Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/110

Natural Language Processing: State-of-the-Art

[1]Aryan Tuteja, [2]Priya Matta, [3]Ms. Akriti Dhoundiyal

[1 2]Department of Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, India.
[3]Assistant Professor, Media and Mass Communication Department, Graphic Era Hill University, Dehradun, Uttarakhand,

ABSTRACT

Natural Language Processing refers to both interaction between computer and human and integration of artificial intelligence into field of linguistics. It enables the human beings to interact with machine using natural languages. Natural Language Processing is a key skill in most of the systems that deal with human–computer interaction. As more and more artificial intelligence is getting embedded into the machinery, it has become a point of attraction to make the machine directly understand the instructions given by human in a simple, more precisely saying natural languages. Natural Language Processing, when integrated with computational modelling, results into the term Computational Linguistics. Computational Linguistics needs the involvement of both the language experts as well as the computational experts (computer scientists). This paper aims to provide the better awareness about the fact how these two fields, Computational Linguistics and Natural language Processing are interrelated to each other. In this research paper we have tried to basically deliberate the past success in the research and development in the field of Natural Language Processing & current status of Natural Language Processing. The other major focus is on new roads towards the research and challenges in the same field.

Keywords: NLP, Computational Linguistics, Artificial Intelligence, HCI, Natural Languages

INTRODUCTION

Two entities or the individuals communicate using different ways, including spoken words, body gestures, written or printed notes. Language is one of the most imperative, common, persistent and easy way of communication.

“The entity that sends the message is called sender and the other party is receiver. When either of them, sender or the receiver is a computing device, or the other one is human being, Natural Language Processing (NLP) comes into the scene.”[1] “NLP is the group of methods, procedures and algorithms whose primary objective is to make the computer systems, more accurately the computing devices, to understand natural languages.”[2] “A natural language is one that is spoken, written or presented by a human being during communication. The input for the communication can be manual or through some input device, like mike or keyboard.”[3] Some researchers defined NLP in their words.

“Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things."[4]

"Natural language processing (NLP) refers to computer systems that analyze, attempt to understand, or produce one or more human languages."[5]

Natural language processing (NLP) is the study of mathematical and computational modeling of various aspects of language and the development of a wide range of systems.

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or application. This research Paper begins with the categorization of NLP on the basis of different criterion in section 2. Section 3 continues with the possible ways in which one can correlate the two fields CL and NLP. This paper further details the motivation towards the current research on NLP in section 4. The history of NLP and its background is discussed in section 5. The primary goal of this research paper is to enlighten the previous work done in this field. This is discussed in section 6. Section 7 gives an overview of the ideas that contains the actual research challenges in the field. The challenges ahead for work in NLP is Finally, our paper leads to our conclusion in section 8.

CATEGORISATION

Basically, NLP is the culmination of both the streams namely Language processing and Language generation. If computing device is playing the role of reader/listener, then this subdivision is termed as Language Processing. In the same way, if computing device is playing the role of writer/speaker, then this subdivision is termed as Language Generation.

The other focus is on the medium of communication, i.e., whether the expert is working on language (written or printed material) or the speech (spoken or recorded words). These two are termed as Language understanding & Speech understanding respectively. Later one is sometimes referred as speech recognition.

Language Understanding may deal with the material having free flow handwriting, cursive writing, or a well-defined patterned writing. It may also comprise the study of typed documents. Similarly, speech understanding may deal with randomly spoken words which can be given as commands to the computing device and it may also interpret the recorded voice. The source of these speeches may be different kind of people, communicating in variety of accents and definitely with different speeds.

The other significant and essential issue in both the cases, either it is language understanding or the speech understanding; is the wide range of languages used for communication by people across the globe. It means if one is developing a NLP system it should be a multilingual one.

HOW ARE THEY INTER RELATED: NLP & CL

While talking about two streams, it is very common to say they are similar or different. But sometimes it becomes tough to say that how they are interrelated; same is the case with Natural

Language Processing (NLP) and Computational Linguistics (CL).
NLP & CL are not two distinct fields; rather NLP is the subset of CL, and therefore CL includes NLP. CL as an area is widely spread area, while NLP forms a part of it.

The other way to distinguish CL and NLP is NLP is engineering discipline and CL is scientific discipline. Computational linguistics forms a scientific discipline that analyses and interprets linguistic processes from a computational viewpoint.

Natural language processing forms an engineering discipline that uses computers or computing devices or computational procedures to work with a language.

NLP majorly focusses on NL parsers, Machine Translations, speech recognition, document clustering, document summarization etc. while CL includes a lot more like language comprehension, language production, language acquisition, information retrieval, information extraction and text mining.

MOTIVATION

Present day scenario is becoming more and more demanding towards user-friendly computation than ever. Every application, which desires human intervention, needs commands and instructions to be given to the computing device via some mode of communication. And therefore, one of the most noteworthy issues in computing environment is to comprehend communication between human being and computing device. It may include the way a human being presents his/her instruction, command or message to the device or vice versa. Such an understandable communication can be provisioned by implementing some computational techniques to analyse, understand the natural language by computing device or generate a language that seems to be natural for human beings. To fulfil such demands, researchers are putting their efforts in the field of NLP. The other factor that drives researchers towards the field of NLP is incorporation of artificial intelligence into the computing devices. The fundamental target is to make the machine artificially intelligent and that too with the natural languages. This may also be thought of as a technological motivation that comprises of the requirement of Intelligent Computing Devices or systems. Such systems may include Machine Translation System, speech Recognition systems, text recognition & analysis systems and systems providing natural language interfaces to the different databases. One more fact is that the printed or handwritten documents are difficult to handle and manage, in comparison to the digital documents. One can keep digital documents more secure than printed material. It is often easier to keep the backups of digital documents. Therefore, conversion of handwritten documents or printed material into digital data decreases the probability of loss of data. These reasons make this conversion a point of attraction for researcher.

HISTORY AND BACKGROUND

The beginning of NLP is formally seen in the 1950s, although efforts can be seen from earlier days. The term NLP came into being in the year 1950 when Alan Turing published his well-known article “Computing Machinery and Intelligence”. This article defined a criterion of intelligence, which is known as Turing Test now-a-days. “This criterion depends on the ability of a computer program to impersonate a human in a real-time written conversation with a human judge, sufficiently well that the judge is unable to distinguish reliably — on the basis of the conversational content alone — between the program and a real human.” This imitation of a human by computing device or algorithm simply leads towards NLP. In 1957, with the advent of “Universal Grammar”, Noam Chomsky laid a revolutionary step in the field of Linguistics.

Certain remarkably effective NLP systems evolved in 1960s. SHRDLU and ELIZA were two famous initial Natural Language systems Some other efforts were done by programmers in 1970s. They started writing ‘conceptual ontologies’, which can translate some sort of real-world information into computer understandable data. Some of the notably success in this effort are MARGIE (Schank, 1975), SAM (Cullingford, 1978), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), and Plot Units (Lehnert 1981). In 1980s, the apparent purpose of the NLP system was considered as base. Instead of using complex set of handwritten rules, research turns toward Machine Translation. Machine translation actually refers to the development of Machine learning Algorithms. Some work has been done to deal with existing multilingual textual corpora. Current research concentrates on the further categorisation of these algorithms as unsupervised and semi-supervised learning algorithms.

STATE OF THE ART AND REALATED WORKS

Guided by a parse tree, Xie et al. [1] proposed a neural architecture where candidate response and their representation learning are constituent centric. By using this architecture, the search pool of candidate answers is reduced while preserving the hierarchical, compositional and syntactic structure among constituents. The model delivers state-of-the-art performance using SQuAD.

For handling tasks involving answering questions on biomedicine, Wiese et al. [2] presented a deep learning method based on domain adaptation methods. Their model outperformed state-of-art approaches in several domains and exhibited the state-of-art on biomedical question replies.

A rational recurrent neural network was presented by Santoro et al. [3] with the ability to learn by classifying data and carry out complicated reasoning based on interactions between segmented data. To manage these interactions, they made use of the relational memory core. Finally, three separate datasets were used to assess the model's ability to model language (GigaWord, Project Gutenberg, and WikiText-103). Additionally, they compared the effectiveness of their model to established methods for handling relational reasoning on compartmentalised information. The outcomes obtained using RMC demonstrate enhanced performance of the model.

To handle the granularity at the character and word levels, Merity et al. [86] expanded traditional word-level language models based on Quasi-Recurrent Neural Network and LSTM. Using the Penn Treebank dataset for character-level modelling and WikiText-103 for word-level modelling, they tuned the modelling parameters. Their model outperformed the state-of-art in both instances. A Memory-Augmented-Machine-Comprehension-Network (MAMCN) created by Seunghak et al. [5] to handle dependencies faced in reading comprehension. The model using TriviaQA and QUASAR-T datasets, and paragraph-level using SQuAD datasets achieved state-of-the-art performance on document-level.

RESEARCH CHALLENGES & FUTURE SCOPE

Data in natural language is directly proportional to the size of the vocabulary that is with increase in vocabulary the data also increases which nearly never comes to an end. How to deal with this kind of problem poses a significant challenge to deep learning. The applications of NLP have been growing rapidly, and with these new challenges are discovered despite a lot of work done in the recent past. Challenges like Contextual words and phrases in the language were same words or phrases can have a different meanings or tone in a sentence which makes it a challenging task for AI, comparatively easy for the humans to understand. Further in language, Homonyms, the words have different meanings but are to be pronounced the same are also a challenge for speech-to-text applications because they aren't able to distinguish the tone. Sentences using sarcasm and irony sometimes may be understood in the opposite way by the humans, and so designing models to deal with such sentences is a really challenging task in NLP. There is no question that models for NLP for the most popular commonly used languages have been performing extremely well and are developing day by day, but there is still a need for models for all people rather than specific understanding of a certain language and technology. Sharifirad and Matwin [6] provide a classification of distinct online harassment types and issues. Complex tasks such as multi-turn dialogue in natural language processing, is not be easily realized with deep learning alone.It involves language understanding, generation, management, knowledge base and inference. Reinforcement learning can play a critical role in dialogue management that can be formalized as a sequential decision process. Obviously, reinforcement learning could be potentially useful for the task, which is beyond deep learning itself.

CONCLUSION

This paper was developed with three goals in mind. The first goal provides insights into the various significant NLP terms that might be valuable for readers who want to work and explore in the domain of NLP and tasks pertaining its applications. The another goal of this study is to discuss the history, applications, and current breakthroughs in the field of natural language processing (NLP). The third goal is to discuss state-of-the-art NLP research. The paper also focused the relevant tasks in the current literature, decisions, and some of the major applications and initiatives in NLP. Significant work on NLP is also present in various literature surveys [7] focusing on one domain such as usage of deep-learning techniques in NLP, techniques used for email spam filtering, medication safety, management research, intrusion detection etc, still there is not much work on regional languages, which can be the focus of future research. To summarise, there are still several unresolved difficulties in deep learning for natural language processing. Deep learning, when integrated with other technologies (reinforcement learning, inference, knowledge), has the potential to push the field's boundaries even farther.

REFERENCES

1. Xie P, Xing E (2017) A constituent-centric neural architecture for reading comprehension. In proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 1: long papers) (pp. 1405-1414)
2. Wiese G, Weissenborn D, Neves M (2017) Neural domain adaptation for biomedical question answering. *arXiv preprint arXiv:1706.03610*
3. Santoro A, Faulkner R, Raposo D, Rae J, Chrzanowski M, Weber T, ..., Lillicrap T (2018) Relational recurrent neural networks. Adv Neural Inf Proces Syst, 31

4. Merity S, Keskar NS, Socher R (2018) An analysis of neural language modeling at multiple scales. arXiv preprint arXiv:1803.08240
5. Yu S, et al. (2018) "A multi-stage memory augmented neural network for machine reading comprehension." Proceedings of the workshop on machine reading for question answering
6. Sharifirad S, Matwin S, (2019) When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NLP. arXiv preprint arXiv:1902.10584
7. Wong A, Plasek JM, Montecalvo SP, Zhou L (2018) Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy 38(8):822–841