2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo) | 979-8-3503-9882-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CINTI-MACRo57952.2022.10029432

The origin and primary areas of application of natural language processing

Zsolt Krutilla
Doctoral School of Applied Informatics and Applied Mathematics
Óbuda University
Budapest, Hungary
krutilla.zsolt@uniduna.hu

Attila Kovari
Institute of Engineering, Alba Regia Technical Faculty
Óbuda University
Budapest, Hungary
kovari.attila@uni-obuda.hu

***Abstract*—Natural language processing with computers, i.e. NLP, is increasingly gaining ground in today's IT. The purpose of this study is to analyze the articles published in the NLP topic and to summarize it from several aspects, especially with regards to the development and practical application of the technology. The idea of NLP as a technology was not formulated in the 21st century, but already in the 1950s, relevant research was carried out and theories were created regarding how a machine can become capable of interpreting human speech. Technology has come a long way from simple algorithms to Deep Learning, but it has not yet solved the problem of machine processing of natural language. The research has confirmed the idea that mathematical modelling and models based on statistical probability dominated the field in the early days of the discipline and before the advent of computers. However, despite technological progress, no solution has yet been found to the comprehensive and all-encompassing problem of natural language processing with computers. To determine the future research area of NLP, we need to take into account the text files on which the technology is based. From this perspective, we can focus on the insurance and banking sectors, where a series of administrative processes take place where centralized paper processing is the incoming data set of the process and a database going back several years may be available.**

Keywords*—Natural Language Processing, Deep Learning, Artificial Intelligence

I. INTRODUCTION

The purpose of this study is to map and summarize the history of natural language processing by computer and to explore the possibility of its current and future application areas, as well as to analyze the published scientific articles related to the applied methods. In today's the natural language processing with computers, i.e. NLP (Natural Language Processing) in information technology is increasingly gaining ground. However, the idea of NLP as a technology is not the 21st century. was formulated in the 19th century, but already in the 1950s relevant research was carried out and theories were developed regarding how a machine can interpret human speech. In the 1960s, King, Masterman, Ceccato and Yngve worked on a translation machine that founded the NLP technology itself. [1] With the advent of IT, the RENDEZVOUS system was the first natural language interface (NLI) database. In the 1970s, two transformational NLP solution methods were more widespread, the generative and statistical algorithm-based methods. In the case of the generative procedure, it takes the structure and rule systems of the two languages as a basis and determine the relationship between the two rule systems. In the case of the statistical method, the probability rules of the teaching sample form the relation. [2] The origins of NLP can therefore be traced back to the fifties, long before the advent of computers. Nowadays, mainly use Deep Learning procedures during NLP, given that now have very high-precision neural network models and huge databases and data warehouses to teach the models, but previously the use of statistical and various mathematical models was more typical. However, the current language models also have their drawbacks, especially considering that although there large a large database for teaching the models, these cannot be called clean data at an early stage, so in several cases use machine-generated student data, which, however, distorts the result from the aspect of "real speech". Our ultimate goal is to explore the path leading to this point and, in the long term, find the NLP technology that can analyze and interpret natural speech in the same way as humans. However, before the technological definition, it is important to clarify the concept of the problem itself and the research objectives in the field of NLP, which path begins with the definition and formulation of the concepts.

NLP itself is a difficult concept to grasp, but do not even have a precise definition that would fully cover what it means. If want to put it into words, it can sum it up by saying that under NLP means computer solutions that are able to linguistically analyze everyday natural texts on several levels in order to achieve human-like language processing for many tasks or applications.

NLP itself is a difficult concept to grasp, but do not even have a precise definition that would fully cover what it means. If want to put it into words, it can sum it up by saying that under NLP means computer solutions that are able to linguistically analyze everyday natural texts on several levels in order to achieve human-like language processing for many tasks or applications. However, the definition is far from perfect, can add several comments. First of all, let's take a closer look at the "computational solutions", because there are many methods or techniques to choose from to perform a specific type of language analysis. Secondly, "natural texts", given that this means that the texts can be in any language, genre, written or spoken. The only requirement is that they be in a language that people use to communicate with each other. Also, the analyzed text should not be constructed specifically for analysis, but the text should be collected from actual use.

The concept of "levels of linguistic analysis" refers to the fact that several types of language processing are at work when people learn or understand language. Conceptually, people tend to use these levels because each level conveys a different type of meaning. However, different NLP systems use different levels of linguistic analysis, or combinations of levels, and this is reflected in the differences between different NLP applications. This leads to a lot of confusion among non-experts as to what NLP actually is, because a system that uses any subset of these levels of analysis can be said to be an NLP-based system. So, the difference may actually be whether the system uses "weak" or "strong" NLP.

"Human-like language processing" highlights that NLP is considered a branch of artificial intelligence (AI) and although the entire line of NLP performs a specific task. Accordingly, there are information retrieval (IR) systems that use NLP, machine translation (MT), and question answering systems that depend on many other disciplines. Since NLP strives for human-like performance, it makes sense to consider it an AI science.

For "numerous tasks or applications", he points out that NLP is not generally seen as an end in itself, except perhaps by AI researchers. For others, NLP is nothing more than a tool to accomplish a specific task. [3]

II. Methodology

The purpose of the analysis is to study the published scientific articles related to the topic of NLP and to map the depth and direction of the discipline. For this, it is necessary to define the given topic and set goals. An essential part of these is an overview of the history and development stages of the NLP discipline, as well as the current areas of application. I have selected the domestic and mainly international literature accordingly.

A. Research methodology

To select the articles and define the direction, I proceeded systematically by applying the classical research methodology, combining empirical experiences and the results of qualitative analyses, for which the main topics were first defined. From the aspect of the specifics of the discipline, the discussed topic can be divided into separate topics, such as the topic of applied methods, the origin and development of the discipline (as well as its stages), the relationship of linguistics with IT, deep learning learning) development and application, the development of mathematical language models, and the structures and storage technologies of teaching data. However, the joint analysis of these topics can take years, so I narrowed the scope to the development of the discipline and the range of applied technologies. During the research, I searched for the following keywords:

- "natural language processing",
- "acceptance of natural language processing",
- "history of nature language processing",
- "technology of nature language processing",
- "methods in nature language processing"
- "algorithms of nature language processing"
- "deep learning in nature language processing"

B. Selection of search engines

To select research topics, I used Google Scholar, Sci-Hub and Z-Library search engines and databases. I searched for the same keywords in all three search engines and, based on reading the topics and abstracts, I narrowed down the scope to the content of those articles and books that are relevant from the aspect of the analysis.

During the literature search, the searchers did not only bring up scientific articles and books that were relevant, but I ignored these search results. Due to the size and topicality of the research area, I narrowed down the results to those results that contained relevant information, which focused mainly on the development of NLP, its rudimentary algorithms and the areas that resulted in today's deep learning process (ignoring the scale of the development of hardware technology, however, cannot ignore the importance of this either).

III. The Development of NLP as a Discipline and its Application Areas and Methods

A typical approach to document indexing and query processing is as follows. First, a tokenization process takes place, and then the stop words are removed. In addition, natural language processing techniques can identify phrases or split compounds. [4]

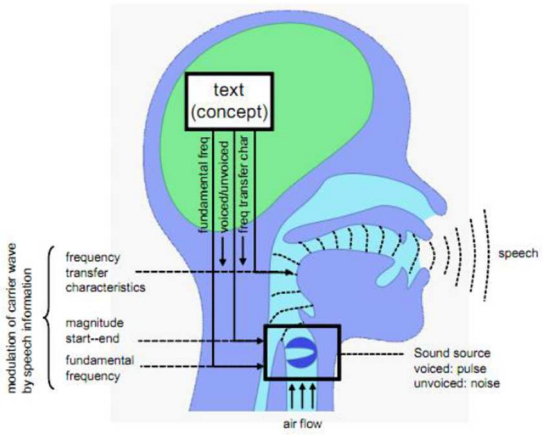

Fig. 1: Natural speech production process [29]

Stop words are words with little meaning that are removed from the index and query. In NLP, stop words can also be considered as negative patterns that are used when teaching LSTM neural networks. Stop words may have little meaning in terms of frequency or concept. Removing very common words does not affect the ranking of documents as much. If words have little conceptual meaning, they can be removed regardless of whether their frequency is high or low in the collection. In fact, it is especially important to remove these words when their frequency is low, because these words affect the ranking of documents the most. Stop words can be removed for conceptual reasons by using a stop list that lists all words with little meaning, typically function words such as "is", "and", "a", etc.

Broadly speaking, two types of text analysis techniques are used in the literature, keyword counting and mapping analysis.

The first method, keyword counting, typically uses a computer to automatically measure the frequency of keywords that researchers have previously determined are related to an interesting construct, and this has been the most widely used technique to date. [5] The popularity of the technique stems from the fact that the frequency of keywords is an objective and intuitive indicator of the construction's size or relative importance. [6]

The second method, the so-called "mapping analysis", during which the relational relationships between concepts are displayed in the texts, and is mostly a complementary technique to the former. [7] Text analysts are usually trained to identify relevant concepts and relationships in texts and read through documents until they arrive at a reliable representation of the cognitive relationships underlying the texts. [8]

Although both techniques have become accepted text analysis methods in the literature, the two traditional techniques represent an obvious compromise. By counting keywords, researchers can extract narrow measures from the content of the text, but they cannot take advantage of the detailed and valuable information that lies in the broader structure of the words of the texts. [9] The latest developments in machine-assisted NLP methods offer new possibilities for the analysis and interpretation of textual data. [10] Modern NLP allows researchers to use computational algorithms to extract deeper meaning structures from large amounts of text and with the help of this, the possibility of analysis opens up.

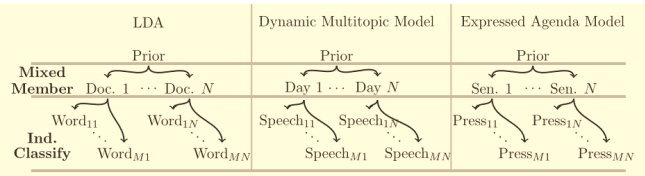

Fig. 2: Common structure across seemingly unrelated text models [10]

In summary can say that text mining and natural language processing means the understanding and analysis of "natural language" with the help of computer algorithms and programs and is an important research direction in the application field of artificial intelligence. With continuous and extensive research in machine learning and data mining algorithms, existing text mining technologies have achieved good results in automatic abstraction, automatic question answering, web relational network analysis, and anaphora resolution. [11]

A. Stages of evolution of NLP

The development of NLP can be divided into three major stages (eras). The first stage is the so-called "rationalism" stage, which can be traced all the way back to the Turing test created by the mathematician Alan Turing. The essence of the test is that a person and a computer communicate with each other, and the test is said to be successful if the subject cannot decide whether he is talking to a machine or a person. In 1954, the Georgetown-IBM experiment presented the first translation machine, which was able to translate 60 Russian sentences into English. The success of rationalist approaches in NLP is mainly due to the widespread acceptance of Noam Chomsky's arguments and his criticism of the N -gram method. [12]. The foundations of the first neural network-based programs and speech recognition procedures were created in this phase.

The second stage is the so-called empiricism stage. The second wave of NLP was characterized by the exploitation of data corpora and machine learning – whether statistical or otherwise – to use computers for such purposes. Much of the structure and theory of natural language has been ignored or discarded in favor of data-driven methods. The main approaches developed during this period were called empirical or pragmatic. [13] In contrast to rationalism, the approaches assume that the human mind begins only with the general operations of association, pattern recognition, and generalization. Rich senses are needed for the mind to learn the detailed structure of natural language. The empiricism that spread in linguistics between 1920 and 1960 was revived from 1990. Since the late 1990s, discriminative models have become a mainstay of many NLP tasks. Representative discriminative models and methods in NLP include the maximum entropy model for support vector machines, conditional random fields, maximum mutual information and minimum classification error, and perceptrons. [14]

The third stage is none other than deep learning itself learning) age. In the first and second stages, NLP systems included speech recognition, language understanding, and machine translation, but they were far from human-level performance and left much to be desired. With few exceptions, they were not suitable for learning through large amounts of teaching data. The teaching algorithms and methods - and not even the technology itself - were strong enough for this kind of teaching. The technological breakthrough came only a few years ago, with the third wave of new type, paradigm-changing deep learning algorithms. In traditional machine learning, features are designed by humans, and feature design itself is the bottleneck, requiring significant human expertise. At the same time, the related shallow models lack the representational ability and, consequently, the ability to form levels of decomposable abstractions that would automatically disentangle the complex factors shaping the observed linguistic data. These problems are solved by deep machine learning, in the form of a layered model structure and often neural networks, and the use of related "end- to-end" learning algorithms. [15]

B. Deep Learning method in the field of NLP

Deep Learning is the subfield of machine learning that deals with algorithms inspired by the structure and function of the brain, known as artificial neural networks. Deep learning excels in problem domains where the inputs (and even the outputs) are analog. This means not some quantities in tabular form, but images of pixel data, documents of textual data or files of audio data. The promise of deep learning in the field of natural language processing is the improved performance of models that may require more data but less linguistic expertise to train and operate. [16] There are five main tasks in natural language processing, these are:

- classification,
- consultation,
- translation,
- structured forecast,
- sequential decision process.

Most of the problems of natural language processing can be formalized with these five tasks. In tasks, words, phrases, sentences, paragraphs, and even documents are usually seen as a series of tokens (strings) and treated similarly, although they differ in complexity. In fact, sentences are the most widely used processing units. Among the NLP problems, the progress made in the field of machine translation was particularly noteworthy (we can experience this when using Google Translate). Neural machine translation, i.e. machine translation using deep learning, significantly outperformed traditional statistical machine translation. [17]

C. The role of NLP in education

NLP generally focuses on the development of educational software systems and instructional strategies that can support the use of natural languages in education (e.g. e-Assessment and Text Adapter). Software systems with NLP can identify the process of language learning under natural conditions. [18]

NLP is also a powerful method for developing a system that processes linguistic input in the natural environment

through various words, sentences and texts. Natural language processing also uses various grammatical rules and linguistic approaches such as derivations, grammatical tenses, semantic system, lexicon, corpus, morphemes, tenses, etc. All of these effective approaches can be used in the educational setting to improve student understanding of educational material and curriculum. NLP is also a widely known solution in the field of language learning all over the world, and it is successfully used as an effective way to improve language teaching systems. Similarly effective and efficient learning can be achieved through visual-based, problem-solving, algorithmic thinking [19][20]. In most studies, English is the most common language, which shows that NLP is effectively used in language learning. NLP is also an effective approach to improve the education system in Arab countries. [21]

IV. Discussion

In the last decade, the technical development in the field of artificial intelligence has brought a significant breakthrough in all areas dealing with research and application of AI technology, including in the field of NLP, however, in this study also saw that the processing of natural language by machines was a concern of the scientists and several solutions (theories) were created in the topic, which were successfully applied. The history of NLP can be traced all the way back to the Touring test created by the famous mathematician, which was only the formulation of a proposal, but even then, it pushed the limits of machine possibilities for natural language processing. The first, so to speak, tangible breakthrough had to wait until the appearance of computers, which were capable of putting theories into practice and proving their success.

During the research, the idea was confirmed that in the early stages of the discipline and before the advent of computers, mathematical modeling and models based on statistical probability began to gain ground in the field. With the advent of information technology (computers), the methods and applied procedures also changed, so to speak, they followed the technological development and increasingly utilized the resources inherent in the technology. With the development of technology, databases appeared, which put the science of NLP on a new foundation. The era of data-driven text analysis began, when it became possible to store, manage and analyze large data sets. A new tool was available to the researchers, which cannot yet be called artificial intelligence, but similar to today's more modern LSTM (Long - Short -Term- Memory) neural model, it was possible to analyze text based on previous data, observe and identify "exceptions" for installation (stop words). In the discipline of NLP, this stage or era can be considered one of the most significant stages, one could say a paradigm shift. During the research, in addition to the stages of development of the discipline, it was revealed that NLP technology was mainly used for language translations in this era.

Although based on previous results, NLP as a discipline achieved significant results even in the era of the data-driven method, but the real breakthrough came with the appearance and successful implementation of artificial intelligence. The newly emerging deep learning method put the work of scientists involved in NLP research into a new perspective and new approaches and methods appeared.

Conclusion

The study gained insight into the starting point of natural language processing by machines, revealed that the theory of the discipline was created long before the birth of computers and despite the fact that are achieving greater and greater breakthroughs in the deep learning using methods, still do not have a procedure or methodology that is really designed to solve the NLP problem. To this day, NLP is still a problem to be solved scientifically, for which there is still no exact and uniform solution.

In addition to translation machines, there is a significant need for natural language analysis, which not only provides the opportunity for machine interpretation of texts written in human language, but also provides us with a communication tool that can take communication with computers to a new level, during which they can even perform a programming task will be carried out by simply formulating an expectation to the computers, which can interpret it and prepare the program that performs the task like other IT applications [22-24].

Eye movement tracking systems can be used to analyze the intelligibility and readability of the source codes created in this way, with the help of which, thanks to an objective measurement system, the subsequent maintenance costs of the systems created in this way can be predicted and estimated [25-27].

Another (and of course much more immediate) use of NLP could be in the interpretation of banking and insurance documents and the further work (task completion). At present, the interpretation of textual documents is very important in these two areas, so NLP can take the office work process to a new level.

From an NLP perspective, another exciting area could be to navigate the maze of legal systems and identify legal cases, either in the private sector or in the area of judicial decision-making. In terms of legal regulation, NLP technology could even make decision-making an automated process, but we cannot neglect the importance and role that NLP could act in this area [28].

From an NLP perspective, the legislative field is more of a future-oriented field than a currently researchable discipline, but at the banking level we are more able to deal with the subject at a scientific level. A bank is engaged in a number of activities in which we can make extensive use of and research the potential of NLP. According to that, I intend to pursue the research direction in the banking sector, with a particular focus on exploring the possibility of automating the complaints handled by the central customer department of back-office using NLP technology. The area has a database and data warehouse that can provide clean learning data for model training, and the opportunity to compare and analyze different deep learning technologies and identify the AI method that is best suited to solve the problem at hand.

References

[1] Y. WILKS, "The history of natural language processing and machine translation." *Encyclopedia of Language and Linguistics*, 2005, 9.

[2] K. László, "Development of a semantic graph-based sentence analysis module for IS-NLI interpreter,"VI. Hungarian Computer Linguistics Conference, pp. 356-359, 2009.

[3] ED Liddy, "Natural language processing," *Library and Information Science Commons*, 2001.

[4] D. Hiemstra and F. d. Jong, "Statistical Language Models and Information Retrieval: natural language processing really meets retrieval," *Center for Telematics and Information Technology*, 2001.

[5] Duriau, VJ, RK Reger and MD Pfarrer, "A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements.," *Organizational research methods*, pp. 5-34, 2007.

[6] D. a. JHK Knoke, Network analysis, Beverly Hills: SAGE Publishing, 1982.

[7] R. Axelrod, "The cognitive maps of political elites," *Structure of decision*, Princeton, Princeton University Press, 1976.

[8] Huff, AS, V. Narapareddy and KE Fletcher, Mapping strategic thought, 1990.

[9] Carley, Kathleen and M. Palmquist, "Extracting, representing, and analyzing mental models," *Social forces*, pp. 601-636, 1992.

[10] J. a. BMS Grimmer, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political analysis*, Oxford University Press, 2013, pp. 267-297.

[11] Zeng, Zhiqiang, H. Shi, Y. Wu and Z. Hong, "Survey of natural language processing techniques in bioinformatics," *Computational and mathematical methods in medicine*, 2015.

[12] C. Noam, Syntactic structures, De Gruyter Mouton, 2009.

[13] C. Kenneth and RL Mercer, "Introduction to the special issue on computational linguistics using large corpora," *Computational linguistics*, 1993.

[14] C. Michael, Three generative, lexicalised models for statistical parsing, 1997.

[15] D. Li and Y. Liu, Deep learning in natural language processing, Springer, 2018.

[16] J. Brownlee, Deep Learning for Natural Language Processing, 2017.

[17] L. Hang, "Deep learning for natural language processing: advantages and challenges," *National Science Review*, 2017.

[18] J. Burstein, "Opportunities for Natural Language Processing Research in Education," *Computational Linguistics and Intelligent Text Processing*, pp. 6-29, 2009.

[19] Gy. Molnár; P. Nyírő (2016), "A gyakorlati programozás tanításának játékfejlesztésen alapuló, élménypedagógiai alapú módszerének bemutatása", Pedagógiai és szakmódszertani tanulmányok, Komárno, Szlovákia, International Research Institute, pp. 89-98, 2016.

[20] J. Francisti et al (2021). "Application Experiences Using IoT Devices in Education", *Applied Sciences*, vol 10, no 20, p. 7286, 2021.

[21] NY Habash, "Introduction to Arabic Natural Language Processing,"Synthesis Lectures on Human Language Technologies, pp. 1-187, 2010.

[22] P. Porteleki, "IoT integration in Microsoft Dynamics ERP", *Computers & Learning*, vol. 3, no. 1, pp. 1-11, 2020.

[23] E. Kocsó and M. Cserné Pekkel, "Using of Dynamic Animations to Illustrate Mathematical Theorems", *Transactions on IT and Engineering Education*, vol. 3, no. 1, pp. 1-15, Dec. 2020.

[24] M. Gaborov, "Comparative analysis of agile and traditional methodologies in IT project management", *Journal of Applied Technical and Educational Sciences*, vol. 11, no. 4, pp. 1-24, ArtNo: 279, Dec. 2021.

[25] J. Katona, "Analyse the Readability of LINQ Code using an Eye-Tracking-based Evaluation", *Acta Polytech. Hung*, vol. 18, pp. 193–215, 2021.

[26] J. Katona, "Clean and dirty code comprehension by eye-tracking based evaluation using GP3 eye tracker", *Acta Polytechnica Hungarica*, vol. 18, no. 1, pp. 79–99, 2021.

[27] J. Katona, "Measuring Cognition Load Using Eye-Tracking Parameters Based on Algorithm Description Tools", *Sensors*, vol. 22, no. 3, p. 912, 2022.

[28] Zs. Riczu, Zs. Krutilla, "The impact of optical character recognition artificial intelligence on the labour market." *International Journal of Engineering and Management Sciences* 6.4, 2021.

[29] Adam, E. E. B. (2020). "Deep learning based NLP techniques in text to speech synthesis for communication recognition." *Journal of Soft Computing Paradigm (JSCP)*, 2(04), 209-215.