International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(1), pp.156-158
DOI: http://dx.doi.org/10.21172/1.101.27
e-ISSN:2278-621X

NATURAL LANGUAGE PROCESSING: A SURVEY

Aditya Gupta[1], Kalpana Dwivedi[2]

Abstract- Natural Language Processing (NLP) is a branch of computer science and sub-branch of artificial intelligence. Natural Language Processing is use to build a machine that interacts with humans in the form of natural language. Natural Language processing is the ability of computational technologies and computational linguistic to process human natural language. Natural Language Processing is an area that explores how computers can be used to understand natural language text or speech. Building a universal machine translation system which can convert one specific language to another specific language, it is a long term goal of Natural Language Processing based system. The application of Natural Language Processing are used in following expert systems such as speech recognition system, translation from one specific language to another specific language, text summarization, sentiment analysis, chat - bots, text classification.

Keywords – Natural Language Processing, Corpus, Machine Learning, Artificial Intelligence, Deep Learning

1. INTRODUCTION

Today in the era of digitalization, our maximum data is created unstructured i.e. audio, video, our social footprints, the data generated from conversation between two representatives and tons of text processed in different sectors.

NLP is a technology of computers is used to understand human speech as it spoken. NLP is a field of computer science, artificial intelligence and computational linguistics. It can be defined as the automatic processing of human natural language.

As a human being, we express our thoughts or feelings via a language. Whatever you speak, read, or listen to is mostly in the form of natural language so it is commonly expressed as natural language.

For example: Whatever we speak, listen and write in our daily life is also in the form of natural language. Our WhatsApp, Facebook, Hike any social networking conversation are also considered a form of natural language. NLP products from the world's top tech companies, such as Google Assistant from Google, Cortana from Microsoft, Siri speech assistance from Apple, Bixby from Samsung, Alexa from Amazon and so on. Since the invention of NLP, the keyboard has been use for the human-computer interface. But that's changing today because of voice recognition via virtual assistants which responds to vocal prompts to do things like finding a store, getting direction to a home, turning on/off lights etc.

From these examples and more, it's clear that NLP has a very important role in new machine – human interfaces and will be an essential tool for leading the future.

2. STRUCTURAL REPRESENTATION OF THE CONCEPT

The basic model, which shows how an expert system can be built for NLP applications.

The below fig. shows the development life cycle:

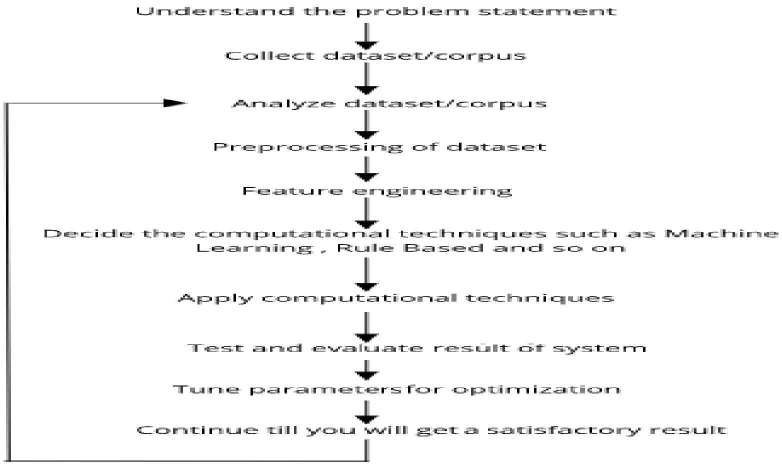
Understand the problem statement
Collect dataset/corpus
Analyze dataset/corpus
Preprocessing of dataset
Feature engineering
Decide the computational techniques such as Machine Learning , Rule Based and so on
Apply computational techniques
Test and evaluate result of system
Tune parameters for optimization
Continue till you will get a satisfactory result

Fig 1. Devlopment Life Cycle

[1] Department of Computer Science Engineering, NIET, Greater Noida, Uttar Pradesh, India
[2] Department of Computer Science Engineering, NIET, Greater Noida, Uttar Pradesh, India

2.1 Corpus and Dataset –

Natural Language Processing based applications required large amount of data. According to Layman's, you can say that a large collection of data is called corpus. To develop NLP applications, we require corpus that is written or spoken natural language. NLP applications can use a single corpus or may use multiple corpus as input. With the help of corpus, we can perform analysis such as frequency distribution and so on. We can define and validate linguistics rule for NLP applications, if we are building a grammar correction system, then we need text corpus. In a corpus, the large amount of data can be in the following formats i.e. Text data (written material), Speech data (spoken material). Corpus is the basic building block of NLP. There are basically three steps for preparing dataset i.e. Selecting data, Preprocessing data, Transforming data.

2.2 Preprocessing –

From the raw data we will preprocess the text and identify the sentences.

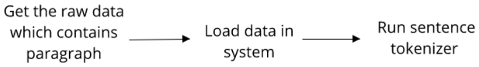

Fig 2. Process of handling corpus-raw text

Sentence tokenization is the process of identifying the boundary of the sentences. It is also called sentence boundary detection. This process identifies starting and ending point of sentence. But there are some challenge which involves some languages such as Urdu, Hebrew, Arabic and so on, they are difficult in terms of deciding the ending of sentences and find out tokens from the sentences.

2.3 Feature Engineering –

Feature engineering has very big role in developing of NLP applications. Features are the input parameter for machine learning (ML) algorithms, the ML generate output based on input features.

Feature engineering is a method which generates the beast possible feature and choosing the best algorithm to develop NLP applications. We have lots of raw data in natural language that computer can't understand, and algorithms don't have the ability to accept the raw natural language and generate expected output.

2.4 Machine Learning for NLP –

Machine Learning and NLP are area of an artificial intelligence which have ability to solve problems by statistical techniques. These techniques are applied to a wide variety of problems. For the implementation of ML techniques to solve NLP problems, we need to convert the unstructured text into a structured format. ML has ability to learn by providing some samples for e.g. if you want to recognize the valid license plates, in traditional programming you need to write code for the shape of the license plate, what color it should be, what fonts are used and so on. These coding steps are too lengthy. Using ML, we will provide some example license plates to machine and the machine will learn the steps so that it can recognize the new valid license plate.

2.5 Deep Learning for NLP

From the last four to five years, neural networks and deep learning techniques have been creating a lot of buzz in the AI area. Tech giants such as Google, Amazon, Apple and so on. Spend a lot of time and effort to create solutions for real-life problems.

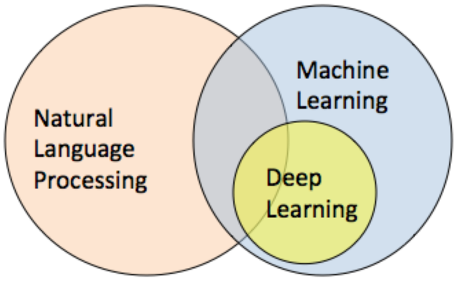

Fig 3. Relation between NLP, ML and Deep Learning

The art of understanding language involves understanding humor, sarcasm, subconscious bias in text, etc. Once we understand that is means to be sarcastic we can encode it into a ML algorithm to automatically discover similar patterns for us statistically.

The early era of NLP is based on the rule-based system, and for many applications because we did not have huge amounts of data. We know language are complex things to deal with and sometimes we also don’t know how to solve certain NLP problems. The reason behind this is that there are many languages in the world that have their own syntactic structure, these reasons and factors lead us in the direction of the usage of DL for NLP applications. There are other more capabilities that DL provides us such as interpretability, modularity, transferability, latency, and security.

3. CONCLUSION

To develop more effective relationship between humans and machines, then we really need NLP systems that can understand the context of human natural language and react and behave more like humans, humanoids robots are the best the best application to describe NLP system. According to me, systems should react more like humans do. Machine reactions should match with real human behavior. After analyzing situations, machine should react the same as human react.

4. FUTURE SCOPE

The NLP and their applications are helps machine to understand the emotions of sentence. It provides advance interface between humans and machines. In the future we can talk to machines like humans do, even more machines also get emotions and take better understanding of human’s emotions.

“A computer would deserve to be called intelligent if it could deceive a human into believing that it was human.” -Alan Turing

5. REFERENCES

[1] Ronan Collobert, Jason Weston, Michael, Koray Kavukcuoglu, Pavel Kuksa (2011). Natural Language Processing (Almost) from Scratch. Jounral of Machine Learning Research.

[2] Edward Loper, Ewan Klein, and Steven Bird (2009). Natural Language Processing with Python.

[3] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural Language Processing: State of The Art, Current Trends and Challenges.

[4] Shemtov, H. (1997). Ambiguity management in natural language generation. Stanford University.

[5] Liddy, E. D. (2001). Natural language processing.

[6] Introduction to Natural Language Processing (NLP). https://blog.algorithmia.com/introduction-natural-language-processing-nlp/ [accessed 20 March 2018].