Chapter 5

Natural Language Processing

It’s a strange world of language in which skating on thin ice can get you into hot water.
– Franklin P. Jones

5.0 Introduction
This chapter deals with the emerging field of computational linguistics, with particular reference to NLP which is new for Gujarati language. It studies the features of natural language and artificial language and tries to find the answer to the question at what level, artificial language can match to natural language. NLP can come as an answer to this question and the first step to NLP is a morph analyzer.

5.1 Natural Language Processing
Natural Language Processing (NLP) is a branch of Artificial Intelligence and deals with the Analysis of Human Languages. Each language consists of a set of words called lexicon and a set of rules called grammar. This enables the use of words to form cognitional valid sentences. NLP studies all the components of human language with a perspective to use it in computer applications. NLP aims to gather knowledge on how humans understand and use natural language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks. The final goal of NLP is to teach the natural language to the computer. The aim is to design a program that can understand the highly ambiguous natural languages and the computer can converse with humans as, humans do it with one another.
NLP has its foundations in core linguistics and computer programming. It is one of the major branches of computational linguistics. And computational linguistics is the youngest member in the linguistics family.

5.1.1 Natural and Artificial language

Natural Language is one of the fundamental aspects of human behaviour and is a crucial component of our lives. In spoken form it serves as our primary means of coordinating our day-to-day behavior with others. In written form it serves as a long-term record of knowledge from one generation to the next.

“Language plays a great part in our life. Perhaps because of its familiarity, we rarely observe it, taking it rather for granted, as we do breathing or walking. The effects of a language are remarkable, and include much of what distinguishes man from animals, but language has no place in our educational program or in the speculations our philosophers.”

This is how Bloomfield\(^1\) (1996) describes language; we rarely observe it. And it is the one that distinguishes humans from animals. And all normal humans possess it uniformly. Language is species-specific and species-uniform.

Only human brain is competent enough to understand the enigmatic nature of language. And can produce infinite set of sentences from finite symbols/sounds. As per Chomsky\(^2\) (1980) it is Poverty of Stimulus; infinite output from finite source.

Language is the fully developed means of communication for the human race. Civilized men can convey and receive millions of messages across the globe. It is through language that they can store knowledge, transfer it to next generation and yoke the present, past and future together. It gives shape to people’s thoughts and guides and controls their entire activity. It is the carrier of civilization and culture. An entire civilization depends on language. It is the most valuable possession of the human beings.
5.1.2 Features of Natural Language

i. Arbitrariness
Language is arbitrary. There is no inherent relation between the words of language and their meanings or the ideas conveyed by them. This nature permits and presupposes change; this is the reason that we have more than one language in the world. The choice of a word selected to mean a particular thing or idea is purely arbitrary but once a word is selected for a particular referent, it comes to stay and is difficult to change as it is accepted by the whole society. In a way, it provides stability to the language.

ii. System of Systems
Though language is arbitrary, it is a system. It is not vague. Every language has its own set of rules, for different level; like phonetic and syntactic. Sounds are arranged and fixed in a systematic pattern to form meaningful words. Similarly, words are arranged in an acceptable frame; according to the rules of the language in question. A systematic arrangement of words will only result into meaningful sentences. Language is not a disorganized and chaotic combination of sounds.

iii. Displacement and Interchangeability
It is one of the important features possessed by natural languages. Humans can refer to any incident; which has occurred in past or is happening at miles away. Humans can combine present, past and future. It is this property of natural language that enables humans to create fiction and describe the past as well as possible future events. Human beings have the capacity to interchange the roles of speaker and the hearer. Hockett (1958, 76) defines interchangeability as “..any participating organism equipped for the transmission of messages in the system is also equipped to receive messages in the same system, and vice versa.”
iv. Social Learning and Cultural Transmission
Language is primarily human. It has to be acquired, and every normal human child is born with inbuilt potential. All the child needs is a social environment. He learns from his elders or peers in the society. Through language the child learns the cultural values. Language can be considered as the mirror of the society. Language is a form of social behavior and is a tool for cultural transmission.

v. Duality
Language displaces two levels of the structure. First it is made up of sound patterns and second; pattern of morphemes. For a language to function well, it needs the agreement of phonemes and morphemes. So duality in language means a combined system of sounds and the pattern of sounds or the system behind the pattern.

vi. Productivity
There are various ways of combining different units of language, and produce desired results. Human beings use few basic principles of construction and create a large number of new and varied constructions. Natural language is creative as well as productive.

vii. Specialization
Any communicative activity by itself involves triggering. Every utterance triggers some mental or physical action or behavior. An utterance leads to direct or indirect consequences. This is the specialty of human language.

These are the features of natural languages. An Artificial language has some features in common and some are not. The final aim of any language is communication so artificial language though does not share all the features of natural language; it still needs to do the same function.
5.1.3 Artificial Language

An artificial language is a language designed to communicate instructions to a machine, particularly a computer. Programming languages can be used to create programs that control the behavior of a machine and/or to express algorithms precisely.  

An artificial language is a language invented by an individual or a group of individuals for a specific purpose. It is not created naturally. It is constructed. The language of mathematics, physics, chemistry and such can be termed as AL. Computer programming languages and international auxiliary languages like Esperanto, Ido and Novial are AL along with fictional languages like Klingon and Elvish.

In the context of NLP an artificial language is a language created by humans to interact with computer. It is more like a set of codes. It is unambiguous. It has a particular set of rules. It is fully logical and rule-bound. There is no change of misleading the commands. Natural language has the quality of prevarication (deception) and human being has a limited memory, they tend to forget. Artificial language does not have the quality of prevarication and has unlimited memory, in a way, a computer is store house. It won’t forget or miss any single, minute detail. AI is absolutely context-free. It needs to be learnt with special efforts, mostly with the help of a tutor or sometimes a book. A child can not acquire it, as it acquires natural language from the society. AL is always growing. There are always new and revised languages available in the market. So if a person is learning a particular AL, within short period a new revised version is launched in the market, hence he has to keep himself regularly updating.

Anyone interested in programming or research related to computers has to compulsorily learn the AL. In a way, only software engineers can do programming. First they have to learn various ALs, keep updating and after that they can create any program or software. It is time consuming, laborious and difficult. And the main issue is that, every single person has to learn it individually. If a computer learns the natural language, then every single computer does not need to learn it individually. As humans have to do. Computers just have to transfer the program. And as a result every
human being who wants to do any sort of programming or research can easily do it. He won’t be in need of computer operators or any such guidance. Computers will be easily accessible in city as well as in villages; without the barrier of education. This will be a huge leap in the technology.

5.2 Approaches to NLP
The complexity of natural languages makes it exceedingly difficult to isolate factors responsible for language learning. For example, in natural language; semantics, syntax, and phonology operate in parallel, in close spatial and temporal contiguity, and because of this, artificial language learning paradigms have to be developed with the objective of controlling the influence of the various elements of natural language. Language researchers have thus turned to the basic concepts of linguistics.

At the core of any NLP task there is the important issue of natural language understanding. The process of building computer programs that understand natural language involves three major problems: the first one relates to the thought process, the second one is for the representation and meaning of the linguistic input, and the third one is the world knowledge. Thus, an NLP system may begin at the word level – to determine the morphological structure, nature of the word – and then may move on to the sentence level – to determine the word order, grammar, meaning of the entire sentence, etc.— and then to the context and the overall environment or domain. A given word or a sentence may have a specific meaning or connotation in a given context or domain, and may be related to many other words and/or sentences in the given context. Natural language is context dependent. The analyzing of the natural language begins at the morphological level. Morphology deals, primarily, with the structure of words. Morphological analysis detects, identifies and describes the meaningful constituent morphs in a word, which function as building blocks of a word. The analysis consists of the identification of parts of the words, or more technically, constituents of the words.

The second level is of the syntactic analysis. Syntax studies the sentence structure. It gathers the words into phrases, depending upon the
agreement. And the third one is semantics and pragmatics. At the apex of triangle is the NLP.

![Diagram of NLP, Semantics, Syntax, and Morphology]

**fig. 5.1**

### 5.2.1 The use of NLP

As discussed earlier the final aim of NLP is to build such a computer which can talk to humans, as humans talk to each other. Another important uses of NLP are in communication- man-machine ie question-answering. Information extraction, text summarization, information retrieval, etc., domain-specific applications; natural language interfaces; web-applications-intelligent search engines, digital libraries, corpus building etc.

There are some unexplained areas of natural languages, like the evolution of language and how does the child acquire the language. This is although a controversial topic. While making the computer learn natural language it is possible that we get an answer to those questions. So NLP also helps in understanding the natural language.
5.2.2 Machine Translation
One of the major tasks on NLP is Machine Translation. Machine translation is the process of translating from source language text into the target language. It is the process of translating documents without further human intervention. The translation is a procedure where a computer program studies a source text and produces a target text. The first and probably most important reason to use computers in translation is that, there is just too much material that needs to be translated and that, human translator cannot cope. There is a major requirement of technical translation in large corporations; they want terms to be translated in the same way every time, the terminology should be used consistently; computers are consistent, but human translators tend to seek variety; and this is no good for technical translation. Another reason is that the use of computer-based translation tools can increase the volume and speed of translation, and companies, organizations like to have translations immediately, the next day, or even the same day, human translators can not do that, so computers are perfect. Large companies want to reduce translation costs and, on the whole, with MT and translation tools they can achieve reductions. Any one of these reasons on its own can be sufficient justification for using and installing either MT systems or computer translation aids. Using computers for translation in different languages is highly recommended and necessary in today's fast pace world. Machine translation is an important and most appropriate technology for localization in a linguistically diverged country like India.

5.2.3 Approaches to MT
Machine translation has various approaches. There are basic three approaches, Direct, Rule-based and Data-driven. In a direct approach, translation is done in almost a word to-word manner. The basic characteristic is very simple; it needs to replace a word of source language to a word in target language using a bilingual dictionary. It also performs some morphological analysis before looking into the bilingual lexicon for the root words. Then it will perform the necessary reordering of the words as according to the target language sentence format. The morphological
processes may improve the quality of the translation but it does not really analyze the structure of the source language. The Rule based approach involves analysis of the source language. It needs countless built-in linguistic rules and bilingual dictionaries for each language pair. This process requires extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules. The software uses these complex rule sets and then transfers the grammatical structure of the source language into the target language. It works well in specific domain. Here, the translations are built on gigantic dictionaries and sophisticated linguistic rules.

It has transfer and Interlingua approaches as sub approaches. In transfer approach, after analysis, the source language structure is transferred into a corresponding target language structure which is then used to generate a target text. Analysis and generation programs are specific for particular languages. Differences between languages, in vocabulary and structure, are handled in the intermediary transfer program. Every approach follows different method of MT, but the basic structure and elements remain the same. Any analysis has to be started with the theories of morphology.

As seen earlier any tasks of NLP need to begin with the analysis of the natural language. And the basic analysis of the language begins at the morphological level. So the first step is a morphological analyzer.

5.3 Morphological Analyzer

A morphological analyzer supplies information concerning morphosyntactic properties of the words it analyses. Morphological Analysis is important component for building computational grammars as well as Machine Translation.

A morphological analyzer breaks sentence into words and suffixes. It identifies each component as PoS and tags it.

The basic principle of morphological analyzer is to get forms from a root and a set of properties (lexical category and morphological properties). A morphological analyzer needs to tackle the different syntactic categories such as nouns, verbs, adjectives, adverbs etc. separately, since the
addition of morphological constituents to each of these syntactic categories depends on different types of information.

Morphological analysis is the segmentation of words into their component morphemes and (usually) the assignment of grammatical information to grammatical categories and the assignment of the lexical information to a particular lexeme or lemma.

5.4 Gujarati verb phrase analyzer

The Gujarati verb phrase analyzer presented here has been designed according to the rules of grammar of Gujarati language. This can be called a hybrid approach as it is a combination of rule based and example based method. A list of rules has been given to identify the words and verbs are categorized in various groups. Each group behaves in the similar manner. Different paradigms are designed for this purpose. It is a language specific analyzer and analysis only the verb phrase portion of a sentence.

The analyzer takes the verb phrase as input and identifies each word. By word the computer understands any token that is between the two spaces. And the computer identifies VP as the phrase encoded in the brackets. The analyzer tags the words with the verbal attributes of tense, aspect, mood, voice and identifies them as main, auxiliary, participles, complement, case, adverb and negation.

5.4.1 Methodology

The computer has been provided with a set of words that belong to the closed category in the grammar like tense auxiliaries, particles, negations and few adverbs. When the computer analyses the VP first it tries to identify the words belonging to the closed class. If it finds any of them, then it tags them with the particular grammatical category allotted to it.

Now, there are those words left which do not belong to the closed category, hence are of open class. Computer identifies them by the suffix stripping method. A huge list of suffix is already stored in the computer. The computer identifies suffixes and cluster of suffixes. And tags the word according to the grammatical category.
The analyzer takes words as input and tries to identify the suffix, and checks whether the suffix is a valid suffix. If the suffix is valid, it is present in the word-list. If it is not valid then it is an unknown word for analyzer. If the suffix is valid then it will check the type of suffix- tense, aspect, mood, passive or causal. If it shows tense then which tense is it- present, future or past? And in same way it is with the aspects and others. If the last word is not there, then it checks the suffixes, it too has the list of suffixes and sees whether it is a valid combination or not. If it is a valid combination then it has to be present in the suffix list. So it tags the word according to the information encoded in the suffix. Then it goes to the adjoining word and repeats the same process, at this stage it does not need to check the list of tense auxiliaries. It will directly check the suffixes. When the word is identified it will move to next word and keep on repeating the process until it encounters the stop symbol.

1. राम (रोल).  
   rAm (bolyo)  
   Ram (spoke)  

In the above sentence, the verb phrase is in the brackets. The computer will analyse the verb phrase only. It will check from the right side of the word, as markings are added on right of the verb. The analysis will be as follows-  
रोल : Masculine Gender Singular Past Tense  
रोल : Masculine Gender Singular Perfect Aspect Of Present/Future/Past Tense  
रोल : Explicator Auxiliary  

2. मोना केरी (माय चे).  
   monA keri (khAy chhe)  
   Mona is eating a mango.  

In the above example, there is a main verb and an auxiliary. The analysis is –  
चे : Second/Third Person Singular, Third Person Plural, Neuter/Masculine/Feminine  
Gender Present Tense, Tense Auxiliary
3. राम (ञऱञऱी चाले छे).

\textit{rAma (jhaDapathi chAle chhe)}

Ram is walking fast.

In the above verb phrase there are three words; adverb, main verb and auxiliary. The analysis is-

\text{ञऱञऱी} : Second/Third Person Singular, Third Person Plural, Neuter/Masculine/Feminine Gender Present Tense, Tense Auxiliary

\text{चाले} : Second/Third Person Singular, Third Person Plural Present Tense

\text{चाले} : Second/Third Person Singular, Third Person Plural Progressive Aspect Of Present Tense

\text{ञऱञऱी} : Adverb

4. आज़, (वरसा० पडवो जैजिएँ)

\textit{Aje, (varasAda paDavo ja joie)}

It should definitely rain today.

Here, (पडवो जैजिएँ) \textit{paDavo ja joie} is the verb phrase. The computer will first look for the words that belong to the closed class. And as a result it will find the \textit{ja} as a particle and will tag it. Now, it will look for the remaining words with the suffix stripping method. It will identify the suffixes and tag the words.

Analysis-

\text{जिएँ} : Potential Auxiliary

\text{जिएँ} : First Person Plural, Masculine/Feminine/Neuter Gender Present Tense Verb

\text{ज} : Particle

\text{पडवो} : Masculine Gender Singular Potential Mood

The same process is applied to another example-
5. छोटां (रमता न हैता)

chokarA~ (ramatA~ na hatA~)
The children were not playing

And the result is

हैता : Neuter/Feminine Gender Plural Past Tense, Tense Auxiliary

ना : Negation

रमता : Neuter/Feminine Gender Plural Progressive Aspect Of Future/Past Tense

रमता : Explicator Auxiliary

5.5 Conclusion

A verb phrase analyzer has been presented with help of grammatical rules and computing knowledge. It identifies components of verb phrase using suffix-striping method as Gujarati is highly dependent on suffixes. TAM and voice are attributed by suffixes only. So identification of suffixes was found most convenient and verbs behave as per their endings, so it was possible to group them according to their behaviour, using paradigms. This is a sample analyzer and needs human filtration for accurate results.

References-

4. Wikipedia.com