM method. There are several important advantages to using a dynamic SVM [4] regression model, especially in the context of data that changes over time.

The dynamic SVM regression model can adapt to new data and variables added over time. This is especially important in cases where the data structure changes or new factors that affect the outcome appear (fig. 1). The use of incremental learning allows the model to update its parameters without the need for complete retraining, which saves computing resources. By incorporating a time parameter or other aspects, a dynamic SVM regression model can take into account the time dependence in the data. This allows the model to more accurately predict future values based on previous observations, which is important for time series analysis and trend forecasting.

Since the functioning of the rehabilitation cycle forecasting system (fig. 1) involves the development of a knowledge base at each stage of the rehabilitation process, the process function can be represented as follows:

$$f(x,z,t) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\cdot K\big((x_i,z_i,t_i),(x,z,t)\big) + b, \tag{1}$$

where α_i, α_i^* are Lagrange multipliers; $K\big((x_i,z_i,t_i),(x,z,t)\big)$ is a kernel function that calculates a scalar product in the feature space; $b \in \mathbb{R}$ is a bias.

The combined kernel function will look like this:

$$K\Big((x_i,z_i,t_i),\big(x_j,z_j,t_j\big)\Big) = K\big(x_i,x_j\big) + \big(z_i,z_j\big) + \big(t_i,t_j\big). \tag{2}$$

Accordingly, the objective function will have the form:

$$\min_{\alpha_i,\alpha_i^*}\left\{\begin{matrix}\frac{1}{2}\sum_{i,j=1}^{n}\Big((\alpha_i-\alpha_i^*)\cdot\big(\alpha_j-\alpha_j^*\big)\Big)\cdot K\Big((x_i,z_i,t_i),\big(x_j,z_j,t_j\big)\Big)+ \\ +\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)-\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)\end{matrix}\right\}, \tag{3}$$

subject to the constraints:

$$\begin{cases} 0 \le \alpha_i \le C \\ 0 \le \alpha_i^* \le C \\ \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \end{cases}. \quad (4)$$

Accordingly, the patient's trajectory can be predicted using the following equation:

$$\hat{y} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) \cdot K\big((x_i, z_i, t_i), (x, z, t)\big) + b, \quad (5)$$

where the kernel function $K\big((x_i, z_i, t_i), (x, z, t)\big)$ depends on the data.

A dynamic SVM regression model can use different kernel functions, which allows it to take into account both linear and nonlinear relationships between variables [5]. Using adaptive kernels that can change over time adds additional flexibility and allows the model to better adapt to new conditions. The following parameters can be used to predict the patient's rehabilitation trajectory:

- $K(x_i, x_j) = e^{\left(-\gamma_x \|x_i - x_j\|^2\right)}$;
- $K(z_i, z_j) = \left(\langle z_i, z_j \rangle + C_z\right)^d$;
- $K(t_i, t_j) = \langle t_i, t_j \rangle$,

where γ_x is a parameter that affects the form of the Radial Basis Function of the kernel; C_z is the control coefficient of the inner product of the polynomial kernel; d is a polynomial degree.

Thanks to the use of regularization, SVM models are resistant to overfitting, which is important in dynamic environments where data can change over time. This ensures the stability and reliability of the model in the long run.

Since the predictive model uses both measured data (medical test results, medical history, etc.) and textual data (e.g., patient history), the main difficulty lies in processing textual information and extracting knowledge from it to be used as input to the model.

Automating the formation of the conceptual structure of the knowledge base. As previously mentioned, after the development of the regression model, the main difficulty arises with the transformation of the input textual information of the patient's history into a numerical space for further analysis [6].

The algorithm for the automated formation of the conceptual structure of the knowledge base involves the implementation of two key stages:

1) Conversion of text information into numerical space;
2) Definition of conceptual concepts and verification of their presence in the knowledge base.

After completing the two above-mentioned stages, the regression model of the patient's rehabilitation is retrained, taking into account changes in the knowledge base [7].

Word2Vec is one of the most popular models for converting words to vectors in a multidimensional space such that words that have similar meanings have similar vectors [8]. The Word2Vec model is good at generating vectors for words that do not occur frequently in a text using the contexts of similar words, and vectors generated on one corpus of text can be used for tasks on other texts, which reduces the need for computing resources and training time.

To automate the processing of incoming textual information, we will use the Skip-Gram architecture [9].

Let ω_i be a given word from the patient's history, ω_{cont} be contextual words, then the probability of a contextual word can be calculated using the following formula:

$$P(\omega_{cont}, \omega_i) = \frac{\exp\left(V_{\omega_{cont}}^T, V_{\omega_i}\right)}{\sum_{i=1}^{\|V\|} \exp\left(V_{\omega_{cont}}^T, V_{\omega_i}\right)}, \quad (6)$$

where V_{ω_i} is a the vector of representation of the word ω_i; V is a word size.

The task of using the method is to maximize the probability of predicting the correct contextual words [10]. Accordingly, the objective function involves minimizing the cost function:

$$\mathcal{L} = \sum_{i=1}^{\|V\|} \sum_{-C \le j \le C} \log P\left(\omega_{i+j}, \omega_i\right), \tag{7}$$

where C is a the size of the context window.

The stochastic gradient descent method is used to train model Skip-Gram [11].

BioBERT (Biomedical Bidirectional Encoder Representations from Transformers) is a pre-trained language model specifically designed for biomedical text processing tasks [12]. The model is based on the BERT (Bidirectional Encoder Representations from Transformers)[13] architecture, but is additionally trained on a large biomedical corpus.

BioBERT is based on the transformer architecture, in particular on the variant that uses encoders [14]. The main elements of the mathematical model include the following:

- Transformer-encoder: $H_0 = [x_i \mathrm{E} | i = \overline{1,n}]$, where H_0 are the input vector representations, x_i are incoming tokens, E is the investment matrix.
- Self-Attention mechanism: $\mathrm{A(Q,K,V)} = \mathrm{softmax}\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}$, where $\mathrm{(Q,K,V)}$ are matrices of queries, keys, and values, d_k is the dimension of the keys.
- MultiHead Self-Attention: $M(\mathrm{Q,K,V}) = \mathrm{concat}(h_1, h_2, \dots, h_\mathrm{m})\mathrm{W}^u$, where h_m is a weight matrix for each head, W^u is a weight matrix for combined heads.
- Feed-Forward layer: $\mathrm{FNN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$, where W_1, W_2 are weighting matrices, b_1, b_2 are the displacement vectors.

The model produces vector representations for each input token that can be used for various NLP tasks, including classification, entity extraction, etc [15]. BioBERT achieves high accuracy on biomedical tasks due to additional training on specialized biomedical corpora [16], which allows the model to better understand the specific terms and contexts that are typical for this field [17].

An example of automated processing of text information for updating the knowledge base. Assume that the knowledge base has the following seven symptoms:

1) «Diminished energy and increased need for rest not proportional to energy exerted»;
2) «Generalized weakness»;
3) «Diminished mental concentration»;
4) «Insomnia or hypersomnia»;
5) «Sleep is often not restorative»;.
6) «Increased emotional reactivity»;
7) «Muscle weakness or heaviness».

Let's look at how the model works on the examples of patient requests.

Case 1: «The patient was diagnosed with "C50: Malignant neoplasm of the breast". He complains of constant fatigue, prolonged headaches and depression. As a result of cognitive decline, the patient was forced to quit his job, which negatively affected his quality of life».

Case 2: «A 49-year-old man complains of acne, which is disturbing him and his wife. The patient is overweight and has been diagnosed with C34: Malignant neoplasm of the bronchi and lungs. As a result of acne, the man is constantly sleep deprived, which negatively affects his performance and cognitive abilities. As a result of the prolonged lack of quality sleep, the patient is constantly tired».

Based on the results of processing the cases, the algorithm identified a subset of disease symptoms: ['fatigue', 'headaches', 'depression'], ['acne', 'acne', 'sleep deprivation', 'lack of quality sleep', 'tired']

As you can see from the results, the result subset may contain both duplicates (e.g., 'acne') and synonyms ('sleep deprived' and 'lack of quality sleep'), which requires the inclusion of duplicate and synonym removal tools in the algorithm. Accordingly, after processing, the updated subset of symptoms will look like this: ['depression', 'fatigue', 'headaches'], ['acne', 'lack of quality sleep', 'tired'].

After forming a subset of potential candidates for inclusion in the knowledge base, it is necessary to check for synonymous symptoms in the knowledge base.

As can be seen from the screenshot of the program (Table 1), the problem of "lack of quality sleep" has a high similarity coefficient with the symptom "Sleep is often not restorative" in the knowledge base, with a value of 0.767688. Accordingly, these two concepts can be considered one variable in the regression model.

TABLE 1. Example of checking for synonymous symptoms in the knowledge base

new_symptom	knowledge_base_symptoms	cosine_similarity
lack of quality sleep	Diminished energy and increased need for rest	0.546932
lack of quality sleep	Generalized weakness	0.454075
lack of quality sleep	Diminished mental concentration	0.446238
lack of quality sleep	Insomnia or hypersomnia	0.628588
lack of quality sleep	Sleep is often not restorative	0.767688
lack of quality sleep	Increased emotional reactivity	0.420785
lack of quality sleep	Muscle weakness or heaviness	0.452938

To determine the threshold on which to base the synonym/non-synonym decision, we generated 10 synonyms for each of the 7 existing symptoms in the knowledge base. On the screenshot of the program, you can see the results of calculating the word nesting representation for each symptom (Table 2).

TABLE 2. The symptoms and their word embedding representations

cluster_id	symp_id	symp_name	symp_name_emb
0	0	Diminished energy and increased need for rest ...	[-0.012109594, 0.0141017195, -0.017949356, ...]
0	1	Feeling tired all the time, even after resting.	[0.03482404, 0.0142817395, -0.021438556, ...]
0	2	Low energy levels that don't improve with rest.	[0.0041652583, 0.019854924, -0.030666716, ...]
0	3	Needing more sleep than usual but still feelin...	[0.0576977, 0.0075445045, -0.05864109, ...]
0	4	Experiencing fatigue that's out of proportion ...	[0.04095247, 0.012383495, -0.027395414, ...]
…	…	…	…
6	72	Experiencing a decline in muscle power and end...	[0.03379547, -0.0057901912, -0.017707108, ...]
6	73	Difficulty performing tasks due to weak or hea...	[0.029822746, -0.02132496, 0.0044572097, ...]
6	74	A feeling of sluggishness and stiffness in you...	[0.04819323, 0.046189692, -0.015851222, ...]
6	75	Muscles that feel tired and sore even after mi...	[0.028195446, 0.04905017, -0.0085209925, ...]
6	76	Reduced muscle function that impacts	[0.00367627, 0.0057160724, -0.011074149, ...]

Based on values from Table 2, we calculated the cosine similarity for each pair of symptoms and the average value between each group of symptom synonyms to estimate which value can be used to find synonyms. Based on the results in the screenshot of the program (Table 3), a value of 0.6 seems to be a suitable threshold for determining whether pairs of symptoms are synonyms or not.

TABLE 3. Average cosine similarity between each group of symptom synonyms

		cluster_id_2						
		0	1	2	3	4	5	6
cluster_id_1	0	0.69854	0.60196	0.52702	0.60470	0.62916	0.50936	0.59716
	1	NaN	0.69185	0.48015	0.45526	0.49701	0.48997	0.69313
	2	NaN	NaN	0.67933	0.48987	0.46258	0.49050	0.45400
	3	NaN	NaN	NaN	0.72060	0.67227	0.46527	0.45859
	4	NaN	NaN	NaN	NaN	0.71108	0.46047	0.50537
	5	NaN	NaN	NaN	NaN	NaN	0.69661	0.45418
	6	NaN	NaN	NaN	NaN	NaN	NaN	0.72387

The cosine similarity value indicates the degree of similarity between two vectors in a multidimensional space. It is defined as the cosine of the angle between two vectors, and takes values from –1 to 1:

- "1" means that the vectors are exactly the same (there is no angle between them, i.e. they have the same direction);
- "0" means that the vectors are orthogonal (perpendicular), i.e., they have no similarity;
- "-1" means that the vectors are completely opposite.

This measure is often used in text analytics to measure the similarity between documents or sentences when they are represented as vectors, such as term vectors or embedding vectors.

The similarity value for the symptoms of Case 1, relative to the symptoms in the knowledge base:
[0.487974, 0.493705, 0.4999892, 0.5078797, 0.4660168, 0.488581, 0.4916319],
[0.596842, 0.526782, 0.4882978, 0.5086504, 0.5237366, 0.395079, 0.6332218],
[0.391560, 0.439345, 0.4755709, 0.5052929, 0.4494474, 0.358021, 0.4153259].

For Case 1, the system adds 'depression' and 'headaches' to the knowledge base, and removes 'fatigue' as it considers it synonymous with 'Muscle weakness or heaviness' (0.63322186 > 0.6) and almost synonymous with 'Diminished energy and increased need for rest not proportional to energy exerted' (0.5968429 ≈ 0.6).

The similarity value for the symptoms of Case 2, relative to the symptoms in the knowledge base:
[0.313179, 0.355903, 0.321295, 0.3315830, 0.3172901, 0.4159766, 0.355421],
[0.563786, 0.433951, 0.512197, 0.6053250, 0.6936647, 0.4411898, 0.475435],
[0.617111, 0.458519, 0.494658, 0.6002520, 0.5899190, 0.4068375, 0.571751].

Only the value "acne" is added to the knowledge base for Case 2, all other values have at least one synonym in the knowledge base.

Thus, after the visit of two patients, the knowledge base of the medical institution will be supplemented with three new concepts that will be added to the regression model of the patient rehabilitation trajectory.

Conclusions and further research. In this study, we have demonstrated the efficacy of using deep learning techniques, specifically BioBERT and word embeddings, to automate the formation of the conceptual structure of a medical knowledge base. By extracting and integrating symptoms from medical notes and comparing them with existing entries, we have shown that this approach can enhance the filling and comprehensiveness of medical knowledge bases. The application of cosine similarity for synonym identification ensures precise updates and maintenance of the database.

Automating the knowledge base formation addresses the multifactorial challenges of medical data and improves the integration of unstructured information from various sources. This advancement is particularly valuable in building predictive models for patient trajectories, essential for effective rehabilitation and resource optimization. The use of deep learning creates adaptable, scalable systems that respond swiftly to changes in patient conditions, enhancing decision support and forecasting accuracy.

Overall, our approach not only streamlines the maintenance and expansion of medical knowledge bases but also contributes to improved patient care and optimized healthcare resource utilization. Future work will focus on refining these techniques and exploring their application in other areas of medical informatics.

This article did not demonstrate the operation of a regression model using knowledge as a variable. This task remains beyond the scope of the current study. Further research will be aimed at developing and testing such models.

Funding. Presented research and authorships were supported by grant 0123U102942.

References

1. Symonov D., Symonov Y. Methods for selecting models of functioning of multicomponent information and environmental systems. *Scientific Journal «Mathematical Modeling»*. 2024. No. 1. P. 57–63. https://doi.org/10.31319/2519-8106.1(50)2024.304943
2. Symonov D.I., Zaika B.Y. Modeling the management of complex information multicomponent systems. *Scientific Bulletin of Uzhhorod University, Series of Mathematics and Informatics*. 2024. No. 1. P. 168–174. (in Ukrainian) https://doi.org/10.24144/2616-7700.2024.44(1)
3. Petrenko M., Palagin O., Boyko M., Matveyshyn S. Knowledge-Oriented Tool Complex for Developing Databases of Scientific Publications and Taking into account Semantic Web Technology. *Control Systems and Computers*. 2022. No. 3. P. 11–28. (in Ukrainian) https://doi.org/10.15407/csc.2022.03.011
4. Veisi H. Introduction to SVM: Learning with Fractional Orthogonal Kernel Classifiers in Support Vector Machines. Industrial and Applied Mathematics. Singapore: Springer, 2023. P. 3–18. https://doi.org/10.1007/978-981-19-6553-1
5. Telalović Hasić J., Salković A. Breast cancer classification using Support Vector Machines (SVM). Advanced Technologies, Systems, and Applications VIII. Cham: Springer, 2023. P. 195–205. https://doi.org/10.1007/978-3-031-43056-5_16
6. Mallik A., Kumar S. Word2Vec and LSTM based deep learning technique for context-free fake news detection. *Multimed Tools Appl*. 2024. No. 83. P. 919–940. https://doi.org/10.1007/s11042-023-15364-3
7. Johnson S.J., Murty M.R., Navakanth I. A detailed review on word embedding techniques with emphasis on word2vec. *Multimed Tools Appl*. 2024. No. 83. P. 37979–38007. https://doi.org/10.1007/s11042-023-17007-z
8. Sharma A., Kumar S. Ontology-based semantic retrieval of documents using Word2vec model. *Data & Knowledge Engineering*. 2023. No. 144. P. 102110. https://doi.org/10.1016/j.datak.2022.102110
9. Chintawar S., Kulkarni R., Patil N. OntoPred: An efficient attention-based approach for protein function prediction using Skip-Gram features. *SN Comput. Sci*. 2023. No. 4. P. 666. https://doi.org/10.1007/s42979-023-02135-y
10. Yu. T. The design of electronic medical records system using Skip-gram algorithm. *Netw Model Anal Health Inform Bioinforma.* 2021. Vol 10, No. 7. https://doi.org/10.1007/s13721-020-00281-4
11. Preethi P., Sharada A. Word Embeddings - Skip Gram Model, ICICCT 2019 – System Reliability, Quality Control, Safety. *Maintenance and Management.* 2019. P. 133–139. https://doi.org/10.1007/978-981-13-8461-5
12. Zhu Y., Li L., Lu H., Zhou A., Qin X. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *Journal of Biomedical Informatics*. 2020. No. 106. P. 103451. https://doi.org/10.1016/j.jbi.2020.103451
13. Turchin A., Masharsky S., Zitnik M. Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*. 2023. No. 36. P. 101139. https://doi.org/10.1016/j.imu.2022.101139
14. Davagdorj K., Park K.H., Amarbayasgalan T., Munkhdalai L., Wang L., Li M. & Ryu K.H. BioBERT based efficient clustering framework for biomedical document analysis. *Genetic and Evolutionary Computing*. 2022. P. 179–188. https://doi.org/10.1007/978-981-16-8430-2_17
15. Paganelli M., Tiano D. & Guerra F. A multi-facet analysis of BERT-based entity matching models. *The VLDB Journal.* 2023. https://doi.org/10.1007/s00778-023-00824-x
16. Jamshidi S., Mohammadi M., Bagheri S., Esmaeili N.H., Rezvanian A., Gheisari M., Ghaderzadeh M., Shahabi A.S., Wu Z. Effective text classification using BERT, MTM LSTM, and DT. *Data & Knowledge Engineering*. 2024. No. 151. P. 102306. https://doi.org/10.1016/j.datak.2024.102306

17. Jatnika D., Bijaksana M.A., Suryani A.A. Word2Vec model analysis for semantic similarities in English words. *Procedia Computer Science*. 2019. No. 157. P. 160–167. https://doi.org/10.1016/j.procs.2019.08.153

Received 07.08.2024

Denys Symonov,
PhD, researcher, V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv.
denys.symonov@gmail.com
https://orcid.org/0000-0002-6648-4736

MSC 68T05:68U05

Denys Symonov

Automating the Formation of the Conceptual Structure of the Knowledge Base Using Deep Learning

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv
Correspondence: denys.symonov@gmail.com

Introduction. The ability to automate processes is a key aspect of modern information technology. The construction and use of the conceptual structure of the knowledge base is becoming an urgent need in the modern world, where the amount of information is growing exponentially. The ability to automate processes, including the construction of ontologies, which requires the extraction of knowledge from full-text sources and their automatic structuring, is important. Knowledge bases are used to manage complex dynamic systems by ensuring the storage, organization, and access to a large amount of information that allows for effective analysis and prediction of the behavior of such systems.

The purpose of the paper. The purpose of the paper is to demonstrate the effectiveness of using deep learning methods to automate the formation of the conceptual structure of the knowledge base. The study also aims to show how the integration of knowledge bases with deep learning methods can improve the quality of forecasts and increase the efficiency of rehabilitation trajectory management.

Results. The algorithm successfully extracted and processed symptom information from the medical cases, effectively handling duplicates and synonyms. The utilization of cosine similarity enabled the identification of synonymous symptoms within the established knowledge base, facilitating the seamless integration of new information while preventing redundancy. The system demonstrated its capability to discern which symptoms should be incorporated into the knowledge base and which should be omitted based on their similarity to existing entries. The outcomes underscore the potential of this automated approach to enhance the knowledge base and contribute to the refinement of predictive models within the healthcare domain.

Conclusions. The study demonstrated the effectiveness of deep learning in automating the formation of the conceptual structure of a medical knowledge base. The approach enhances the filling and comprehensiveness of the knowledge base, which is crucial for building predictive models for patient trajectories and improving healthcare decision support.

Keywords: Knowledge-Oriented Management Systems, knowledge base, Support Vector Machine, Word2Vec, Skip-Gram, BioBERT.

УДК 004.8:004.91

Д. Симонов

Автоматизація формування понятійної структури бази знань з використанням глибокого навчання

Інститут кібернетики імені В.М. Глушкова НАН України, Київ
Листування: denys.symonov@gmail.com

Вступ. Можливість автоматизації процесів – це ключовий аспект сучасних інформаційних технологій. Побудова та використання концептуальної структури бази знань стає нагальною потребою у сучасному світі, де кількість інформації зростає в геометричній прогресії. Важлива можливість – це автоматизація процесів, зокрема побудови онтологій, що вимагає вилучення знань з повнотекстових джерел та їх автоматичного структурування. Бази знань використовуються для управління складними динамічними системами, забезпечуючи зберігання, організацію та доступ до великого обсягу інформації, що дозволяє ефективно аналізувати та прогнозувати поведінку таких систем.

Мета роботи – продемонструвати ефективність використання методів глибинного навчання для автоматизації формування концептуальної структури бази знань. Дослідження також має на меті показати, як інтеграція баз знань з методами глибинного навчання може покращити якість прогнозів та підвищити ефективність управління траєкторією реабілітації пацієнтів.

Результати. Алгоритм успішно вилучав і обробляв інформацію про симптоми з медичних документів, ефективно справлявся з дублікатами та синонімами. Використання косинусної подібності дозволило ідентифікувати синонімічні симптоми в існуючій базі знань, що полегшило інтеграцію нової інформації, запобігаючи при цьому надлишковості та дублікатів. Система продемонструвала свою здатність визначати, які симптоми слід включити до бази знань, а які слід вилучити на основі їхньої схожості з уже існуючими записами. Результати підкреслюють потенціал цього автоматизованого підходу до розширення бази знань і сприяють вдосконаленню прогностичних моделей у сфері охорони здоров'я.

Висновки. Дослідження продемонструвало ефективність глибинного навчання для автоматизації формування концептуальної структури медичної бази знань. Підхід підвищує наповнюваність та повноту бази знань, що має вирішальне значення для побудови прогностичних моделей траєкторій реабілітації пацієнтів та покращення підтримки прийняття рішень у сфері охорони здоров'я.

Ключові слова: знання-орієнтовані системи управління, база знань, Support Vector Machine, Word2Vec, Skip-Gram, BioBERT.