TASKS AND METHODS OF NATURAL LANGUAGE PROCESSING

Svitlana Bobrovnyk

Lecturer,

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

As we live in the world of information and from birth people have to accumulate information, process it and use it in their life. At first children begin to acquire their native language, accumulate, process the information and learn the environment. Later other acquired skills are added such as learning foreign languages, and so on. To master a foreign language in full is not enough to learn just words and grammar. Fluent speaking implies knowing not only grammar structures and enriched vocabulary, but also knowing phrasal verbs, idioms, slang, stylistic devices and understanding the right lexical choice of words. The process of learning a foreign language includes listening to audio-recordings, watching videos with dialogs where people interact in different everyday situations, doing it the learners remember the tone of voice, speech tempo, facial expressions, gestures, i.e. verbal language that help to understand the speakers. Thus, gaining experience of mastering of all the speech activities people can understand foreign speech. Human brains accumulate the memory processes to use it later on. Natural-language processing as well as machine translation face similar challenges. The advent of technology has changed communication of modern society. The International Data Corporation (IDC) estimates that by 2025, people will interact with data-driven technologies on average every 18 seconds.

Natural language processing, NLP is a general direction of computer science, artificial intelligence and mathematical linguistics. It studies the problems of computer analysis and synthesis of natural language. For artificial intelligence, analysis means understanding the language, and synthesis means generating intelligent text. Solving these problems will mean creating a more convenient form of computer-human interaction. Natural language is the language commonly used for speaking and writing, which identifies a popular way for people to communicate with devices and people. The texts preserve what has been said and have long been an important format for

preservation of human knowledge. Books, newspapers, texts are the common information resources on the World Wide Web. Especially in social networks, users generate a huge amount of text data every second. In order to make human languages understandable to machines, research in natural language processing has been conducted since the late 1940s. At the beginning of NLP, machine translation was the focus of the study. Meanwhile, more and more research is being conducted in the areas of NLP and machine learning caused by the deep learning boom (DL), which happens due to improved performance of computer resources.

Natural language processing (NLP) is related to other areas, such as artificial intelligence, machine learning. Artificial intelligence is a very broad term and a way to describe systems that are able to "think". Artificial intelligence consists of four main parts: machine learning, reasoning, planning and NLP. Reasoning allows the machine to make hypotheses based on data, while planning gives the systems the ability to act autonomously when interpreting data. Natural language processing deals with the use of human languages by computer. It has many different applications that relate to the unstructured natural language of man. For example, its areas of application are machine translation, language recognition, and dialog systems, named object recognition, information retrieval, and text classification. Thus, the field of natural language processing covers all interactions between computer and person using written or oral natural language. The field of research and application is related to the manipulation and understanding of natural languages. Human language processing is based on understanding the intended meaning of the message, which is not always easy to understand even for people, for example, when irony is used. All components of natural language, such as phonetics, phonology, morphology, syntax, semantics, and pragmatics must be taken into account to gain a full understanding of the message (Bliznyuk, Vasilieva, Strelnikov & Tkachuk, 2017).

According to Elizabeth Liddy natural language processing is a computerized approach to text analysis based on a number of theories and a set of technologies. This industry does not have a common definition, as it is in a state of constant research and

development. However, there are certain aspects that would unite all existing definitions (Liddy, 2001).

Significant results were achieved in the process of studying natural language processing, including the development of powerful lexicographic systems, machine translation programs, electronic dictionaries, and others. However, there is a problem that has not been solved yet; it is rooted in the very nature of human language. The problem of understanding of human speech is precisely its ambiguity. You can select the following types of ambiguities: syntactic ambiguity which can be met in proverbs, for the processing of natural language it will be completely unclear what exactly is said in the sentence; semantic ambiguity, where certain words can have two completely different meanings, depending on the emphasis; the case ambiguity, where the word order can change the meaning of the phrase; reference ambiguity, e.g. in the phrase "Open the bag and get a wet umbrella, I want to dry it", the pronoun "it" will be semantically related to a wet umbrella, but for a machine that has no understanding of reality, this pronoun will refer to both the bag and to the umbrella.

Recognition of natural language requires a huge knowledge of the system about the environment and the ability to interact with it. Defining the meaning of the word "understand" is one of the main tasks of artificial intelligence. Scientists have developed methods of language processing, such as function selection and pre-processing, tokenization, stemming, lemitization, deleting of stop words. Removal of stop words is a very important approach for reduction of huge raw input space in NLP. Most languages have specific words that appear more often than others or do not contain much information about the content of the text, for example, auxiliary verbs or articles. In this regard, it often makes sense to exclude these so-called stop words in further analysis.

GPT-3 (Generative Pre-trained Transformer 3) is the third generation of the algorithm of natural language processing from Open AI. Large, natural-language computer models that learn to write and speak are the big step toward AI that can better understand and interact with the world. GPT-3 today is the largest and most competent model. GPT-3 can imitate handwriting with supernatural and sometimes bizarre

realism, making it the most impressive language model created using machine learning. But GPT-3 does not understand what it is written about, so sometimes the results are distorted and meaningless. Training requires a huge amount of computing power, data and money, which creates a large carbon footprint and limits the development of similar models by laboratories with exceptional resources (Massachusetts Institute of Technology [MIT], 2021).

Natural Language Processing is the general name of the area that spans many subsections. All of them usually use machine learning models, mostly neural networks, and the data of many conversations between people. Since human languages are constantly and spontaneously evolving, and computer needs clear and structured data, certain problems arise during processing and accuracy suffers. In addition, text analysis methods are highly dependent on the language, genre, topic, so additional configuration is always required. However, today many tasks of natural language processing are still being solved using deep learning of neural networks. One of the challenges that arises in the process of natural language processing can be considered the problem of synonymy, as a result of which one concept can be expressed in several different words. As a result, relevant documents that use synonyms for the terms specified by the user in the enquiry may not be defined by the system. The influence of the above phenomena is especially noticeable at creating machine translation systems. The problem is the difficulty of establishing a specific reflection of the actual semantic-syntactic structure of the sentence in its internal logical representation, which is automatically generated by the system. The solution of these types of ambiguities is possible by introducing additional values that will increase the knowledge of the program about a particular industry. Today, there are no programs that "understand" all types of ambiguities in a wide range of industries, but there are programs that can respond correctly to ambiguities in very narrow areas.

Summing up, we can conclude that in today's world, in an environment of ever-growing volumes of information, the analysis of textual data has great potential and wide application for speech processing. Restrictions on the use of speech processing systems in the most traditional applications allow us to conclude that potentially new

solutions in the field of speech recognition need to be found. In the next decade, the task of recognizing and understanding natural speech, regardless of language and speaker, will be central to speech technology. For translation into a foreign language, the technical means cannot yet take into account all aspects of the language in full, as the machine does not see non-verbal language and can only take into account the tone of the speaker's voice. To avoid misunderstandings and inaccuracies in translation, it is necessary to develop additional word processing systems for a more accurate understanding, which gives motivation to further research of the problem.

References

Bliznyuk, B. O., Vasilieva, L. V., Strelnikov, I. D., & Tkachuk, D. S. (2017). Modern methods of natural language processing. *Bulletin of V. N. Karazin Kharkiv National University*, 14-25.

Liddy, E. D. (2001). Natural Language Processing. *In Encyclopedia of Library and Information Science* (2nd ed.). New York: Marcel Decker, Inc.

Massachusetts Institute of Technology (2021). *MIT's report 2021.* Retrieved from: https://ain.ua/2021/02/25/10-texnologij-2021-budushhee/