



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Natural Language Processing Basic Tasks for Text Mining

Dhanashree K. Thakur

Department of Computer Engineering,

Dr. Babasaheb Ambedkar Technological University,

Lonere, Mangaon, 402 103, Dist: Raigad, Maharashtra, India

Abstract: *Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. NLP is among the most heavily research areas in computer science today. Historically, NLP has been concerned with developing machine translation and natural language generation systems. For processing natural language, we have to go through several phases such as from meaning of each word with its utterance to the meaning of whole sentence to paragraph. This paper contain different phases, tasks, applications and challenges of NLP and there use in real life.*

I. Introduction

Language is the primary means of communication used by humans. It is a tool we use to express the greater part of our ideas and emotions. It shapes thoughts, has a structure, and carries meaning. Learning new concepts and expressing ideas through them is so natural. But there must be some kind of representation in our mind, of the content of language. Natural languages are languages used by human to communicate with each other like English, Hindi, and Turkish other than computer languages such as C, C++, java, etc.

The basic goal of Natural language Processing is to enable a person to communicate with a computer in a language that they use in their everyday life.

Natural Language Processing (NLP) means processing of natural language text using computational models which uses different algorithms to solve different tasks of that language which is applied to computer machine or useful to make software which behaves with human like human being. It means that we can use these tools or machines by giving natural language commands. So, we can say that it is a field of computer science, human-computer interaction, Artificial Intelligence, Linguistics, Text Mining, etc.

II. Phases of NLP

It is not an easy task to teach a person or computer a natural language. The main problems are syntax, and understanding context to determine the meaning of a word. To interpret even simple phrases requires a vast amount of knowledge. In NLP, we have to consider text form of the Language and the content of it as Knowledge. Hence, to process a language means to process the content of it. As computer is not able to understand natural language, methods are developed to map its content in formal language. The Language and speech community on the other hand, considers a Language as a set of sounds, conveys meaning to listener. Therefore, to process natural language, we have to go through number of phases. Each phase contain particular ambiguity which is nothing but one thing has two meanings and in our NLP task, computer must understand exact meaning of that thing otherwise whole meaning of that paragraph may be change.

A. Phonetics and Phonology

Phonetics deals with the physical building blocks of a language sound system. E.g. sounds of 'k', 't' and 'e' in 'kite'. Phonology deals with organization of speech sounds within a language. E.g. (1) different k sounds in kite verses coat. (2) different t and p sounds in top verses pot.

B. Lexical Analysis

The simplest level which involve analysis of words. Word-level processing requires morphological Knowledge, i.e. Knowledge about structure and formation of words from basic units called morphemes. The rules for forming words from morphemes are language specific. It is also concerned with derivation of new words from existing ones, e.g. light-house (formed from light & house).

C. Syntactic Analysis

The next level which considers a sequence of words as a unit, usually a sentence, and its structure. It decomposes the sentence into word and identifies how they relate to each other. Not every sequence of words results in a sentence. For example, 'I went to the market' is valid sentence while 'went the market to' is not. Similarly, 'She is going to the market' is valid, but 'She are going to the market' is not. Thus, this level of analysis required detailed knowledge about rules of grammar.

D. Semantic Analysis

Semantic is associated with the meaning of language. Semantic analysis map natural language sentences into some representation of meaning. Defining meaning of components is difficult, as grammatically valid sentences can be meaningless. Example is, 'Colorless green ideas sleep furiously'. The sentence is syntactically correct but

semantically anomalous. Unfortunately, many words have several meanings, for example, the word 'diamond' might have the following set of meanings:

- (1) A geometrical shape with four equal sides.
- (2) A baseball field
- (3) An extremely hard and valuable gemstone

To select the correct meaning for the word 'diamond' in the sentence "John saw Susan diamond shimmering from across the room." It is necessary to know that neither geometrical shapes nor baseball fields shimmer, whereas gemstones do. The concept of meaning is quite broad. A word can have number of possible meanings associated with it. But in a given context, only one of the meanings participates. The process of determining the correct meaning of an individual word is call word sense disambiguation or lexical disambiguation. It is done by associating, with each word in the lexicon, information about the contexts in which each of the words senses may appear.

E. Discourse Analysis

It is higher level of analysis attempt to interpret the structure and meaning of even larger units, e.g. at the paragraph and document level, in terms of words, phrases, clusters and sentences. It requires resolution of anaphoric reference and identification of discourse structure. That is knowledge of how the meaning of sentence is determined by preceding sentence. E.g. how a pronoun refers to the preceding noun and how to determine function of a sentence in a text. There are many important relationships that may hold between phrases and parts of their discourse context. Consider "'Bill had a red balloon. John wanted it.'" The word 'it' should be identified as referring to the red balloon. References such as this are call anaphoric or anaphora.

F. Pragmatic Analysis

The highest level of processing is pragmatic analysis which deals with the purposeful use of sentences in situations. It requires knowledge of the world, i.e. knowledge that extends beyond the contents of the text. This is an additional stage of analysis concerned with the pragmatic use of the language. This is important in the understanding of texts and dialogues. For example, suppose teacher giving a lecture and forgot to bring book. So, he asked to one student to go in his cabin and to see if his book is there or not. Student went and came without bringing book and say that his book is there. It means that student cannot understand the situation.

III. NLP Activities

To natural language processing contain different basic tasks or activities required to process any language which are done by computer programming languages such as java, python, etc. These basic tasks of NLP are explained here.

A. Tokenizer

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. For a language like English, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese and Japanese do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language.

E.g. Suppose a sentence is "Sony sang a song." Then tokens are 'Sony', 'sang', 'a', 'song' '.' and 3 space tokens.

B. Sentence Splitter

It segments the text into sentences. It assigns annotations of type split to sentence boundaries in text based on punctuation. A question mark or exclamation mark always ends a sentence. A period followed by an upper-case letter generally ends a sentence, but there are a number of exceptions. For example, if the period is part of an abbreviated title (e.g. "Mr.", "Gen."), it does not end of the sentence. A period following a single capitalized letter is assumed to be a person's initial, and is not considered the end of a sentence. Sentence Detector can detect that a punctuation character marks at the end of a sentence or not. The sample text below should be segmented into its sentences. Ex. Suppose text is "My friends are Sony and Sangita. Sony sang a song. She is very happy today." Then output is in three separate strings:

"My friends are Sony and Sangita."

"Sony sang a song."

"She is very happy today."

C. POS Tagger

It assigns parts of speech such as noun, verb, pronoun, preposition, adverb, adjective, etc. to each word in a sentence. The sequence of words in natural language sentence and specified tag sets is input to tagging algorithm. Number of tags in a tag set varies substantially. The Penn Treebank tag set contains 45 tags while C7 uses 164. Output is single best POS tag for each word.

Ex. Suppose input sentence is "Sandesh sang a song."

Then output is:

Sandesh_NP sang_VBD a_DT song_NN.

Here NP for proper noun, VBD for verb in past tense, DT for determiner, NN for noun [3].

D. Named Entity Recognition

Named Entity Recognition (NER) is a technology for recognizing proper nouns (entities) in text and associating them with the appropriate types. Common types in NER systems are location, person name, date, city, organization, etc. It uses Dictionary, lexicon or Gazetteer. E.g. Suppose input sentence is "Lalita sang a song. She is friend of Sangita. She lives in Mumbai." Then output is Lalita-person, Sangita-person, Mumbai-city.

E. Coreference Resolution

In linguistics, coreference (sometimes written co-reference) occurs when two or more expressions in a text refer to the same person or thing; they have the same referent. Coreference is the main concept underlying binding phenomena in the field of syntax.

The theory of binding explores the syntactic relationship that exists between coreferential expressions in sentences and texts. E.g. "Saniya and I bought a camera. We like capturing nature scenes." Here, 'We' refers to 'I' and 'Saniya'. Another E.g. is: "Neha bought a printer. She is printing now." Here, 'She' refers to 'Neha' not a 'printer'. Another E.g. is 'Neha bought a printer. It is printing now.' Here, 'It' refers to 'printer'.

F. Chunker

It divides a text in syntactically correlated parts of words, like noun groups, verb groups, but does not specify their internal structure nor their role in the main sentence. Chunking is an alternative to parsing that provides a partial syntactic structure of a sentence, with a limited tree depth, as opposed to parsing.

E.g. Suppose input sentence is "Amey, Soniya and Tejas sang a song.", then chunking is:

[NP Amey, Soniya and Tejas] [VP sang] [NP a song].

Here, NP is Noun Phrase and VP is Verb Phrase.

G. Parser

A parser is a software component that takes input data (frequently text) and builds a data structure often some kind of parse tree, abstract syntax tree or other hierarchical structure giving a structural representation of the input, checking for correct syntax in the process. Parsing or syntactic analysis is the process of analyzing a string of symbols, either in natural language or in computer languages, according to the rules of a formal grammar. Example is, "My dog also likes eating sausage." after parsing, it gives output as:

```
(S
(NP (PRP My) (NN dog))
((ADVP (RB also))
(VP (VBZ likes)))
(VP (VBG eating)
(NP (NN sausage)
)))
```

Here, S-Sentence; PRP-Personal Pronoun; RB0-Particle; VB -Verb, gerund or present participle; VBG-Verb, 3rd person singular present; VBZ-Verb, third person singular present; NN-Noun, singular or mass.

H. Document Reset

Some tools can take large number of documents for annotations. So they may tend to employ weak referencing in the JVM, thus overriding garbage collection. Consequently there is need to explicitly delete resources mostly when using these documents. This has two consequences:

- 1) Old annotations residing in memory from a previous document may be taken into account when processing the next document.
- 2) The JVM heap will begin to fill up with old annotation sets generated from previous documents which can severely impact on performance.

Hence the purpose of the Document Reset PR is to clean out any previous annotations, essentially releasing the annotation set objects for garbage collection.

IV. Applications of NLP

There are huge amounts of data in Internet, at least 20 billion pages. Applications for processing large amounts of texts require NLP expertise. Machine translation is the first application area of NLP.

A. Speech Recognition

This is the process of mapping acoustic speech signals to a set of words. The difficulties occur due to the wide variations in pronunciation of words (e.g. dear and deer). Example is to get flight information or book a hotel over the phone.

B. Speech Synthesis

Speech Synthesis refers to the automatic production of speech (utterance of natural language sentences) from text. Such systems can read out mails on telephone or even read out storybook. In order to generate utterances, text has to be processed.

C. Information Extraction

It captures and outputs factual information contained within a document and responds to the users information need. Example is to discover names of people and event they participate in, from a document.

D. Information Retrieval

This is concerned with identifying documents relevant to user's query. It is activity of obtaining information resources relevant to an information need from a collection of information resources. Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are most visible IR applications.

E. Machine Translation

This refers to the automatic translation of text from one human language to another. In order to carry out this translation, it is necessary to have an understanding of words and phrases, grammars of the both languages involved, semantics of languages and world knowledge.

F. Natural Language Interfaces to Databases

Natural language interfaces allow querying a structured database using natural language sentences. Intelligent databases systems are systems that manage information in natural way, making that information easy to access and use. This system translates a natural language sentence to database query. Anyone can gather information from the database by using such systems.

G. Question Answering System

Question Answering system find answers to natural language questions in a text collection or database. It requires not only precise analysis of questions and portions of texts but also semantic as well as background knowledge to answer certain type of question.

H. Text Summarization

This deals with the creation of summaries of documents. It generates a short biography from one or more news articles. Automatic summarization is the process of reducing a text document with a computer program in order to create summary that retains the most important points of the original document. As the problem of information overload is grown, and as quantity of data has increased, so has interest in text summarization.

V. Some successful NLP Systems

A. ELIZA (Weizenbaum 1966)

It uses syntactic patterns to mimic human conversation with the user. Here is a sample conversation

ELIZA: Hello, I am ELIZA. How may I help you?

User: I am feeling a little bit sleepy.

ELIZA: How long have you been feeling a little bit sleepy?

User: For almost half an hour.

ELIZA: Please go on.

B. SysTran (System Translation)

The first SysTran machine translation system was developed in 1969 for Russian-English translation. It also provided the first online machine translation service called Babel Fish, which is used by AltaVista search engines for handling translation requests from users.

C. TAUM METEO

This is natural language generation system used in Canada to generate weather reports. It accepts daily weather data and generates weather reports in English and French.

D. SHRDLU (Winograd 1972)

This is natural language understanding system that simulates actions of robot in a block word domain. The user can ask the robot to manipulate the blocks, to tell the blocks configuration, and to explain its reasoning.

E. LUNAR (Woods 1977)

This was an early question answering system that answered questions about moon rock.

F. Google Translator

Now-a-days, Google translator is one of the pioneer applications supporting a number of languages to translate from one to another. Although, it has been successfully implemented for many languages, but many languages still in developing phase. The others translators e.g. Yahoo Babel Fish, SDL Free Translation, Systran Language Translation etc. support multi language translation like Danish, English, Chinese, Italian, Japanese, French, Greek, Korean etc.

VI. Conclusion

Human languages are interesting & challenging and therefore NLP is one of the most comprehensive and most challenging tasks. NLP is among the most heavily researched areas in Computer Sciences today. NLP also has been part of the ongoing research in the field of information extraction. NLP programs can carry out a number of very interesting tasks. These programs have impacts on the way we communicate. Various tools are available for different NLP tasks. It is very difficult to develop perfect NLP System. So, development in this area is helpful for individual human, different companies and research areas.

References

- [1] Tanveer Siddiqui, U. S. Tiwary, Natural Language Processing and Information Retrieval.
- [2] <http://en.wikipedia.org/wiki/Naturallanguageprocessing>
- [3] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- [4] <http://www-nlp.stanford.edu/>
- [5] http://en.wikipedia.org/wiki/TSext_mining