Environment. Technology. Resources. Rezekne, Latvia
Proceedings of the 13th International Scientific and Practical Conference. Volume 2, 38-43

Research of Approaches to the Recognition of Semantic Images of Scientific Publications Based on Neural Networks

Iuliia Bruttan
Institute of Engineering Sciences
Pskov State University
Pskov, Russia
bruttan@mail.ru

Igor Antonov
Institute of Engineering Sciences
Pskov State University
Pskov, Russia
igorant63@yandex.ru

Dmitry Andreev
Institute of Engineering Sciences
Pskov State University
Pskov, Russia
dandreev60@mail.ru

Victor Nikolaev
Institute of Engineering Sciences
Pskov State University
Pskov, Russia
nvv60@mail.ru

Tatyana Klets
Institute of Humanities and Linguistic Communications
Pskov State University
Pskov, Russia
kte63@yandex.ru

***Abstract* - The paper is devoted to the problems of orientation and navigation in the world of verbal presentation of scientific knowledge. The solution of these problems is currently hampered by the lack of intelligent information retrieval systems that allow comparing descriptions of various scientific works at the level of coincidence of semantic situations, rather than keywords. The article discusses methods for the formation and recognition of semantic images of scientific publications belonging to specific subject areas. The method for constructing a semantic image of a scientific text developed by Iuliia Bruttan allows to form an image of the text of a scientific publication, which can be used as input data for a neural network. Training of this neural network will automate the processes of pattern recognition and classification of scientific publications according to specified criteria. The approaches to the recognition of semantic images of scientific publications based on neural networks considered in the paper can be used to organize the semantic search for scientific publications, as well as in the design of intelligent information retrieval systems.**

Keywords - semantic image, pattern recognition, semantic search, classification of scientific publications, neural network.

I. INTRODUCTION

In modern conditions, orientation in the continuously increasing volume of scientific publications without the use of automated tools is becoming more and more difficult. Scientists and specialists in different fields do not always manage to successfully track publications containing new significant results in their area of knowledge. The development of automated systems for the classification of scientific publications and the organization of semantic search for scientific information in actual areas of research will increase the efficiency of research work.

Modern approaches to the organization of automated analysis of scientific texts are mainly associated with the use of neural network technologies. The application of machine learning for the classification of scientific publications according to given features will allow creating an automated system for searching scientific publications, as well as increasing the efficiency of searching for the latest publications in a given field of knowledge.

When solving this task, a problem arises related to eliminating the contradiction between the context-dependent representation of texts in natural language and

Online ISSN 2256-070X
https://doi.org/10.17770/etr2021vol2.6628

the context-independent algorithms for their computer processing. Therefore, the authors of the article investigated and proposed approaches to the computer representation of the semantic content of texts in a context-dependent language. The application of these approaches to the construction of a semantic image of a scientific publication can be considered as a variant of the input data formation, taking into account the semantic component for a neural network, the training of which will solve the problem of classifying publications according to given features at a higher quality level and organize a semantic search for texts of scientific content in a given subject area.

It can be argued that at present the global problem of recognition of the semantics of texts has not been solved. Its full solution would lead to a genuine scientific breakthrough. But even a partial solution to this problem, proposed by the authors of the article, seems to be very relevant.

Algorithmically solvable procedures for the recognition of semantic images of texts of scientific publications allow the implementation of intelligent information retrieval systems.

II. Problems of Constructing a Machine Representation of Text in a Natural Language

The texts of scientific publications are unstructured datasets. For automated processing, these unstructured text sequences must be transformed into a structured feature space. The problem lies in developing an approach to constructing a DataSet that is suitable for processing by a neural network and at the same time provides an acceptable level of quality of the results obtained. Obviously, considering the semantics of the source text when forming the DataSet will help to improve the quality.

The existing approaches to processing the text of scientific publications to create a DataSet is described in this paper. Initially, it is required to collect a domain-specific corpus of publications for building models. At the stage of preprocessing the collected text data, it is necessary to perform a number of tasks: removing non-alphabetic characters from the text, splitting the text into a set of tokens, removing stop words, reducing words to their word stem, base or root form. Removing non-alphabetic characters and stop words is a standard procedure. It allows to delete unnecessary elements that have little effect on its general subject matter. Text preprocessing includes the removal of functional words (semantically neutral words such as conjunctions, prepositions, articles, etc.). Next, morphological analysis is performed. One of two ways is used to cast a word to the stem: stemming and lemmatization. In the first case, the ending of the word is cut off according to a certain algorithm, in the second case, the word is reduced to the base or dictionary form of a word in accordance with the applied language grammar. These measures can significantly reduce the dimension of space. As a result, all significant words that appear in the document act as features of the document. After cleansing the data, formal feature extraction methods can be applied.

Document indexing is the construction of a certain numerical model of the text, which translates the text into a representation that is convenient for further processing. The bag-of-words model allows to represent a document as a multidimensional vector of words and their weights in the document [1]. In this case, each document is a vector in a multidimensional space, the coordinates of which correspond to the word numbers, and the values of the coordinates correspond to the values of the weights. Another common indexing model is Word2Vec [2]. It represents each word as a vector that contains information about the context (companion) words. Another indexing model is based on taking into account n-grams [1, 3, 4], that is, sequences of adjacent words.

In the works of D. A. Pospelov, V. Sh. Rubashkin, V. K. Finn, I. A. Melchuk, M. Minsky [5] – [9], classical approaches to the presentation of textual descriptions with a possible level of preservation of semantics for solving search problems and a comparative analysis of these descriptions were proposed:

- frames;
- semantic networks;
- logical models;
- model "Meaning => Text".

To identify the meaning of text-based documents, it is necessary to use semantic analysis, which is realized thanks to a linguistic analyzer. There are a number of problems that arise at the stage of semantic analysis of text-based documents:

- standardization of knowledge representation languages;
- resolution of syntactic and lexical homonymy;
- coreference of relations between units of the text;
- analysis of contexts characterized by semantic incompleteness;
- development of semantic dictionaries required to support semantic analysis algorithms.

It should be borne in mind that for a sufficiently complete understanding of the text from the linguistic analyzer, in addition to the ability to identify and formalize the semantics of the text, the ability to implement logical inference from the text is also required.

The authors' review of publications on this topic allows us to conclude that the mechanism for taking into account the semantic component in existing formalized text models does not make it possible to use even well-known methods to solve the problem of forming a context-sensitive DataSet in which the semantics of the original publication is preserved. Therefore, the authors of the article propose for discussion their approach to the formation of a semantic image of a text, based on the method of spatial representation of text descriptions.

III. Method of Spatial Representation of Text Descriptions

Let us consider an approach to the formation of a semantic image of a text of scientific content, based on the method of spatial representation of text descriptions, developed by the author of the article [10] – [12]. Objects that are specified by textual descriptions can be represented as an area of colored dots in the *N*-dimensional model space of a certain subject area. Thus, a graphic image of the text or a graphic model of a linguistically specified object will be obtained. Consequently, digital image processing methods can be applied to this graphic image, in particular, image recognition methods [13, 14], which are currently well studied and implemented.

Application of the method of spatial representation of text descriptions allows:

- to partially preserve the semantics of the original text, given in a natural language; when it is represented in a computer;
- to use digital image processing methods;
- to implement comparative analysis and search for texts at a higher quality level.

When implementing an approach based on the application of the method of spatial representation of text descriptions, each source text in a natural language will be associated with its graphic image. But in this graphic image of the text it is necessary to add the semantics of the original description, and then it can be called the semantic image of the text description.

One of the best options for presenting texts with preserving the semantics of description in a natural language is the predicate representation of text records in the form of *ARB* syntagmas [12]. The professional language of a scientific direction can be translated into the language of predicates. The authors of the monograph [15] have proved this fundamental possibility.

The predicate language proposed in this paper consists of a descriptor dictionary and a specific description structure, which is a set of elementary statements (syntagmas) of the standard *ARB* form, where *A* and *B* are term codes along with connection pointers, and *R* is a code of a binary relation reflecting the relationship of objects, or features of the subject area under consideration.

The following approaches to the formation of the corresponding dictionaries are proposed.

The lexical composition of the predicate language must be recorded using a descriptor dictionary. A vocabulary should be developed for each domain. Its thematic scope should be such that any texts belonging to the considered subject area are translated into the predicate language. Dictionaries are proposed to be formed on the basis of the constructed domain ontology [16, 17].

The choice of binary relations begins with finding the main types of simple natural language sentences that implement them. In such a sentence, information is recorded about two objects, an object and a feature, or two features and a relation connecting this pair. Each object or feature is most often expressed in one word, which in the sentence plays the role of a subject, addition, definition or circumstance. Attitude is also expressed most often in one word, which in the sentence plays the role of a predicate. Further, for binary relations, formal and logical properties must be established, the operations underlying the identical transformations of elementary statements must be determined. After analyzing the formal-logical properties of the selected binary relations, it becomes clear that many of them are capable of generating new relations. Therefore, it is necessary to conduct a study of the rules for transforming binary relations. Then a dictionary of binary relations is formed with the following structure [12]:

1) descriptor code;
2) head descriptor;
3) synonyms;
4) formal logical properties.

To implement a successful (and adequate) translation of a text description into the predicate language, it is necessary to carry out preprocessing of the text: translation of sentences into a simple form and replacement of pronouns with the corresponding terms.

The translation of a text description from a natural language into a predicate language takes place in two stages:

1) The translation of lexical units of the text.
2) The translation of the links that exist between them.

As a result of translation into the language of predicates of a text in a natural language, a set of syntagmas is formed, consisting of a set of *ARB* that reflect the original description. Thus, we obtain a formalized representation of the text description, which retains the meaning of the text at the level of interrelation of terms of a specific domain.

Further formalization of the text makes it possible to represent its image in the *N*-dimensional model space of the subject area. Suppose the axes of the *N*-dimensional model space are designated $X_1, X_2, \dots, X_N$. Along the axes of the *N*-dimensional space, *m* identical terms of the thesaurus are located, and if necessary, you can add axes to reflect the corresponding quantitative characteristics. Let us assume that the terms in position *A* will be plotted on the odd axes, and the terms in position *B* will be plotted on the even axes. The points in the *N*-dimensional space represent multiple named relationships (*R*) between the corresponding terms of the thesaurus. Then, for example, a text description that belongs to a certain domain of knowledge will represent a separate area filled with named relations. When assigning a specific color to each type of predicate relationship, we obtain different graphic images for the interval. They can be considered models of a linguistically given object with a predicate representation of its context-dependent description (see Fig. 1).

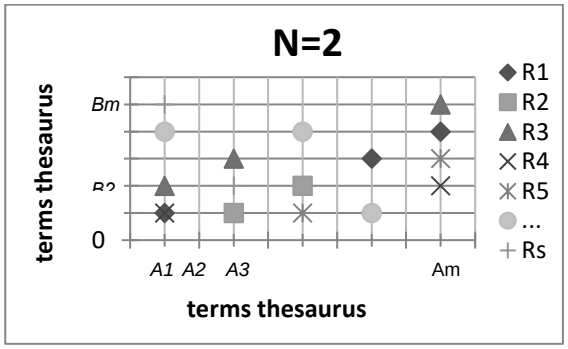

Fig. 1. Semantic image of the text.

It should be emphasized that some relationships often breed others. The order of transformation of statements is determined by the logical connections that exist for predicate relations of textual descriptions of a certain subject area. As a result, simple statements are supplemented with a set of consequences arising from them, which are new statements that were previously absent in the text. But they are necessary to preserve the semantics in the predicate representation, i.e. predicate extensions of the original description appear, containing the results of possible logical transformations of the original predicates. The predicate extensions obtained in this way, in turn, provide coloring of additional subintervals of the thesaurus space.

To build a graphical image of a text description, in most cases (with the exception of text descriptions containing quantitative data), a 2-dimensional space is sufficient (which we will call the base plane). Nevertheless, when constructing a predicate extension of the original description in the model space, a situation may arise when it is required to increase the number of its dimensions, since as a result of performing formal logical transformations, other types of relations between the existing terms may additionally come to light. The maximum number of measurements in the model space *Nmax* can be determined as in (1):

$$Nmax = 2 + 3 + k \quad (1)$$

where 2 – is the number of measurements of the base plane,

3 – is the number of additional dimensions that may appear as a result of performing logical implications of statements containing binary relations,

k – is the number of additional dimensions that are needed to display quantitative data of the original text description.

It should be noted that texts in any natural language belonging to a certain domain will have similar semantic images in the model space of terms of this domain.

IV. Method of Recognition of Semantic Images of Scientific Publications Based on a Neural Network

Let us consider the approach to the recognition of semantic images of scientific publications, created by the authors of the article, based on the method of recognition of semantic images of scientific publications based on a neural network. The input of the neural network receives DataSets, which are semantic images of scientific publications belonging to a certain scientific field. Each semantic image is formed using the above method of spatial representation of text descriptions. This formalization of the texts of scientific publications provides a higher quality of the machine learning model, since it preserves the semantics of the original text description.

The authors propose an approach to the automated formation of a DataSet – a labeled (classified) corpus of semantic images of publications. The DataSet obtained in this way will be used at the next stage for training and testing the neural network. Classification of documents is implemented based on the application of comparison of images of individual documents with a reference image for a category for the degree of their similarity. In recognition methods based on matching, each class is represented by a vector of features of the image that is the prototype of this class. An unfamiliar image is assigned to the class whose prototype is the closest in the sense of a predetermined metric. Let us adapt the approach proposed in [13] for solving the problem of recognizing the semantic image of the text of a scientific publication.

Initially, it is necessary to form semantic images of the corpus of scientific publications belonging to specific domain. At this stage, as a result of the application of the method of spatial representation of textual descriptions, the formation of semantic images of scientific publications is carried out.

Semantic images, as shown above, are a collection of points in the *N*-dimensional model space with axes X_1, X_2,… , X_N.

Each point of each *Yg* semantic image is characterized by coordinates and color (let's assume that the color is specified in the form of *RGB* model). Therefore, each semantic image of a linguistically given object – the text of a scientific publication – can be represented as a matrix of features of the following form (2):

$$Y_g = \begin{pmatrix} y_{11}^g & \cdots & y_{1N}^g & RGB_1^g \\ \vdots & \ddots & \vdots & \vdots \\ y_{M_g 1}^g & \cdots & y_{M_g N}^g & RGB_{M_g}^g \end{pmatrix} \quad (2)$$

where g = 1,2,…,W,

y_{ij}^g *– coordinates of points of the g-th semantic image in the model space (i=1, 2, …, Mg; j=1, 2, …, N),*

RGB_i^g *– colors of points of the g-th semantic image, specified in the form of RGB color model (i =1, 2, …, Mg)*

W – total number of semantic images,

N – the number of dimensions of the model space,

Mg – the number of points in the g-th semantic image of a text description – includes points representing the totality of ARB syntagmas of the original text description and points representing the predicate expansion of the original ARB array obtained as a result of performing formal logical transformations over it.

The set of all semantic images of the text corpus of scientific publications will be a set (3):

$$\{Y\} = \{Y_1, Y_2, \cdots, Y_W\} \quad (3)$$

Let us assume that the points in the matrix of features of the semantic image are listed in ascending order of the coordinates of the points of the image, and if the coordinates coincide, in ascending order of the color values of the corresponding points.

At the next stage, it is required to provide the correspondence of the set of indexes of semantic images to a specific class (domain) of publications. As a result, a DataSet will be obtained from the labeled-up semantic images.

The procedure for recognizing the semantic image of a text description with such a representation of linguistically defined objects can be implemented by comparing the recognizable image of the text, given in the form of matrices of features of the model space, with the available images-standards, specified in the form of matrices of features of the same model space. That is, each matrix of the recognizable object, representing the image of this object on the plane of the model space, must be compared with a similar matrix of the reference object. In this case, the elements of each matrix of the recognizable linguistically specified object are compared with the values of the corresponding elements of each matrix of the reference object, and the number of matches is summed up. As a result of the comparison, the number of points of the semantic image of the recognizable object will be determined, which coincide with the points of the reference image.

To determine the class to which the studied linguistically specified object belongs, we introduce the set of parameters *result* (4), which shows the percentage of coincidence of the analysed text with the available reference texts (5).

$$result = \{result_1, \cdots, result_W\} \quad (4)$$

where W – the total number of semantic images,

$$result_g = \frac{S_g}{M_0} \cdot 100\% \quad (5)$$

where g = 1..W,

Sg – the number of points of the recognizable pattern coinciding with the points of the g-th reference pattern,

M_0 – the number of points in the recognizable pattern.

Thus, in this case, the maximum value of $result_g$ means the best match of the *g*-th reference description of an object that characterizes a specific class of objects (domain) with a recognizable publication. As a result, we can conclude that the publication in question belongs to a specific class of the domain.

As a result of this stage, a labeled corpus of semantic images will be formed, which can be used for training and testing a neural network. Let us assume:

- 95% of the corpus of semantic images will be used as a training data for the neural network;
- 5% — as a testing data.

At the next stage, it is necessary to train the neural network selected for solving the problem of classification of scientific publications [18] – [21] based on the prepared training dataset of semantic images of publications.

Then it is necessary to test the trained network on the available testing dataset. After the successful completion of this stage, the neural network is ready to recognize new publications of specific domain.

This is the essence of the method proposed by the authors for the recognition of semantic images of scientific publications based on neural networks. Using the described method, it is possible to determine the belonging of a linguistically given object (text of a scientific publication) to a specific class of objects (domain).

On the basis of this method, the authors have developed an algorithm for determining the class of the object under study, which can be used in information retrieval systems to determine the belonging of a linguistically given object (scientific publication) to one of the selected classes of objects (domain). This algorithm can be used when designing a search engine for an information retrieval system.

V. Organization of Search Engine of a New Type

The standard search engine does not consider the semantic content of natural language texts. Such text, from the point of view of the search engine, is simply strings of characters separated by spaces. It is such “words” that are preselected from the text and entered into the search index, which allows the search engine to find documents. At the same time, the query language of a good search service permits to set various restrictions on the desired combinations of words in the document, which allows, in principle, to formulate very complex queries, describing the desired meaning in the text.

However, the problem of creating good information retrieval systems based on standard search engines is that the user wants to formulate his request in the form of a simple set of words or phrases in a natural language, expecting the machine to understand these words that can be written in the text. In such a situation, if the words of the query do not match the words of the search text, it will be almost impossible to find the required document.

Therefore, on the basis of the approaches proposed by the authors of the article to the formation and recognition of semantic images of scientific publications, it is possible to design an information retrieval system of a new type. It will search not by keywords, but by coincidence of

semantic situations. Such an information retrieval system of a new type will form semantic images of documents and store them in its database (or storage). The information retrieval system should build these images using the method of spatial representation of text descriptions. And the search engine should be built on the basis of using the method of recognition of semantic image of a text description. It will compare the semantic image of the user's search query, built on the basis of the method of spatial representation of text descriptions with the semantic images available in the database of the information retrieval system of text-based documents. Then, based on a predefined successful search criterion, a list of documents relevant to the query (if any were found) is displayed. To reduce the percentage of losses when translating original sentences in natural language into the formalized representation proposed by the authors (semantic image of a text description), we formulate the requirements for the structure of queries:

- they should be written in the form of simple sentences;
- pronouns must be absent (or replaced by their corresponding concepts).

Search engines of a new type, which use the method of recognition of the semantic image of a text description, can also be used on the Internet, because they implement a semantic search for documents, which can improve the relevance of the search. But at the same time, preliminary work should be carried out to create dictionaries of descriptors and predicates for the corresponding subject areas. Semantic images of information resources of the Internet are formed on the basis of the method of spatial representation of text descriptions. These semantic images can be used to index documents on the Internet.

VI. CONCLUSIONS

The approach to the construction of semantic images of scientific publications, proposed in the article, makes it possible to theoretically substantiate the fundamental possibility of the existence of an algorithmic solution for problems of comparison, classification of these images. This approach to formalizing text for its further processing by a neural network initially has a semantic focus and potentially improves the quality of training a neural network. The algorithm developed by the authors for determining the class of a scientific publication based on neural networks can effectively solve the problem of determining whether texts of scientific content belong to one of the specified classes. This makes it possible to use the proposed approaches and algorithms in the design of new types of search engines that will carry out semantic search for documents.

REFERENCES

[1] X. Zhang, J. Zhao, Ya. LeCun, “Character-level convolutional networks for text classification,” Proc. Neural Inform. Processing Systems Conf. (NIPS 2015), Montreal, Canada, 2015. [Online]. Available: https://arxiv.org/abs/1509.01626 [Accessed: May 20, 2020].

[2] R. Ju, “An Efficient Method for Document Categorization Based on Word2vec and Latent Semantic Analysis,” 2015 IEEE Int. Conf. on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Auto-nomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK, 2015, pp. 2276-2283.

[3] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, “SemEval-2014 Task 4: Aspect based sentiment analysis,” Proc. 8th Int. Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 2014, pp. 27–35.

[4] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532-1543.

[5] D. A. Pospelov, Logical-linguistic models in management. Moscow: Energoizdat, 1981. (in Russian)

[6] V. Sh. Rubashkin, Representation and analysis of meaning in intelligent information systems. Moscow: Nauka, 1989. (in Russian)

[7] V. K. Finn, “Information systems and problems of their intellectualization,” NTI, Ser. 1, No. 1, pp. 1-14, 1981. (in Russian)

[8] I. A. Mel'chuk, Experience in the theory of linguistic models "meaning-text". Moscow: Nauka, 1982. (in Russian)

[9] M. Minsky, Frames for knowledge representation. Moscow: Energiya, 1979. (in Russian)

[10] Iu. V. Bruttan, “Intellectualization of the behavior of computers based on the use of a cellular automaton of a new type,” Nauchno-tekhnicheskie vedomosti SPbSPU, No. 2, pp. 225-229, 2007. (in Russian)

[11] Iu. V. Bruttan, “Linguistic processor for processing scientific knowledge,” Energy - XXI century, No. 4 (104), pp. 82-85, 2018. (in Russian)

[12] Iu. V. Bruttan, Methods for spatial representation and analysis of text descriptions for information retrieval systems. Monograph. Pskov: Pskov State University, 2016. (in Russian)

[13] R. Gonzalez, Digital image processing. Moscow: Technosphere, 2005. (in Russian)

[14] Ya. A. Fomin, Pattern Recognition: Theory and Applications. 2nd ed. Moscow: FAZIS, 2012. (in Russian)

[15] V. V. Alexandrov, Automated information processing in the predicate language. Moscow: Nauka, 1982. (in Russian)

[16] I. Antonov, Iu. Bruttan, L. Motaylenko, and D. Andreev, “The Method of Automated Building of Domain Ontology,” in proceedings of the 12th International Scientific and Practical Conference on Environment. Technology. Resources, Rezekne, 2019, vol. II, pp. 34-37.

[17] D. A. Andreev and M. V. Voronov, “Method for constructing an ontology of technological actions,” Bulletin of Saratov State Technical University, No. 3 (67), pp. 160–168, 2012. (in Russian)

[18] B. Benjamin, B. Rebecca, and O. Tony, Applied analysis of text data in Python. Machine Learning and Building Natural Language Processing Applications. Saint-Petersburg: Peter, 2019. (in Russian)

[19] Y. Goldberg, “A primer on neural network models for natural language processing,” Journal of Artificial Intelligence Research, vol. 57, pp. 345–420, 2016.

[20] A. Tor, R. A. Shirvani, Y. Keneshloo, N. Tavvaf, and E. A. Fox, “Natural language processing advancements by deep learning: A survey,” 2020, arXiv:2003.01200. [Online]. Available: http://arxiv.org/abs/2003.01200 [Accessed: May 24, 2020].

[21] M. Ghiassi, M. Olschimke, B. Moon, and P. Arnaudo, “Automated text classification using a dynamic artificial neural network model,” Expert Systems with Applications, No. 39, pp. 10967–10976, 2012.