NATURAL LANGUAGE PROCESSING

March 25, 2019

Nikita Masand
Mentor: Prof Pranav Nerurkar
Veermata Jijabai Technological Institute
Matunga, Mumbai
Information Technology
ID: 171081054

March 25, 2019

Contents

List of Figures

List of Tables

Chapter 1

Abstract

Natural Language Processing (NLP) is a way of analyzing texts by computerized means. NLP involves gathering of knowledge on how human beings understand and use language. This is done in order to develop appropriate tools and techniques which could make computer systems understand and manipulate natural languages to perform various desired tasks. This paper reviews the literature on NLP. It also covers or gives a hint about the history of NLP. It is based on document analysis. This research paper could be beneficial to those who wish to study and learn about NLP. Keywords: NLP, machine translation, machine learning, computational techniques, linguists

graphicx

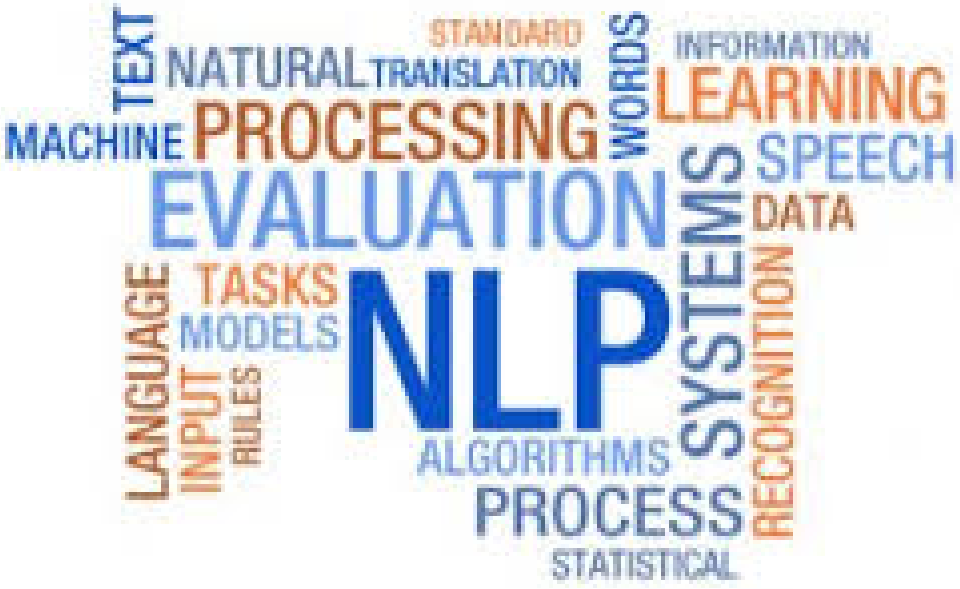

Figure 1.1: NLP

hackernoon.com

Chapter 2

What is NLP?

Artificial intelligence (AI) technologies such as machine learning (ML) and deep learning (DL) are dazzling in and of themselves, but believe it or not, leveraged in isolation, they are limited in their potential. These technologies do not interpret data by themselves: they are tied either to deterministic, hard coded software programs created by humans or they are linked to a form of artificial intelligence that can interpret human language into a form ML and DL algorithms can understand. The umbrella term for this gateway AI technology is natural language processing (NLP).

Various researchers have explained Natural Language Processing (NLP) as 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 The term NLP is normally used to describe the function of software or hardware components in a computer system which analyze or synthesize spoken or written language.

2.1 Understanding Terms in a NLP Project

The research and development in NLP over the last sixty years as stated by Church and Rau can be categorized into the following five areas:
Natural Language Understanding
Natural Language Generation
Speech or Voice recognition
Machine Translation
Spelling Correction and Grammar Checking

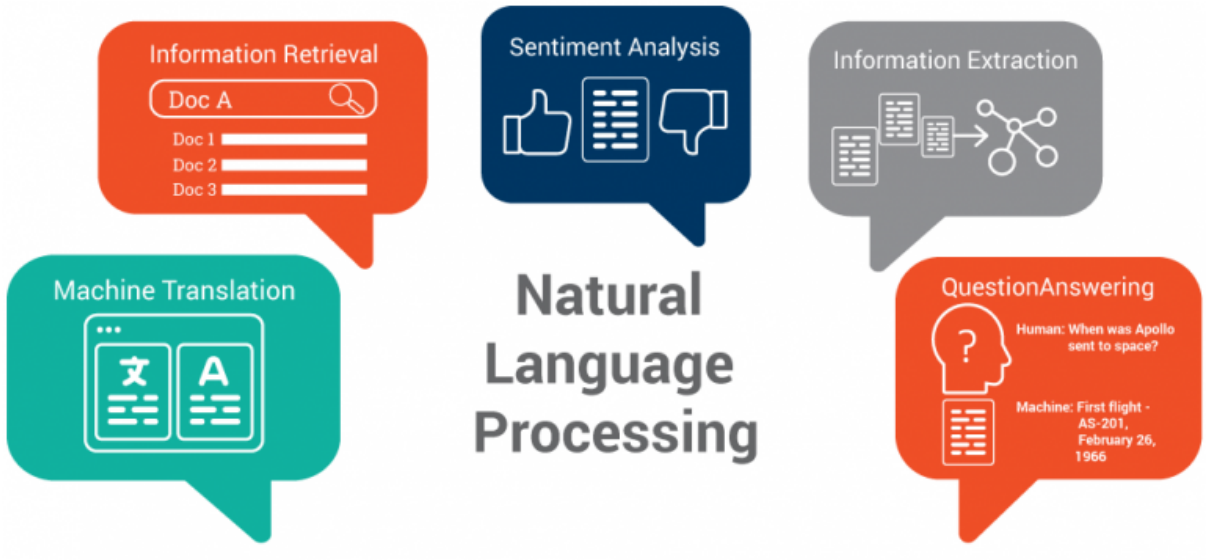

Figure 2.1: Process

https://miro.medium.com/max/1036

term	definition
Lemmatization	grouping together the inflected forms of word
Stemming	taking list of common prefixes and suffixes can be found word
Co-reference resolution	words used to refer to the same objects
nltk	toolkit NLP libraries containing packages

Table 2.1: Concepts in NLP

2.1.1 Tokenization

Tokenization is a process to split longer strings into smaller pieces.Large documents can be tokenized into paragraphs,Paragraphs can be tokenized into sentences and sentences can be tokenized into phrases,words or letters.

2.1.2 Stemming and Lemmatization

Stemming is a process to eliminate affixes(prefix,suffix,infix,circumfix) from a word in order to obtain a word stem or root word.

going -> go , happily -> happy , am/are/is -> be.

A common term associated with stemming is Lemmatization. There is a slight differ-

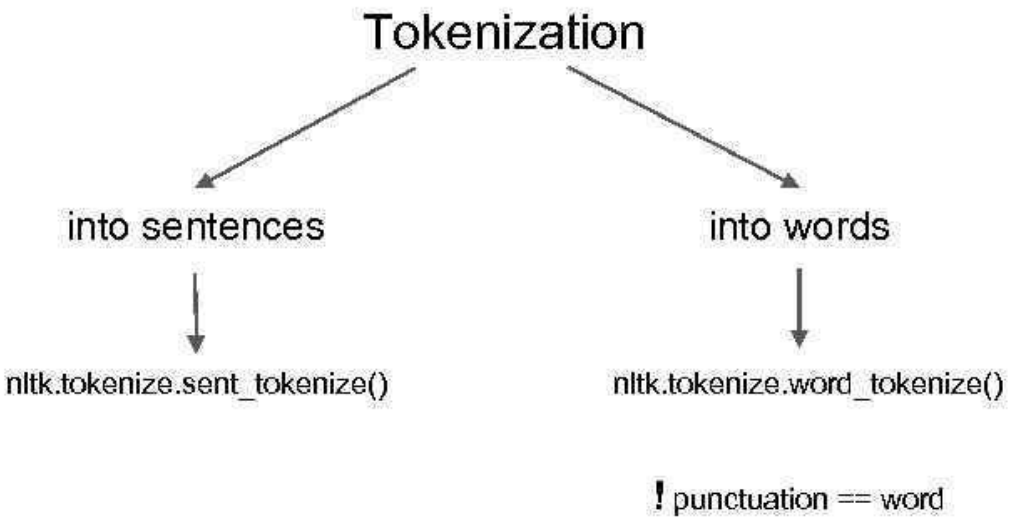

Figure 2.2: tokenizing

https://cdn-images-1.medium.com

ence between stemming and Lemmatization

Stemming cuts off the end or beginning of the word,taking into account a list of common prefixes and suffixes

Form : Studies Suffix : es Stem : Studi

Form : Studying Suffix : ing Stem : Study

Lemmatization takes into consideration morphological analysis of the words.

Form : Studies Lemma : Study

Form : Studying Lemma : Study

Lemmatization definitely has an edge over stemming but building a Stemmer is far easy then the latter as deep linguistic knowledge is required to look for the proper form of word.

Chapter 3

Applications of NLP

- Text Classification and Categorization
  Text classification is an essential part in many applications, such as web searching, information filtering, language identification, readability assessment, and sentiment analysis. Neural networks are actively used for these tasks.

- Named Entity Recognition (NER)
  The main task of named entity recognition (NER) is to classify named entities, such as Guido van Rossum, Microsoft, London, etc., into predefined categories like persons, organizations, locations, time, dates, and so on. Many NER systems were already created, and the best of them use neural networks

- Part-of-Speech Tagging
  Part-of-speech (POS) tagging has many applications including parsing, text-to-speech conversion, information extraction, and so on. In the work, Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network a recurrent neural network with word embedding for part-of-speech (POS) tagging task is presented

- Semantic Parsing and Question Answering
  Question Answering systems automatically answer different types of questions asked in natural languages including definition questions, biographical questions, multi-lingual questions, and so on. Neural networks usage makes it possible to develop

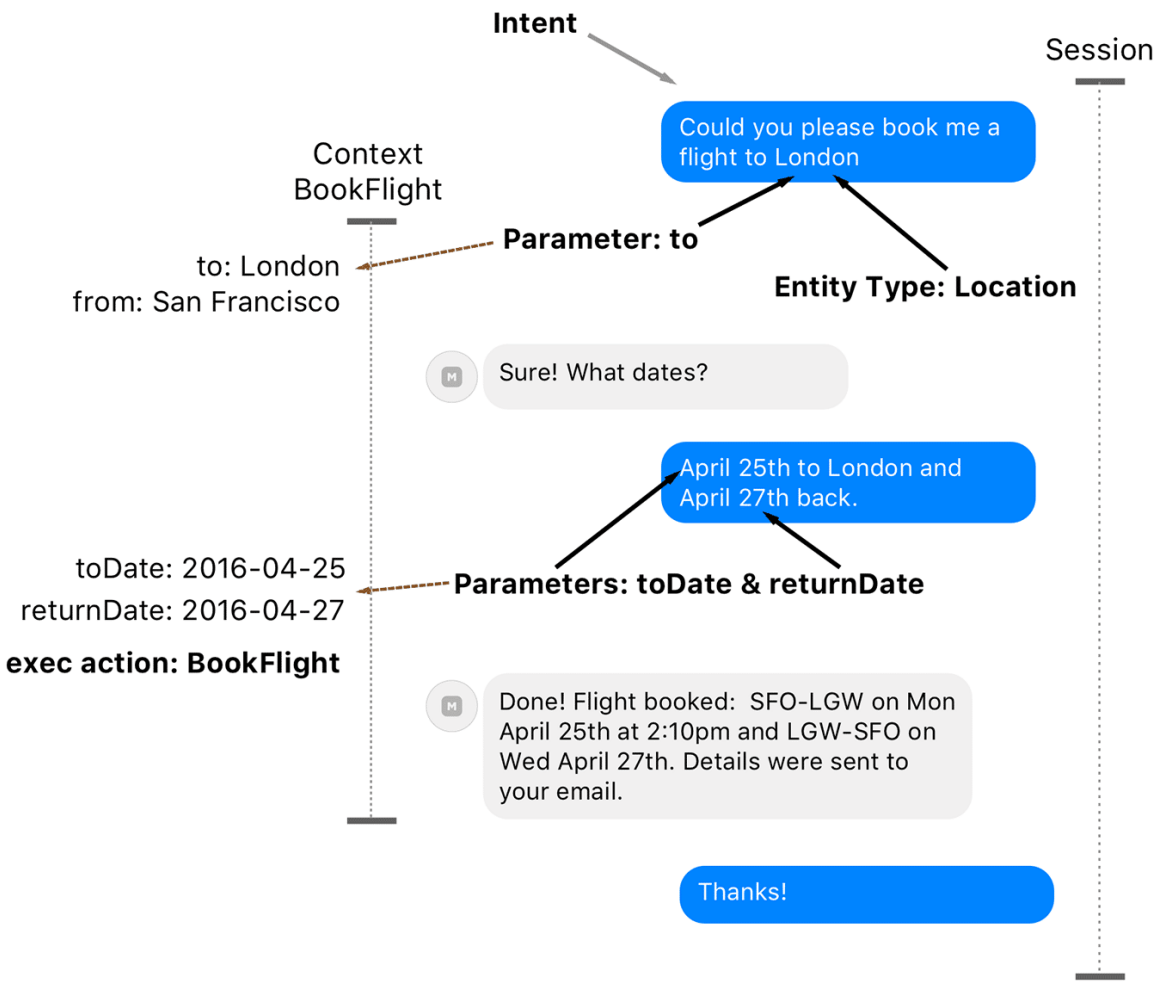

Figure 3.1: Working

http://www.nlp.com/what-is-nlp/

high performing question answering systems.

- Paraphrase Detection
  Paraphrase detection determines whether two sentences have the same meaning. This task is especially important for question answering systems since there are many ways to ask the same question.Detecting Semantically Equivalent Questions in Online User Forums suggests a method for identifying semantically equivalent questions based on a convolutional neural network.

- Language Generation and Multi-document Summarization
  Natural language generation has many applications such as automated writing of reports, generating texts based on analysis of retail sales data, summarizing electronic medical records, producing textual weather forecasts from weather data, and even producing jokes

- Machine Translation
  The purpose of Neural-based Machine Translation for Medical Text Domain study is to inspect the effects of different training methods on a Polish-English machine translation system used for medical data. To train neural and statistical network-based translation systems The European Medicines Agency parallel text corpus was used.

- Character Recognition
  The article Character Recognition Using Neural Network presents a method for the recognition of handwritten characters.

Chapter 4

Conclusion

As a computerized approach of analyzing text, NLP is continually striving forward. Researchers are continually trying to gather knowledge on how human beings understand and use various languages. This aid in the development of appropriate tools and techniques which make computer systems understand and manipulate natural languages to perform the various tasks. Technologies, such as string matching, keyword search, glossary lookup are now on the past as, to more forward looking technologies such as grammar checkers, conceptual search, event extraction, interlingual on going and striving forward

[1]E.D. Liddy, Natural Language Processing, 2001 \\
[2] S. Vijayarani1, J. Ilamathi and Nithya, âĂIJPreprocessing Techniques for
\\International Journal \\of Computer Science & Communication Networks, Vol.5
[3] J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litm
âĂIJRecognizing and Organizing Opinions Expressed in the World PressâĂİ. In \\Proceed
[4] P. Jackson and I. Moulinier,âĂİNatural Language Processing for Online Applicatio
[5] R. Bose. âĂIJNatural language processing: Current state and future direc