# DEVELOPMENT OF THE STRUCTURAL INTERACTING TECHNOLOGIES SCHEME FOR TRAINING OF NEURAL NETWORKS

**Sabina RUZUDZHENK**

V.N. Karazin Kharkiv National University, Kharkiv, Ukraine,

## ABSTRACT

An artificial neural network (NN) is a device of parallel computing, which consists of many interacting processors. Such processors are usually very simple, unlike those used in computers. Each of them works only with signals that it receives and sends at certain intervals. However, when combining such locally simple elements into a sufficiently large network with controlled interaction, it is possible to solve a wide range of quite complex problems.

The most important property of NN is the ability to learn based on the processing of environmental data and as a result of learning to increase the level of system performance. Productivity increases over time, according to certain rules. NN training takes place through an interactive process of adjusting synaptic weights and thresholds. At best, NN acquires new knowledge about the environment at each iteration of the learning process.

Natural language processing (NLP) is the application of computational methods to model and extract information from human language. With the rise of social media, conversational agents, and personal assistants, computational linguistics is increasingly relevant in creating practical solutions to modeling and understanding human language.

In the course of this work, a study of technologies used to teach neural networks without a teacher (unsupervised learning), features and methods of teaching neural networks to solve NLP (Natural Language Processing). Features of natural language processing (NLP) for learning neural networks, methods of morphological, lexical, syntactic, semantic, discourse analysis, methods of classification and clustering of text data, the process of vocabulary formation were considered. Technologies of vector representation of the text, the process of predicate formation (Claim) and its negation (Claim Negation), as well as the process of synthesis of new predicates (Claim Synthesis) are described.

The aim of the work was to develop a structural scheme of interacting technologies for learning neural networks on the example of solving two basic problems of natural language processing – text analysis (Natural Language Understanding, NLU) and text generation (Natural Language Generation, NLG). The proposed scheme allows to understand the principles of training the network to work with NLP and demonstrates the basic technologies used for this purpose.

**Keywords**: neural network, unsupervised learning, natural language processing.

## INTRODUCTION

An artificial neural network (AR) is a device of parallel computing, which consists of many interacting processors. Such processors are usually very simple, unlike those used in computers. Each of them works only with signals that it receives and sends at certain intervals. However, when combining such locally simple elements into a sufficiently large network with controlled interaction, there is an opportunity to solve a wide range of quite complex problems.

The most important property of NM is the ability to learn based on the processing of environmental data and as a result of learning to increase the level of system performance. Productivity increases over time, according to certain rules. NM training takes place through an interactive process of adjusting synaptic weights and thresholds. At best, NM acquires new knowledge about the environment at each iteration

of the learning process. Natural language processing (NLP) is the application of computational methods to model and extract information from human language. With the rise of social media, conversational agents, and personal assistants, computational linguistics is increasingly relevant in creating practical solutions to modeling and understanding human language.
The aim of the work was to develop a structural scheme of interacting technologies for learning neural networks on the example of solving two basic problems of natural language processing - text analysis (Natural Language Understanding, NLU) and text generation (Natural Language Generation, NLG).

**NEURAL NETWORK ARCHITECTURE**

In natural language processing tasks, due to the possibility of using internal memory to process sequences of arbitrary length, the most common are recurrent neural networks (RNN). Recurrent neural networks are a type of neural networks where connections between elements form a directed sequence. This property allows you to process a series of events over time or successive spatial chains. Of the many architectural solutions for recurrent networks, the most commonly used are Long-Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). Long-term short-term memory networks are a special type of RNS capable of learning long-term dependencies (Fig. 1).

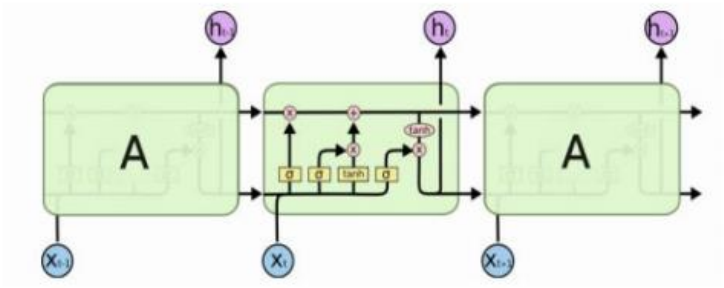


Fig. 1 – LSTM network architecture

Each line transmits an entire vector from the output of one node to the inputs of others. Small circles represent point operators, such as addition of vectors, while small rectangles are trained layers of a neural network. Solid lines indicate concatenation, while branched lines indicate that their contents are being copied and copies are being sent to different locations. Networks of the managed recurrent unit type (Fig. 2) combine LSTM gateways to create a simpler data update rule, which reduces system complexity and increases its efficiency.

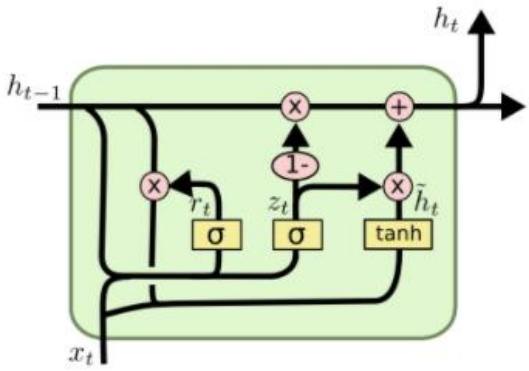


Fig. 2 - Architecture of the controlled recurrent block

Comparing these architectures, it is difficult to say which is more efficient, so the choice of architecture is motivated by the specific task for which the network is used. In the general case, other things being equal, use a controlled recurrent unit, because it requires fewer parameters.

## METHODS OF NATURAL LANGUAGE PROCESSING

Natural language is a language that has evolved naturally in the process of everyday use by humans over time, without formal construction. It covers a wide range, including spoken and sign languages. Natural language is ambiguous, both in written and oral form. Because of this ambiguity, it is difficult for computers to process and understand natural language. This complexity is especially true of aspects of language such as emotional coloring, as well as the use of various paths. In any language there is ambiguity in semantics, grammatical structure of words and sentences, so use different methods of natural language processing, which allow to solve each of these ambiguities.

Natural language processing (NLP) is a term that refers to various methods of computational processing of human languages. There are several types of classification problems typical of NLP: determining the beginning and end of words by characters; identifying the beginning and end of a sentence by words; identifying the beginning and end of word combinations by words in a sentence; identifying the word belonging to a certain part of speech; identifying the beginning and end of named entities by words in a sentence; detection of emotional coloring by words in a sentence.

In the analysis of natural language, language characteristics are often grouped into a set of categories also named the NLP pyramid (Fig. 3). With the rise of social media, conversational agents, and personal assistants, computational linguistics is increasingly relevant in creating practical solutions to modeling and understanding human language. For text analysis, these categories are morphology, vocabulary, syntax, semantics, discourse, and pragmatics. Morphology refers to the form and internal structure of a word. Lexical analysis involves segmenting the text into meaningful units, such as words. Syntax refers to the rules and principles that apply to words, phrases and sentences. Semantics – to the context that gives meaning to the sentence. It is semantics that ensures the effectiveness of human language. Discourse analysis is applied to conversations and the relationships that exist between sentences. Pragmatics – to external characteristics, such as the intention to convey the context.

For language analysis, characteristics are usually grouped into categories of acoustics, phonetics, phonemics, and prosody. Acoustics are the methods we use to represent sounds. Phonetics analyzes how sounds are compared to phonemes, which serve as the basic units of language. Phonemes, also known as phonology, refer to how phonemes are used in a language. Prosody analyzes the nonverbal characteristics that accompany speech, such as tone, stress, intonation, and pitch. Computational and statistical methods that can be applied to each component of the synchronous model form the basis of natural language processing.

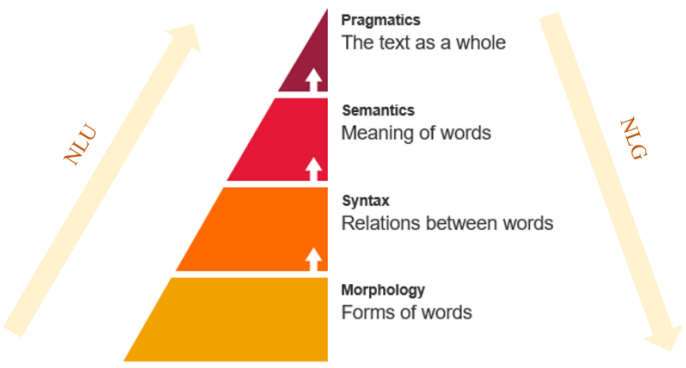


Fig. 3 – Stages of the language processing by categories

Natural language processing seeks to compare language with representations that capture morphological, lexical, syntactic, semantic, or discursive characteristics that can then be processed by learning methods. The choice of representation can have a significant impact on further tasks and will depend on the chosen learning algorithm for analysis.

## DEVELOPMENT OF THE SCHEME OF INTERACTING TECHNOLOGIES

Based on the research of methods of text analysis and generation, a structural scheme of interacting technologies was developed, which demonstrates the sequence of stages for learning the neural network used to solve the problem of text processing and generation (Fig. 4). According to the given scheme the analysis and generation of the text is divided into the following stages: preliminary processing of the given corpus (marked database); presentation of text data in vector form, on the basis of which a dictionary is formed for further work; direct text generation.

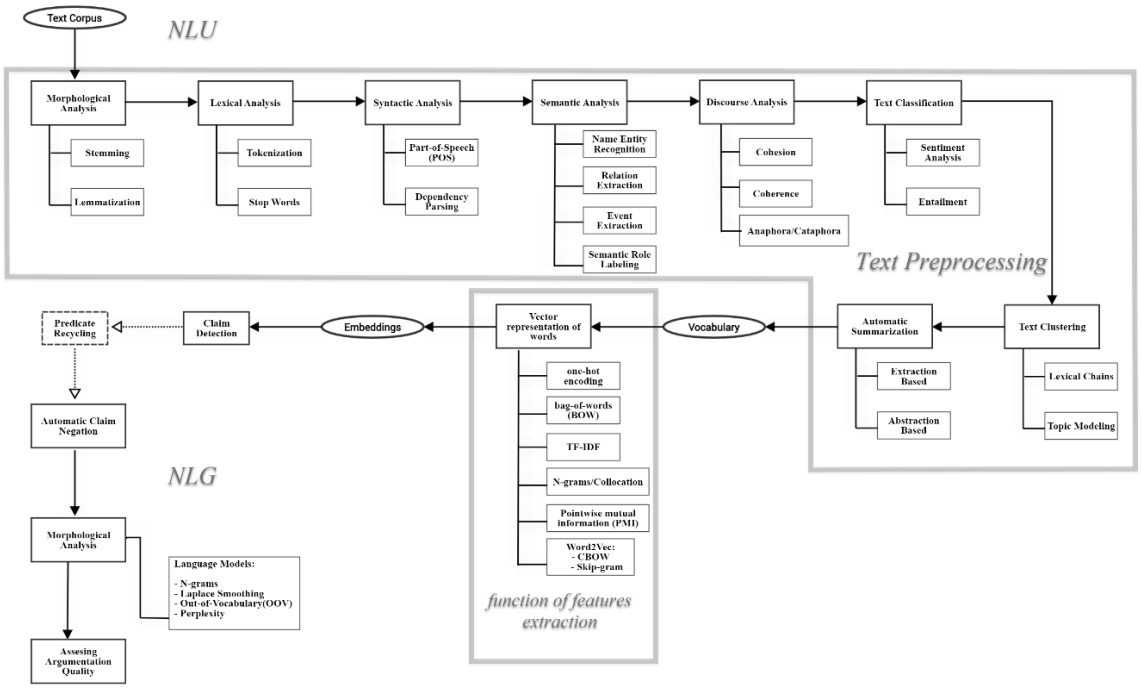


Fig. 4 – Scheme of interacting technologies for neural network training

During preprocessing, the corpus (marked database) goes through several stages of analysis:

1. Morphological analysis. The two most common approaches to solving this problem are stemming and lemmatization.

Often, the word ending is not as important as the root word itself. This is especially true of verbs, where the verb root may hold significantly more meaning than the verb tense. If this is the case, computational linguistics applies the process of word stemming to convert words to their root form (Fig. 5).

Lemmatization is closely related to stemming in that it is an algorithmic process that removes inflection and suffixes to convert words into their lemma. Lemmatization results are very similar to those of stemming, except that the results are actual words. Whereas stemming is a process where meaning and context can be lost, lemmatization does a much better job (Fig. 5).

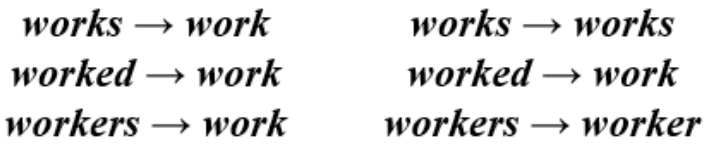


Fig. 5 – Examples of stemming and lemmatization

2. Lexical analysis. Tokenization and removal of stop words are used for analysis.

Tokenization is segmenting text into relevant words or units of meaning. Tokens may be words, numbers, or punctuation marks. In simplest form, tokenization can be achieved by splitting text using whitespace, as on the example (Fig. 6). Tokenization serves also to segment sentences by delineating the end of one sentence and beginning of another.

In the written English language, common functional words like "the," "a," or "is" provide little to no context, yet are often the most frequently occurring words in text. By excluding these words in natural language processing, performance can be significantly improved.

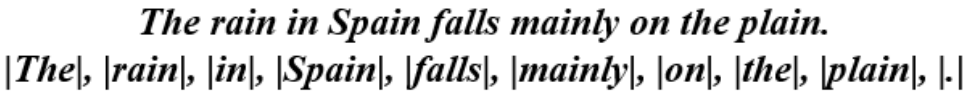


Fig. 6 – Examples of tokenization

3. Parsing. Computational linguistics uses a variety of approaches to extract these contextual cues, such as part-of-speech tags, partitioning, and parsing.

A part-of-speech (POS) is a class of words with grammatical properties that play similar sentence syntax roles. There are 9 basic part-of-speech classes (Fig. 7). It can be difficult to identify which category the word belongs. Part-of-speech tagging is the process of predicting the part-of-speech category for each word in the text based on its grammatical role and context within a sentence.

| | | |
|---|---|---|
| N | Noun | Dog, cat |
| V | Verb | Run, hide |
| A | Article | The, an |
| ADJ | Adjective | Green, short |
| ADV | Adverb | Quickly, likely |
| P | Preposition | By, for |
| CON | Conjunction | And, but |
| PRO | Pronoun | You, me |
| INT | Interjection | Wow, lol |

Fig. 7 – Basic part-of-speech classes

Parsing is the natural language processing task of identifying the syntactic relationship of words within a sentence, given the grammar rules of a language. The way is to link individual words together based on their dependency relationship. Dependency is a one-to-one correspondence, which means that there is exactly one node for every word in the sentence.

4. Semantic analysis. Semantic analysis is interested in understanding the meaning, first of all, in terms of the relationship between words and sentences, using technologies such as recognition of named entities, recognition of relationships and events, analysis of semantic roles.

Named entity recognition (NER) is a task in natural language processing that seeks to identify and label words or phrases in text that refer to a person, location, organization, date, time, or quantity (Fig. 8).

Semantic role labeling (SRL), also known as thematic role labeling or shallow semantic parsing, is the process of assigning labels to words and phrases that indicate their semantic role in the sentence. A semantic role is an abstract linguistic construct that refers to the role that a subject or object takes on with respect to a verb. These roles include: agent, experiencer, theme, patient, instrument, recipient, source, beneficiary, manner, goal, or result. Semantic role labeling can provide valuable context, whereas syntactic parsing can only provide grammatical structure.

| | |
|---|---|
| Person | George Washington |
| Location | Washington State |
| Organization | General Motors |
| Date | Fourth of July |
| Time | Half past noon |
| Quantity | Four score |

Fig. 8 – Named entity recognition example

5. Discourse analysis and classification and clustering of the text, which are carried out through the analysis of tonality, lexical chains and thematic modeling.

Discourse analysis is the study of the structure, relations, and meaning in units of text that are longer than a single sentence. Cohesion is a measure of the structure and dependencies of sentences within discourse. It is defined as the presence of information elsewhere in the text that supports presuppositions within the text. That is, cohesion provides continuity in word and sentence structure.

Coherence refers to the existence of semantic meaning to tie phrases and sentences together within text. It can be defined as continuity in meaning and context, and usually requires inference and real-world knowledge. Coherence is often based on conceptual relationships implicitly shared by both the sender and receiver that are used to construct a mental representation of the discourse. An example of coherence can be seen in the following example which presumes knowledge that a bucket holds water.

Anaphora refers to the relation between two words or phrases where the interpretation of one, called an anaphor, is determined by the interpretation of a word that came before, called an antecedent. Cataphora is where the interpretation of a word is determined by another word that came after in the text. Both are important characteristics of cohesion in discourse.

Since we must supply a vector to the input of the neural network, we need to convert each word in the dictionary into a vector form. To do this, you can use the following technologies. CBOW (Continuous Bag-of-Words) – architecture is based on a projection layer that is trained to predict a target word given a context window of words to the left and right side of the target word. Skip-gram – architecture is based on a projection layer that is trained to predict words around the target word.

The next stage of the processing is Claim Synthesis, which uses the following technologies. Claim Detection is an information extraction task, whose goal is to identify those sentences in a document that contain the conclusive part of an argument. Automatic Claim Negation – from the main predicate, after its detection, the predicate-negation is formed.

Then everything comes in a set – (claim, claim-negation). Now in the newly created text, simple predicates are selected and a match with the old simple predicates (claim or claim-negation) is sought. Links to confirmation or denial are also generated during this step.

The last stage of processing is assessing of arguments' quality. To assess the quality of the arguments generated by the machine is used modeling the quality of each individual argument as a real value in the range of [0, 1], by calculating the fraction of «yes» answers.

To ensure the annotators will carefully read each argument, the labeling of each argument started with a test question about the stance of the argument towards the concept (pros or cons).

## CONCLUSION

Neural networks provide an efficient learning engine that is extremely attractive for use in natural language processing tasks. The main component of a linguistic neural network is the immersion layer, that is, the mapping of discrete symbols to continuous vectors in a relatively small space. Immersion transforms words from isolated discrete symbols into mathematical objects that can be manipulated in various ways. In particular, if for the measure of the distance between words to take the distance between vectors, it will be easier to generalize the influence of one word on another. The network finds such a representation of words by vectors in the learning process. As it climbs up the hierarchy, the network also learns to combine word vectors in ways useful for prediction. This feature compensates to some extent for the discreteness and sparseness of the data.
In the course of this work the researching of technologies, which are used to neural networks' unsupervised learning, features and methods of training neural networks to solve the problems of natural language processing has been conducted.
The suggested scheme of interacting technologies allows to understand the principles of neural network training for natural language processing, using the example of solving two basic problems of natural language processing (NLP) – text analysis (Natural Language Understanding, NLU) and text generation (Natural Language Generation, NLG), even for people who do not specialize in working with neural networks, and demonstrates the main technologies for its implementation.

## REFERENCES

1. Basic concepts of neural networks. R. Callan, 2001 – 287 p.
2. Recurrent neural networks [Electronic resource]:
https://habr.com/ru/company/wunderfund/blog/331310/ (Date of application 25.01.2021).
3. Neural network methods in natural language processing. J. Goldberg, 2019 – 282 p.
4. Deep Learning for NLP and Speech Recognition. U. Kamath, J. Liu, J. Whitaker, 2019 – 637 p.
5. Natural Language Processing in Action. H. Lane, C. Howard, H. M. Hapke, 2020 –576 p.