41. NATURAL LANGUAGE PROCESSING: CURRENT STATE AND PERSPECTIVES

Svinkouski I.S.

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Kaspiarovish N.G. – Senior Lecturer

An overview of natural language processing is presented. Current status and test results of modern language models are summarized. Different types of language models are described. The capabilities and implications of the modern NLP models are discussed. Special attention is given to key features of transformer models.

Natural Language Processing (NLP) is an interdisciplinary subfield of computer science and linguistics associated with interaction between computers and humans via a natural language. This technology is aimed at solving a broad spectrum of tasks such as speech recognition, text analysis, and text generation. NLP technology has various real use cases including creating chat-bots, sentiment analysis, machine translation, etc. The example of Generative Pre-trained Transformer (GPT) model shows that investigation of new approaches of developing language models leads to a significant improvement in models performance. Therefore, a continual need exists for reviewing and updating the state-of-the-art in the area of NLP due to its significant practical benefits.

Scientists distinguish these main types of NLP models types [1]:

Rule-based models are historically the first type of architecture. They use prescripted algorithms and linguistic rules, based on syntax or grammar to process text. This approach can be used only for developing highly specialized systems but it is not possible to be used in general intelligence development.

Statistical models use probabilistic methods to process and generate texts by predicting the next word in sequence [2]. There are various types of statistical models, from simple N-Gram models, that try to predict the word sequence with length of *n*, to complex models, such as continuous space, that use machine learning. All statistical models are based on Markov assumption, which says that each word in the text depends on all previous words.

Neural networks models belong to the most common approach nowadays. They use deep learning algorithms to analyze text data. These models are often based on recurrent neural networks, convolutional neural networks, and transformer models.

Transformer models should be noted separately, as at the moment it is the most popular architecture for developing general language models such as BERT made by Google and GPT made by OpenAI. The key feature of the Transformer model is a self-attention mechanism [3] that allows the model to process the entire text at once. Transformer model can be applied to the sphere of machine translation. In this case, input and output hidden state sequences are connected with each other, which in turn creates the alignment between the source and target languages.

The invention of Transformer models creates a new group of large language models (LLM), such as GPT, BERT and PaLM, which shows a significant improvement in both classic and new implications [4]. Due to this survey, GPT-4 can solve specific problems without special prompting: from math and coding to medicine and law. Such result can be qualified as unexpected because the core of GPT model is a combination of simple algorithmic components – gradient descent and large-scale transformers.

The researches that would tackle the reasons why quantitative changes in amount of neural network parameters and amount of input data lead to significant qualitative changes in developing flexible intelligence. The success of GPT model is therefore a matter of the combination of such factors as making gradient descent more effective by connecting different minima and forcing neural networks to search for a deeper connection between tokens in the language [4]. Despite the impressive results, which were shown by OpenAI GPT models, the current LLMs have the following weaknesses: confidence calibration which means that the model does not know when it should be confident and when it is guessing; continual learning which implies that the model has no ability to learn new information; long term memory which states that the memory is limited to only 8000 tokens.

Although, researches noticed considerable empowerment of language models implications, nevertheless, the weaknesses should still be considered.

References:

1. Guide to Natural Nanguage Processing, an Introduction to NLP [Electronic resource]. – Mode of access: kili-technology.com/data-labeling/nlp#nlp-tasks-&-model/. – Date of access: 21.03.2022.

2. Statistical Language Modeling [Electronic resource]. – Mode of access: engati.com/glossary/statistical-language-modeling. – Date of access: 22.03.2022.

3. Recent trends int deep learning based natural language processing / T. Young [et al.] 2018. – 15 p.

4. Sparks of Artificial General Intelligence: Early experiments with GPT-4 / S. Bubeck [et al.] 2023.