DOI https://doi.org/10.30525/978-9934-26-303-3-1

FEATURES OF DATA COLLECTION AND PROCESSING FOR NEURAL NETWORK MODELING OF LANGUAGE CATEGORIES

Dovhan O. V.

INTRODUCTION

The term "technology" (from the Greek *techne*) means skill, ability, etc. At the same time, this term carries a certain procedural nature, as it can mean both "a set of knowledge, information about the sequence of individual production operations in the process of producing something" and "a set of methods of processing or processing materials, manufacturing products, conducting various production operations"[1]. Therefore, the term can be used to refer to a certain set of data, as well as the ways in which they are used, collected and processed.

In the context of scientific activity, which is differentiated, relevant and up-to-date, it is advisable to talk about both possible interpretations of the above concept. After all, scientific activity is a tool for cognition of ontological reality through gnosiological (from the Greek *γνώσις* – "knowledge"), i. e. cognitive, which, in turn, is aimed at studying episteme (from the Greek *έπιστήμη* – "knowledge"), i. e. information or data in the ontological (from Latin *ontologia* and ancient Greek *ών* – being or existing) plane. From the point of view of linguistics, this means the research (epistemology) of the categories of the language polysystem (epistemology) in the context of linguistic pragmatics (ontology)[2].

Today, the tendency to develop, deepen, and actualize interdisciplinarity has transformed the conventional forms of scientific activity. The aforementioned modification is not related to the transformation of scientific practices, the integration of the mathematical paradigm into humanities research, or anything else (although all of this is happening), but rather to the change in the flow of data. This process refers to the arithmetic growth of the volume of data flow, as well as its rapid differentiation (stratification) and the constant inclusion of more and more new types (log files, email, social networks, etc.).

Since data management is the core of scientific activity, the above causes a number of problems that impede the conduct of high-quality and thorough

[1] Технологія – Технологія. *Горох* : вебсайт. URL: https://cutt.ly/44Cqnyp (дата звернення: 31.03.2023).

[2] Філософський енциклопедичний словник : енциклопедія / НАН України, Інститут філософії ім. Г. С. Сковороди ; головний редактор В. І. Шинкарук. Київ : Абрис, 2002. 742 с.

research without the use of new tools. This is due to the fact that the aforementioned tools limit the process of processing incoming data to their capacity. The point is that the tendency to increase the amount and types of data that are constantly collected, stored and transmitted by information technology is changing the priorities of modern science, economics and culture[3].

In order to avoid confusion in the context of our research, we should position science itself as a certain set of data on the sequence of certain actions to achieve the desired result (ideal nature). Instead, scientific activity itself should be presented as the embodiment of its concept, which has a secondary (non-ideal) nature, which is not due to the quality or breadth of coverage, but to the primacy of localization in the temporal (temporal) field. We are talking about the epistemological basis of scientific activity – data and work with them: generation, analysis, structuring, visualization, etc. It is clear that data can be produced, interpreted, and accumulated, but scientists first process them in a certain way, creating something new from them[4].

At the same time, it is necessary to take into account such trends in the use of modern data as their intermedial, intertextual, interauditory nature, etc. In turn, these features create the need to develop a fundamentally new approach to data collection and processing – a new scientific toolkit that would be universal for all fields of knowledge. We are talking, first of all, about tools adapted for analyzing various interpretations, transformations and other of sense, which are essential for working with language categories[5].

In our opinion, the best option in this case is to use Data Science methods, a science that combines Big Data analysis, elements of mathematical statistics, rhetoric, etc., as well as a special approach to analyzing and building correlations between different types of data. This versatility of its tools allows for a thorough analysis based on data in the fullness of their interrelationships with the environment (discourse). The objects of Data Science analysis are data of various kinds: relational (ordered, structured, or tabular) and non-relational (multimedia data of various kinds and types).

---

[3] Джус С. І. Потреба використання DATA SCIENCE & BIG DATA ANALYSIS (Наука про дані та аналіз великих даних) у сучасному статистичному та фінансовому світі. *Бізнес-аналітика в управлінні зовнішньоекономічною діяльністю* : матеріали IV Міжнар. наук.-практ. конф. Київ, 2017. С. 53.

[4] Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification / Banerjee I. et al. *Artificial intelligence in medicine*. 2019. № 97. P. 79–88. DOI: https://doi.org/10.1016/j.artmed.2018.11.004.

[5] Sethia D., Singh P., Mohapatra B. Gesture Recognition for American Sign Language Using Pytorch and Convolutional Neural Network. In: *Intelligent Systems and Applications: Select Proceedings of ICISA 2022*. Singapore : Springer Nature Singapore, 2023. P. 307–317. DOI: https://doi.org/10.1007/978-981-19-6581-4_24.

The main thing is that the aforementioned science is in line with the process of integrating linguistics into the mathematical paradigm through the use of multi-tools for processing data of any nature[6]. We are not talking about a specific field of science in general and linguistics in particular. Undoubtedly, modeling language categories will be especially useful for computer and mathematical linguistics. However, the point here is to approach the language polysystem from the standpoint of algorithmization of language units, categories, etc., in particular, the peculiarities of data collection and processing for this purpose.

Naturally, the issues of creating, collecting, structuring and processing (primarily analytical and synthetic) data are relevant due to the aforementioned increase in their number[7]. Data are being accumulated, classified, analyzed, processed, etc., but all these processes (from searching to finding correlations in them) are becoming more and more expensive. That is why the issue of using appropriate tools for scientific activities is becoming more and more acute, as they enable the reuse of updated data. We are talking about cases when already known data have an alternative interpretation in a different context.

Thus, by using new tools that are focused on working with different types of data and adapted to track correlations between them, we get more complete and thorough research results. In the context of data collection and processing, such tools are productive because they will be adapted to work with this type of information, taking into account the fullness of the latter's existence and the network of cultural and linguistic practices associated with it, which is the relevance of our research.

1. Data collection and processing for neural network modeling

The problem of collecting and processing data to build neural network models of language categories is important and relevant for scientific linguistic research, machine learning, Data Science, and Big Data analysis. First of all, this is due to the universal nature of data processing tools (in particular, language data), since methods and models for such operations have cross-cutting potential[8]. We are talking about their use in translation studies (in particular, machine translation), natural language analysis, etc., as well as in working with data sets in general.

[6] Cielen D., Meysman A. D. B., Ali M. Introducing Data Science. Big Data, Machine Learning, and more, using Python Tools. New York, 2016. 322 p. P. 1–2.

[7] Heidari M., Rafatirad, S. Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews. In: *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization SMA*. IEEE, 2020. P. 1–6. DOI: https://doi.org/10.1109/SMAP49528.2020.9248443.

[8] Pater J. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language 95.1* 2019. P. 41–74. DOI: https://doi.org/10.1353/lan.2019.0009.

Data Science is the art of uncovering ideas and trends that are hidden behind a certain set of data, and in this sense, it is an information technology. That is, a process that uses a set of tools and methods for collecting, processing and transmitting data (primary information) to obtain new quality information about the state of an object, process or phenomenon (information product). Data Science is an extension of statistics that is able to cope with the huge volumes of data produced today[9].

At the same time, Data Science tools are virtually unlimited: clustering, neural networks, boosting, trees, Natural Language Processing, etc. The advantage of new methods for research in any subject area is that a well-constructed system based on well-chosen Data Science tools allows for large-scale research with a minimum of hardware resources.

This is due to the fact that most of the tools work on the GitHub principle: all data, tools, and the experiment itself are in the cloud and can be used locally on the desktop version if desired. A good example of this is Watson Studio, Jupiter Notebook, Apache, Scala, etc. Naturally, collecting and processing data to build neural network models is only one of the options for implementing the process of analyzing and processing the occurrence of language categories, in particular, natural language. At the same time, neural network modeling is a universal tool for scientific research, as it has a multi-subject nature of its use (in particular, in linguistics, it can be the creation, classification, recognition, processing, etc. of data)[10].

Nevertheless, it is more productive than others due to the integration of its work: by using neural network modeling in the spirit of the Data Science methodology, i. e., approaching the process as analysts rather than linguists, researchers obtain results that can be integrated into multi-subject grant research, which is a definite plus. However, the main problem remains the rather complicated process of preparing for the collection and processing of data for neural network modeling. The core difficulty in this process is the design of approaches and methods for collecting and processing data: we are talking about the selection, discovery, structuring, and phasing of a large amount of data to be used for machine learning.

It is clear that the selection of data sets in this case depends on the linguistic categories that are the object of analysis of a particular linguistic research. Thus, when studying the stylistics of certain data sets, it is worth using a variety of data sources (text corpora, social networks, Internet resources: websites, web portals, etc.). In addition to the resources for training, as

[9] Cielen D., Meysman A. D. B., Ali M. Introducing Data Science. Big Data, Machine Learning, and more, using Python Tools. New York, 2016. P. 1.

[10] Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification / Banerjee I. et al. *Artificial intelligence in medicine*. 2019. № 97. P. 79–88. DOI: https://doi.org/10.1016/j.artmed.2018.11.004.

mentioned above, it is necessary to design the stages of updating certain sources, because in case of incorrectly built machine learning, the researcher will have to retrain, which means a loss of time and reduced efficiency.

As you know, neural network training takes place in layers, and completion of each layer means acquiring certain skills in working with these neural network models. To move from training on one layer to training on another, you need to think about a logical connection, and the information on each layer should be characterized by continuity and consistency, otherwise the neural network will not successfully assimilate the information[11]. Particular attention should also be paid to the specifics of the neural network model's work with data, namely, approaches to working with it to improve the parsing process. For example, it is productive to use technologies such as data pruning and preprocessing to make the data more representative.

In addition, it is advisable to improve the adaptability of the neural network model to the linguistic design of the data, in particular, for derivatological research, it is advisable to take into account their features by updating lemmatization (turning a word into a dictionary form or lemma, in other words, transforming the form of words into the original one), stemming (reducing words to the base by dropping the ending or suffix), etc.[12].

In our opinion, the key to preparing for the collection and processing of data for neural network modeling of language categories is the localization of characteristics that reflect the desired linguistic features. Thus, it is advisable to use word-formation, stylistic, morphological, syntactic, semantic, lexical, etc. features that will help the neural network model focus on the necessary language units or levels, thereby improving the accuracy and efficiency of its work.

As mentioned above, the Data Scince toolkit, which includes neural network modeling, characterizes inclusion in the discourse, which, in turn, necessitates taking into account the specifics of the linguistic and cultural polysystems of a particular ethnic group. For example, it is advisable to take into account the peculiarities of semantics, temporal representation, the originality of syntax and phraseology, etc. of a language that will affect the process of neural network modeling. In particular, it is worth predicting the

[11] Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques / Anand M. et al. *Theoretical Computer Science*. 2023. № 943. P. 203–218. DOI: https://doi.org/10.1016/j.tcs.2022.06.020.

[12] Jayasudha J., Thilagu M. A Survey on Sentimental Analysis of Student Reviews Using Natural Language Processing (NLP) and Text Mining. In: *Innovations in Intelligent Computing and Communication: First International Conference, ICIICC 2022, Bhubaneswar, Odisha, India, December 16–17, 2022, Proceedings*. Cham : Springer International Publishing, 2023. P. 365–378. DOI: https:// doi.org/10.1007/978-3-031-23233-6_2.

peculiarities of realities and gaps in the data sets of the language being studied[13].

In general, the preparation of data collection and processing for neural network modeling of language categories is an important milestone in the context of studying the phenomenon of the language polysystem and the evolution of both methods and tools for such research. At the same time, the aforementioned research is significant from the standpoint of the powerful development of machine learning and automatic data processing for any field of knowledge, including linguistics. Thus, neural network modeling is a significant step in the context of scientific activity, as it affects entire research areas: from machine translation to speech recognition, as well as data mining of any nature.

In terms of linguistics, the ability to automatically classify text arrays by topic, genre, type, etc. is significant[14]. Neural networks with the Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) architecture, which can distinguish the above characteristics, are productive for such work. The use of such tools will provide the ability to initially screen textual data coming from various sources (social networks, the Internet, etc.) to speed up the processing process[15].

The use of neural-frontier models for automatic text translation (e. g., Google Translate, DeepL Translate, which are used in online services or can be installed as a browser extension) is illustrative. Such neural network models can automatically translate texts, updating the knowledge base of semantics, grammar, morphology, etc. of the language polysystems (the language from which they are translated and the language they are translated into) to provide more accurate and complete translation.

Thus, studying the peculiarities of data collection and processing for neural network modeling of language categories is an important step in line with the trend toward the integration of mathematical and humanities sciences, which allows for high-quality and thorough interdisciplinary research.

As for the actualization of neural network models for modeling language categories, the use of tools in the vein of Data Science (with an emphasis on measurable methods and techniques) allows to reduce the time for conducting research in general, improve the quality and representativeness of data

[13] Scientometric analysis and classification of research using convolutional neural networks: A case study in data science and analytics / Daradkeh M. et al. *Electronics*. 2022. № 11 (13). DOI: https://doi.org/10.3390/electronics11132066.

[14] Sentiment strength detection with a context-dependent lexicon-based convolutional neural network / Huang M. et al. *Information Sciences*. 2020. № 520. P. 389–399.

[15] Sentiment analysis on twitter data by using convolutional neural network (CNN) and long short term memory (LSTM) / Gandhi Usha Devi et al. *Wireless Personal Communications*. 2021. № 1/10. DOI: https://doi.org/10.1007/s11277-021-08580-3.

processing, and initiate a number of international projects on this basis[16]. The results of research conducted with the use of such tools can be updated to improve the performance of a number of programs and services related to machine learning, automatic translation, and others that deal with algorithmization of levels, categories, and units of the language polysystem.

2. Neural network modeling of language categories in the context of language data

Neural network modeling of language categories is a promising area of research in modern linguistics. It has been noted above that the effectiveness of implementing neural network models in linguistic research lies in the ability to cover a large amount of data (Big Data), to find a network of regularities in it, implicit and explicit correlations between the analyzed linguistic material, etc.[17].

It is natural that neural network models are productive for classifying texts according to various criteria (morphological, semantic, stylistic, etc.). In this case, the "starting points" are the linguistic categories of sense and absurdity as elements of a binary code (1 and 0), since here they act as a boundary for certain properties under analysis.

Collecting and processing data to create neural network models is a complex process, as it limits the results of the neural network in general and their completeness, correctness, appropriateness, etc. in particular. Thus, the quality of the materials used for training, the correctness of the task formulation, and the relevance of the materials analyzed by neural network models to this task are the basic milestones of its successful operation.

This is due to the fact that it is necessary to present a representative sample of a text or other dataset for the successful training of neural network models, which would allow them to efficiently classify texts into the desired categories. At the same time, training on large corpora of texts or other data has its own peculiarities: for example, these datasets may contain duplicate information or information that is not representative of a particular language category, and there is a possibility of errors in texts collected from different sources[18].

[16] Transfer learning based recurrent neural network algorithm for linguistic analysis / Sasikala S. et al. *Concurrency and Computation: Practice and Experience 34.5*. 2022. DOI: https://doi.org/10.1002/cpe.6708.

[17] Rahman S., Chakraborty, P. Bangla document classification using deep recurrent neural network with BiLSTM. In: *Proceedings of International Conference on Machine Intelligence and Data Science Applications*: MIDAS 2020. Singapore: Springer Singapore, 2021. P. 507–519. DOI: https://doi.org/10.1007/978-981-33-4087-9_43.

[18] Tanantong T., Yongwattana P. A convolutional neural network framework for classifying inappropriate online video contents. *IAES International Journal of Artificial Intelligence*. 2023. Vol. 12, Iss. 1. P. 124–136. DOI: https://doi.org/10.11591/ijai.v12.i1.

It is clear that depending on the purpose and needs of the researcher (in our case, a linguist), approaches to training neural networks are chosen. For example, a neural network model can be trained on the basis of large corpora of texts (the aforementioned Big Data) from open sources or own datasets created through crowdsourcing.

In our case, a neural network model involved in the research of the basic interdisciplinary problem of sense is productive, which produces the use of the entire range of Data Science tools (we are talking, first of all, about the use of Machine Learning, Neural Networks, natural language research using frameworks and applied analysis of textual data in Python)[19] in the research, whose training is based on the principle of plagiarism detection programs as the main strategy for working with text. It is about finding arrays of texts that are plagiarized and using them as training data for a neural network. In this case, it is advisable to update various techniques of text processing and data vectorization to ensure optimal conditions for its functioning and training.

When it comes to data for neural network modeling of language categories, annotation is an important component of their performance. This process can be implemented using a variety of methods, one of which is manual (classical) annotation or analytical and synthetic text processing. Another option is automatic annotation, which is performed through the use of special software tools[20].

The first option involves the manual selection and extraction of language categories in data sets; this process is usually characterized by high quality output, but it is costly and time-consuming. That's why an automated option is preferable when working with Big Data. First of all, this is due to the fact that automatic annotation is performed with the help of various software tools, among which the most productive are partial language and dependent analyzers.

These tools use language grammars and formal rules for the existence and localization of language categories, which leads to a fairly high speed of the process. Despite the high speed and efficiency of the above methods of automatic annotation in the process of working with Big Data, they have their limitations. For example, the accuracy of automatic annotation has certain limitations when it comes to ambiguous or difficult-to-understand language categories, the use of which has no unambiguous interpretation.

It should be noted that the problem of working with large corpora of texts, i. e. Big Data, for training a neural network is the actual collection and processing of such arrays for its training. This is not about annotating data by

[19] Cielen D., Meysman A. D. B., Ali M. Introducing Data Science. Big Data, Machine Learning, and more, using Python Tools. New York, 2016. P. 4.

[20] Дранишников Л. В. Нечітке і нейромережеве моделювання в системах управління. *Міжнародний науковий журнал «Грааль науки».* 2021. № 5. С. 153–159.

linguists, archivists, or document specialists, but about the relevance of the material it analyzes, its relevance for the analysis of a particular language category.

Another option for automatic data collection and processing for neural network modeling of language categories is to use Big Data as a kind of database of materials without preliminary processing. We are talking about the possibility of training a neural network on the frequency of the compatibility or combinability of certain lexical units in the context of a certain semantic environment. In this case, the context of a particular meaning and language practice will be selected based on the frequency of certain "beacons" that are located nearby[21].

For example, the neural network will choose the variant of the word "braid" as a woman's hairstyle if it localizes the key concepts "hairstyle", "hair", "long", etc. In fact, this option is a research of the use of language categories in the context of discourse, that is, the context in which they are actualized. Thus, it is possible to track the change in the meaning of certain language categories depending on the environment, which will help improve the degree of understanding of the language polysystem and the features of its recognition by the neural network model.

Typically, text processing is based on N-gram models, which calculate the occurrence of a particular element based on its frequency in the text. The probability of occurrence is estimated after calculating the intensity of repetition of the corresponding n-grams. This processing includes all levels of the language polysystem: phonological, morphological, lexical, syntactic, semantic, etc. For example, the semantic and lexical levels were the basis for the work of a number of research institutions in the United States and Japan. Researchers from these countries have developed a method for detecting fake news by measuring its thematic differentiation. According to the researchers, fake news is characterized by less thematic diversity of messages.

The researchers' method of detecting fake news is to calculate the diversity of topics using "micro-clustering". Micro-clustering collects data into small homogeneous groups and creates a set of topics, each of which is composed of one or more clusters. Micro-clusters were extracted from these clusters using data polishing, after which the researchers analyzed the changes in themes within the clusters. This process included identifying the degree of topic diversity by taking into account the number of clusters and words in one cluster[22].

---

[21] Rahman S., Chakraborty, P. Bangla document classification using deep recurrent neural network with BiLSTM. In: *Proceedings of International Conference on Machine Intelligence and Data Science Applications*: MIDAS 2020. Singapore: Springer Singapore, 2021. P. 507–519. DOI: https://doi.org/10.1007/978-981-33-4087-9_43.

[22] Терещенко Л., Гладьо С. Неправдивість письмового тексту: шляхи її визначення. *Psycholinguistics.* 2022. № 31 (2). С. 116–136. URL: https://cutt.ly/Q42Rfwg (дата звернення: 31.03.2023).

Another option for collecting and processing data for neural network modeling is to create data that is representative of a particular problem. For example, to analyze the problems of linguistic morphology, it is advisable to update linguistic corpora, each word of which has an annotation of its morphemic structure. In this case, the use of such data will allow us to achieve thorough and representative results in the context of building accurate and efficient morphological models.

It is clear that the collection and processing of data for neural network modeling of language categories is determined by the type and kind of neural network model. For example, in the case of active learning, in which the neural network model is trained on a small amount of data and then immediately used to classify data sets, it is necessary to prepare a fixed amount of data that does not allow for different interpretations.

Such unambiguity is extremely important in this case, since the main emphasis here is on a special algorithm that helps the neural network model identify the most significant texts that require manual analysis by the researcher. In turn, a resolution from a scientist who will conduct the above analysis and provide feedback in the form of a specific resolution will improve the performance of the neural network. At the same time, the researcher should understand that such a correction will not replace a properly built structure and training plan for a neural network model, so it is advisable to focus on data collection and processing[23].

A peculiar interpretation of the previous option is productive for data collection and processing – updating the methods of training a neural network model with and without a teacher. In the first case, we are talking about using labeled text data, which allows the neural network model to track correlations between input and output labels. For example, to classify text data by topic, it is advisable to use labeled data, which will allow the neural network to quickly learn to classify it by this parameter. This method requires relatively less data collection and processing by the researcher, as most of the actions for this purpose do not take much time. In the second case, the process is similar, but there is no labeling, instead, the neural network model independently builds the necessary connections and dependencies between the data, features of their structure, etc. The second case is productive for automatic clustering of texts by similarity.

The variability of data collection and processing methods for neural network modeling of language categories is also interesting. We are talking, first of all, about the vector representation of words (word embedding), syntactic and semantic analysis of text data, etc. Remember that the choice of

[23] Shaji B., Singh R., Nisha K. L. High-performance fuzzy optimized deep convolutional neural network model for big data classification based on the social internet of things. *The Journal of Supercomputing*. 2023. DOI: https://doi.org/10.1007/s11227-022-04974-7.

a particular approach to data collection and processing should correlate with the object, purpose, hypothesis, etc. of our linguistic research. In this context, the variety of tasks and evaluation metrics is important: for example, text classification tasks require fine-tuning (accuracy, sensitivity, specificity, etc.), while generation tasks only require representation of the quality of the above process (for example, BLEU and ROUGE)[24].

It is worth noting that data collection and processing for neural network modeling do not always precede the construction of a neural network model. This is due to the fact that some training methods update the data processing methods available in the neural network model by default: For example, information about its latent state, which allows for automatic text analysis and contextualization in the final decision-making process.

CONCLUSIONS

Thus, the research of the peculiarities of data collection and processing for neural network modeling of language categories is an important and urgent problem of linguistic science, and the research of possible approaches to data collection and processing in the context of Data Science methods and tools is in line with current trends towards building an integrated research with a practical component.

At the same time, the creation, training, and maintenance of neural network models for any purpose is an extremely complex and multi-stage process. Starting from the planning of the research itself (its object, subject, hypothesis, materials, etc.), continuing with the peculiarities of data collection and processing for this work, in particular, a neural network model as a tool for implementing the plan, ending with the structure and type of neural network, as well as the construction of its training, its stages, and materials for it.

Each component of the aforementioned process directly affects the data sought to deploy the neural network model, limiting the results of its work. That is why the choice of approach to data collection and processing, the quality of the materials used (prepared, created, presented, etc.) are the key factors that affect the results, completeness, and quality of a scientific research.

Therefore, continuing research into the peculiarities of neural network models in the context of linguistics will ensure a rapid increase in their efficiency and accuracy in the entirety of linguistic issues. In addition, such work will produce a deeper understanding of the language polysystem as

[24] Jayasudha J., Thilagu M. A Survey on Sentimental Analysis of Student Reviews Using Natural Language Processing (NLP) and Text Mining. In: *Innovations in Intelligent Computing and Communication: First International Conference, ICIICC 2022, Bhubaneswar, Odisha, India, December 16–17, 2022, Proceedings*. Cham : Springer International Publishing, 2023. P. 365–378. DOI: https:// doi.org/10.1007/978-3-031-23233-6_27.

a phenomenon of ontological reality, a more thorough understanding of the peculiarities of automatic textual data processing and the specifics and prospects of working with them.

At the same time, linguistic research in this area will allow us to develop optimal scenarios for combining different types and kinds of data collection and processing methods, as well as to develop an optimal sequence of working with them to achieve better results of neural network models. In our opinion, it is promising to combine conventional (traditional) methods and types of work with text data with neural network modeling, mathematical statistics, etc. This approach will provide a better understanding of the peculiarities of language categories and improve the quality of work with them.

It is also productive to use neural network models in the context of solving specific linguistic tasks related to translation studies, language analytics, automatic text processing, etc. In addition to making a significant contribution to the development of linguistics, such research will contribute to the evolution of information technology and ensure the efficiency and completeness of its actualization in various spheres of life.

In this research, we have considered various options for collecting and processing data for neural network modeling of language categories: one of the most effective ways is to use large corpora of texts. In addition, it is also productive to create your own textual data for training neural network models: for example, to build neural network models that are highly accurate and efficient, it is advisable to use specially designed data (linguistic corpora). Another option is active and deep learning, which involves less preparatory work but requires constant response from the researcher, which ensures greater accuracy of the results.

Thus, data collection and processing is an important stage of research using neural network modeling. The choice of the type of neural network, architecture of the neural network model and its parameters, data processing methods for it, peculiarities of data collection, etc. is important because it affects the research results.

That is why the above-mentioned decisions should be determined by the specific research task, available resources, peculiarities of the language polysystem under research, etc. Thus, for complex linguistic categories that do not have a single interpretation, such as the peculiarities of semantic relations between linguistic units, it is advisable to use large and diverse data corpora in terms of nature, structure, content, etc. At the same time, for more obvious and unambiguous tasks, it is worth using specially generated or written data, which will affect the project budget and the timing of its implementation.

For successful neural network modeling of language categories, it is advisable to use a variety of methods: tokenization, vectorization, etc. It is

necessary to determine the features of the optimal architecture and hyperparameters that will ensure high efficiency and accuracy in the modeling process. Pre-training is also productive, when a neural network model is trained on large corpora before its potential is actualized for a specific linguistic research task. This approach provides greater flexibility and versatility of the neural network model used, which directly affects the accuracy and efficiency of its results.

Perspective

Collecting and processing data for neural network modeling of any category is also important in the context of a predictive comparison of the effectiveness of different methods. We are talking about comparing the performance of neural network models with other data processing options: in our opinion, it is productive to compare them with mathematical statistics in general and statistical methods in particular (the latter can be extremely useful in analyzing text arrays, finding trends and paradigms in them).

It is also productive to compare the architectures of different types of neural networks and track the correlations between their structure, functionality, and the results that can be obtained with their help. It is also possible to compare neural network architectures and hyperparameters to find the most optimal model for a particular linguistic research or a number of models for analyzing each specific problem with a particular type of neural network.

In the context of the above-mentioned issues, it is relevant to research the influence of various factors on the process of analyzing a neural network model of language categories. Thus, it is advisable to research the correlation between the amount of input data and the efficiency of the neural network model (speed, accuracy, errors, etc.), the impact of data collection and processing on the result of such a model, the relationship between reducing the dimensionality of feature vectors and the completeness of their analysis, etc.

The prospect of our research is also to research the correlations between linguistic categories and the peculiarities of their occurrence in the linguistic polysystem. For example, it is productive to research the correlations between the phonetic and phonological form of a unit of a language polysystem and its semantic shades, between the peculiarities of its syntactic representation and the breadth of meanings in language practice.

Thus, the research of the peculiarities of data collection and processing for neural network modeling of language categories is an important area of linguistic research in general and linguistic analytics in particular. The specifics of the analysis of efficient operation and parametric data representation are promising in view of the possibility of improving the quality of neural network models in the context of their actualization of research on different levels of the language polysystem, in particular, the analysis of language categories.

SUMMARY

The article analyzes the features of data collection and processing for neural network modeling of language categories. Attention is focused on different methods of data collection and processing from the perspective of mathematical statistics, Data Science, mathematical and computer linguistics. The author outlines the correlation between the objectives of the research and the methods and approaches to data processing, and identifies productive models of data collection and processing. The importance of studying the peculiarities of data collection and processing for linguistic science is emphasized. The specifics of parametric data representation as a direction of actualization of research on different levels of the language polysystem are presented. The influence of various factors on the process of analyzing the neural network model of language categories is analyzed. The prospects for studying the correlation between the amount of input data and the efficiency of the neural network model (speed, accuracy, errors, etc.), the impact of data collection and processing on the result of such a model, the relationship between reducing the dimensionality of feature vectors and the completeness of their analysis, etc. are highlighted.

Bibliography

1. Cielen D., Meysman A. D. B., Ali M. Introducing Data Science. Big Data, Machine Learning, and more, using Python Tools. New York, 2016. 322 p.

2. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification / Banerjee I. et al. *Artificial intelligence in medicine*. 2019. № 97. P. 79–88. DOI: https://doi.org/10.1016/j.artmed.2018.11.004

3. Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques / Anand M. et al. *Theoretical Computer Science*. 2023. № 943. P. 203–218. DOI: https://doi.org/10.1016/j.tcs.2022.06.020

4. Heidari M., Rafatirad, S. Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews. In: *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization SMA*. IEEE, 2020. P. 1–6. DOI: https://doi.org/10.1109/SMAP49528.2020.9248443

5. Jayasudha J., Thilagu M. A Survey on Sentimental Analysis of Student Reviews Using Natural Language Processing (NLP) and Text Mining. In: *Innovations in Intelligent Computing and Communication: First International Conference, ICIICC 2022, Bhubaneswar, Odisha, India, December*

16–17, 2022, Proceedings. Cham: Springer International Publishing, 2023. P. 365–378. DOI: https:// doi.org/10.1007/978-3-031-23233-6_27

6. Pater J. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language 95.1* 2019. P. 41–74. DOI: https://doi.org/ 10.1353/lan.2019.0009

7. Rahman S., Chakraborty, P. Bangla document classification using deep recurrent neural network with BiLSTM. In: *Proceedings of International Conference on Machine Intelligence and Data Science Applications*: MIDAS 2020. Singapore: Springer Singapore, 2021. P. 507–519. DOI: https://doi.org/ 10.1007/978-981-33-4087-9_43

8. Scientometric analysis and classification of research using convolutional neural networks: A case study in data science and analytics / Daradkeh M. et al. *Electronics*. 2022. № 11 (13). DOI: https://doi.org/ 10.3390/electronics11132066

9. Sentiment analysis on twitter data by using convolutional neural network (CNN) and long short term memory (LSTM) / Gandhi Usha Devi et al. *Wireless Personal Communications*. 2021. № 1/10. DOI: https://doi.org/ 10.1007/s11277-021-08580-3

10. Sentiment strength detection with a context-dependent lexicon-based convolutional neural network / Huang M. et al. *Information Sciences*. 2020. № 520. P. 389–399.

11. Sethia D., Singh P., Mohapatra B. Gesture Recognition for American Sign Language Using Pytorch and Convolutional Neural Network. In: *Intelligent Systems and Applications: Select Proceedings of ICISA 2022*. Singapore: Springer Nature Singapore, 2023. P. 307–317. DOI: https:// doi.org/10.1007/978-981-19-6581-4_24

12. Shaji B., Singh R., Nisha K. L. High-performance fuzzy optimized deep convolutional neural network model for big data classification based on the social internet of things. *The Journal of Supercomputing*. 2023. DOI: https://doi.org/10.1007/s11227-022-04974-7

13. Tanantong T., Yongwattana P. A convolutional neural network framework for classifying inappropriate online video contents. *IAES International Journal of Artificial Intelligence*. 2023. Vol. 12, Iss. 1. P. 124–136. DOI: https://doi.org/10.11591/ijai.v12.i1

15. Transfer learning based recurrent neural network algorithm for linguistic analysis / Sasikala S. et al. *Concurrency and Computation: Practice and Experience 34.5*. 2022. DOI: https://doi.org/10.1002/cpe.6708

16. Джус С. І. Потреба використання DATA SCIENCE & BIG DATA ANALYSIS (Наука про дані та аналіз великих даних) у сучасному статистичному та фінансовому світі. *Бізнес-аналітика в управлінні зовнішньоекономічною діяльністю* : матеріали IV Міжнар. наук.-практ. конф. Київ, 2017. С. 51–56.

17. Дранишников Л. В. Нечітке і нейромережеве моделювання в системах управління. *Міжнародний науковий журнал «Грааль науки».* 2021. № 5. С. 153–159.

18. Технологія – Технологія. *Горох* : вебсайт. URL: https://cutt.ly/44Cqnyp (дата звернення: 31.03.2023).

19. Терещенко Л., Гладьо С. Неправдивість письмового тексту: шляхи її визначення. *Psycholinguistics*. 2022. № 31 (2). С. 116–136. URL: https://cutt.ly/Q42Rfwg (дата звернення: 31.03.2023).

20. Філософський енциклопедичний словник : енциклопедія / НАН України, Інститут філософії ім. Г. С. Сковороди ; головний редактор В. І. Шинкарук. Київ : Абрис, 2002. 742 с.

Information about the author:
Dovhan Oleksii Valentynovych,
Candidate of Philological Sciences,
Doctoral student at the Department of Slavic Languages
Drahomanov Ukrainian State University
9, Pyrohova str., Kyiv, 01601, Ukraine