COMPARATIVE LITERATURE STUDIES

DOI https://doi.org/10.30525/978-9934-26-348-4-32

METHODOLOGICAL FEATURES OF NEURAL NETWORK MODELING OF THE PROCESSES OF RECOGNITION OF LINGUISTIC MARKERS OF THE CATEGORIES OF SENSE AND ABSURDITY

МЕТОДОЛОГІЧНІ ОСОБЛИВОСТІ НЕЙРОМЕРЕЖЕВОГО МОДЕЛЮВАННЯ ПРОЦЕСІВ РОЗПІЗНАВАННЯ МОВНИХ МАРКЕРІВ КАТЕГОРІЙ СМИСЛУ Й АБСУРДУ

Dovhan O. V.
*Candidate of Philological Sciences,
Doctoral Student at the Department of
Slavic, Romance and Oriental Languages
Drahomanov Ukrainian State University
Kyiv, Ukraine*

Довгань О. В.
*кандидат філологічних наук,
докторант кафедри слов’янських,
романських і східних мов
Український державний університет
імені Михайла Драгоманова
м. Київ, Україна*

The language polysystem is the pivotal means of human communication, which is why understanding the nature of its functioning, the peculiarities of the processes that take place in it (in particular, sense generation, transformation, modification of senses, etc.) is of great importance for human ontology. We are talking about a number of fields in general and Natural Language Processing (hereinafter – NLP), its vectorization, tokenization, lemmatization, speech interface (Google Assistant and others), automatic translation (Google Translate, Reverso, DeepL, etc.) and other areas [1].

The above makes it possible to position neural network modeling as a relevant tool for processing linguistic data, carried out with the aim of further updating the latter in a number of areas of linguistic research (analysis, classification, structuring and identification of hidden patterns in large volumes of heterogeneous complexly structured data) [3].

The core stages of methodological approaches to neural network modeling of the processes of recognizing linguistic markers of the categories of sense and absurdity are:

1. Textual data preprocessing.

During this stage, the linguist-developer processes the data set in a certain way, highlighting the aspects that are important for the research. This stage precedes the direct classification of language data, consisting in the process of tokenization (instance of a sequence of characters in the part of the analyzed data set that will be accumulated for actualization in the process of semantic processing) of the text. The latter consists in splitting the text into token words with simultaneous filtering of information noise – special characters (primarily punctuation).

It is noteworthy that in order to implement this stage, it is necessary to understand the typological features of the above elements (tokens): thus, a *type* is a class of all tokens containing a certain symbolic identity, while a *term* is a type included in the *Information retrieval system's dictionary*. Thus, the tokens appearing in a document are derived from terms by updating different approaches to normalization, while a number of term indices may be different from such tokens [5].

In addition to the above-mentioned tokenization, it is also advisable to update at this stage the *cleaning of language data from non-alphabetic characters* (for example, replacing all non-lexical characters with spaces, etc.), *lemmatization* (the process of transforming word forms into *lemmas* – their normal dictionary form, similar to the selection of the base of each lexical unit in a sentence: for example, for nouns and adjectives – nominative case, and for verbs, participles, and adverbs – verbs in the infinitive), *removal of stop words* (removal of articles, interjections, conjunctions, etc. that are not representative of the sense), *vectorization* (this strategy is called "Bag of words" representation, its essence is that that the data is represented by the occurrence of words, while completely ignoring information about their relative position in the text data, this strategy finds the number of occurrences of each word in the data set) of text data, which will ensure the transformation of the latter into numerical data, and those, in turn, are processed by the selected type of neural network [2].

2. The process of preparing (representing) language data.

This is a key stage of the neural network modeling process, because the correctness of the algorithm depends on the representativeness of the data set, which contains sense and absurd language markers. Naturally, a neural network should be trained on representative data that is fundamental in terms of covering the problem under research. It is noteworthy that data representation correlates with a specific application task, because for each of these tasks it is necessary to choose specialized methods that are productive for its solution.

The most common methods of data representation in the process of training a neural network or machine learning (hereinafter – ML) are: *feature description of an object* (the features that are fed as input should be relevant to the label that we get as output – the result of a neural network or a complex of neural networks: for example, morphological features will not be or are not sufficiently representative for lexicographic research, while they are relevant for semantics research) and *selection of key features* (the decision of relevant features in relation to the research goal is individualized, since tracking correlations between certain features is an assumption of the developer's linguistic research, within which some features are suitable for prediction and others are not).

3. Selection of the neural network architecture (type).

This step is the second most important after the process of data preparation (presentation), because the success of neural network modeling depends on the choice of the tool for analysis, as well as its feasibility, functional characteristics, etc. That is why it is necessary to take into account the type of neural network, which is determined by their typological and functional specificity. In particular, it is advisable to use *recurrent neural networks* (hereinafter – RNNs) for contextualization, *convolutional neural networks* (hereinafter – CNNs) or *transformers with attenuation* for text comprehension, and a combination of RNNs and CNNs will be productive for studying the peculiarities of neural network modeling of the processes of recognizing language markers of the categories of sense and absurdity.

In the above modeling, lexical items are represented as points in hyperspace that correspond to certain senses, and sentence construction is positioned by moving along the above points in a multidimensional space. We are talking about the vectorization of language data, i.e., the transformation of textual data (textual input) into vectors (vector representation) that can be potentially processed by a neural network.

4. The process of algorithm design.

The choice of actualized methods correlates with the specific task of the research carried out by the developer-linguist: for example, to study the recognition of linguistic markers of the categories of sense and absurdity, the methods of classification, clustering, regression, ranking, as well as genetic algorithms (despite the significant drawback of the latter in the form of the lack of induction (data analysis and model building on their basis) learning) will be productive [4]. It is noteworthy that ML is based on training data, and when updating it, it is also necessary to use methods such as *backpropagation* and optimization algorithms (e.g., gradient descent, etc.) to improve the results.

5. The process of training the algorithm on the available (prepared by us) data with its subsequent validation on them.

At this stage, the speech data should be divided into two groups: *training data* (intended for calibrating the system's weights, when the required value is reached, it is advisable to determine the loss function (representing the difference between the expected and correct results of the neural network model), since it is representative of the error in the classification of markers) and *testing data* (during which we validate the results of the neural network model, which can be determined using special metrics: accuracy, F1-mean, and confusion matrix, which represent the efficiency of the system) sample.

Bibliography:

1. Comparing natural language processing (NLP) applications in construction and computer science using preferred reporting items for systematic reviews (PRISMA) / S. Chung et al. *Automation in Construction*. 2023. https://doi.org/10.1016/j.autcon.2023.105020 *ScienceDirect* : web-site. URL: https://goo.su/RcApWzc (date of application: 31.08.23).

2. Guo F. Revisiting Item Semantics in Measurement : A New Perspective Using Modern Natural Language Processing Embedding Techniques : Doctoral dissertation / Bowling Green State University. Ohio, 2023. 137 p. *OhioLINK* : *ETD Center* : web-site. URL: https://goo.su/78fUQ (date of application: 31.08.23).

3. Heimann M., Hübener A. F. Circling the Void : Using Heidegger and Lacan to think about Large Language Models. https://doi.org/10.21203/rs.3.rs-3023378/v2 *Research Square* : web-site. URL: https://goo.su/vRNiU (date of application: 31.08.23).

4. Jones M., Neumayer C., Shklovski I. Embodying the Algorithm : Exploring Relationships with Large Language Models Through Artistic Performance. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023. P. 1–24. https://doi.org/10.1145/3544548.3580885 *ACM Digital Library* : web-site. URL: https://goo.su/vzhN1 (date of application: 31.08.23).

5. Яцишин В. В., Давидов А. О., Подолян Д. О. Класифікація та препроцесинг текстових даних. *Збірник тез доповідей VII Міжнародної науково-технічної конференції молодих учених та студентів «Актуальні задачі сучасних технологій», 28-29 листопада 2018 року*. Тернопіль : ТНТУ, 2018. Том 2. С. 200. *CORE* : web-site. URL: https://goo.su/X9z8 (дата звернення: 31.08.23).