International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979, Impact Factor: RJIF 5.72
www.allsubjectjournal.com
Volume 4; Issue 1; January 2017; Page No. 124-125

Parts of speech tagging in NLP

[1] R Umagandhi, [2] R Ramya

[1] M.A., M.PHIL. MCA, M.ED., Nadar Saraswathi College of Arts and science, Theni, Tamil Nadu, India
[2] M.A. English, Nadar Saraswathi College of Arts and science, Theni, Tamil Nadu, India

Abstract
Language as defined a tool of Communication. Grammar has the foremost position in using the Language. For translating a sentence from one Language to other Parts of speech was the main aspect. While translating a sentence from English it only gives ambiguous. Natural Language Processing is a scientific field that consist of Artificial Intelligence and Computational Linguistics concerned with the interactions between computers and human (natural) Language. Brill's Tagger was the major theory of *Post (Parts of Speech Tagging*). In this article we can converse about Language and Linguistics and how Linguistics plays its task in Grammar, the practice of NLP using Linguistics, Tagging of Parts of speech with several taggers and models.

Keywords: language, NLP, parts of speech tagging

Introduction
"Language is the expressions of ideas by means of speech sounds combined into words. Words are combined into sentences, this combination answering to those ideas into thoughts"

-Henry sweet

Language is the ability to obtain and use multipart systems of communication predominantly the human ability to do so. Language is essentially a form of human behavior. The scientific study of language as well as its structure, including the study of Grammar, Syntax and Phonetics is Linguistics. NLP (natural language processing) was a trending zone which deals with the language translation. With the help of natural languages moreover the linguistic patterns, the process of translation was categorized.

Natural Language Processing
"Natural language processing refers to computer systems that analyze, attempt to understand or produce one or more human languages, such as English, Japanese, Italian or Russian. The input might be text, spoken language or keyboard."

-J.F. Allen (2003)

NLP is ability of computers to understand the human language as spoken. It is ruling to make it easier for humans to alter their language into machine which is translated one. It is easier said than done to make computer understand the human speech. Through the process of NLP, we can effortlessly address through the computer as if addressing a person or even communicating with them. Natural context be converted into a verb or verb can be converted into a noun. But machine won't change noun and verb. This is a drawback of Natural Language Processing. Word to word translation is not relevant for machine translation. Tamil sentence when translated into English, the structure will change the full meaning completely. There are certain tribulations while handling NLP. Such as

- *Ambiguity problem.*
- *Language variability.*

In English there were certain rules for speaking and writing. The rules are simply known to us as Grammar.

Grammar
"*The set of rules that explain how words are used in a language"*

The core of Grammar is to organize words and sentences. We could say, words can be structured into sentences in many different ways. English the word Grammar is derived from Greek which means "Art of Letters". This art should be developed to enhance our language fluency. Grammar has undeniable work to classify the words in Language. To organize it, we can use parts of speech to proclaim the exact formation and denotation of a sentence.

Parts of Speech
"Parts of speech, is the sense in which the word is used and not the letters of which it is composed, that determines the part of speech to which it belongs"

-William Cobbett

In contemporary Linguistics, the label parts of speech discarded usually in favor of the term word clause or syntactic category. Part of speech is a term in Traditional Grammar for one of the eight main categories into which words are classified according to their functions in sentences: *Noun, Pronoun, Verb, Adverb, Adjective, Preposition, Conjunction, Interjection.* Though some Traditional Grammar have handled *Articles* (the, a, an) as a distinct part of speech, Modern Grammars most often comprise Articles in the category of *Determiners.* Each part of speech elucidate not the word is but how the word is used. As a matter of fact the same word can be as Noun in one sentence and a Verb or Adjective in the next. In NLP Parts of speech tagging was the key approach that deals with the Translation process. In English Language a word can be used as Noun, Verb or Adjective by the usage of the word in a sentence. While Translation it may give trouble. Meaning would change. So Parts of Speech Tagging were introduced in NLP to avoid ambiguous problems while translation.

Parts of speech tagging in NLP
Post was coming under Corpus Linguistics. It was also called

Grammatical Tagging or Word-Category disambiguation, is the process of attaching a label to a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition furthermore its context. POS Tagging is a course of accomplishment in which syntactic categories are consigned to words. It can be seen as a mapping from sentences to strings of the tags. POS tagging is now done in the context of Computational Linguistics using algorithms. POS Tagging algorithms fall into two distinctive groups: Rule-based and Stochastic. E. Brill's Tagger, one of the first and most widely used English POS-Taggers employ Rule-based algorithms.

Principle

Parts of speech tagging is hard-hitting than just having a list of words and their parts of speech, because some words can epitomize more than one part of speech at different times, A large percentage of word - forms are ambiguous. For instance, "Dogs" the plural noun can also be a verb.

"*The captain dogs the hatch"*

1) In the nautical context (dogs)
2) An action applied (hatch) to the object.

This two reasons were given by the semantic analysis. In this context "dogs" is a nautical term meaning "fastens (a water tight door) securely". In Parts of speech tagging by computer, it is typical to distinguish from 50 to 150 separate parts of speech for English. For example, NN for Singular Common Nouns, NNS for Plural Common Nouns, NP for Singular Proper Nouns work on Stochastic methods for tagging Koine Greek (DeRose 1990) has used over 1000 parts of speech, and found that about as many words were ambiguous there as in English. Research on POST has been intimately coupled to Corpus Linguistics. The earliest major Corpus of English for computer analysis was the Brown Corpus developed at Brown University by Henry Kucera and W.Nelson Francis, in the mid.1960's. Quiet a bundle of approaches have been anticipated to erect automatic taggers. In general Statistical methods has used *n-gram* models or *Hidden Markov Model based taggers*. POS tagging systems are CLAWS, VOLSUNGA.

Models in Parts of Speech Tagging

- *Hidden Markov Models*
- *Brown Corpus*

Hidden Markov Model

In the mid 1980's, researchers in Europe launched to us Hidden Markov Models (HMMs) to disambiguate parts of speech, when implement on a tag the Lancaster-Oslo-Bergen Corpus of British English. HMMs consist of counting cases (such as from the Brown Corpus), and assembling a table of probabilities of undeniable sequences. As, once you've seen an Article such as 'the', perhaps the next word is a Noun 40% of the time, an Adjective 40% and a Number 20%. Knowing this, a program can decide that "can" in "the can" is far more likely to be a Noun than a Verb or a Modal. HMMs learn the probabilities, not only of pairs, but triples or even larger sequences. So for example, if you've just seen a Noun followed by a Verb, the next item may be very likely a Preposition, Article or Noun, but much less likely another Verb. The European group developed CLAWS, a tagging program that did exactly this and achieved accuracy in the 93-95% range**.** Eugene Charniak points out in Statistical techniques in Natural Language Parsing (1997) [2], that merely assigning the most common tag to each known word and a tag "Proper Noun" to all unknown will approach 90% accurate because many words are unambiguous. CLAWS pioneered the field of HMM-based part of speech tagging. HMMs be positioned beneath the functioning of Stochastic taggers and are worn in various algorithms nearly all widely used being the Bi-directional inference algorithm [1].

Brown Corpus

The Brown Corpus of standard American English was the earlier of the modern, Computer readable, general corpora. It was set to one side by W.N. Francis and H. Kucera, Brown University, providence, RI. The Corpus consists of one million words of American English texts printed in 1961. Brown Corpus encloses a case with 17 ambiguous words in a row plus there are words are words such as "still" that can represent as many as 7 distinct parts of speech (De Rose 1990, p.82). Other taggers are

- *Viterbi Algorithm*
- *Brill Tagger*
- *Baum-Welch Algorithm*
- *Hidden and Visible Markov Model*

Conclusion

This journal is essentially dealing with what was meant by Natural Language Processing and how Parts of speech tagging plays its vital role in it. The Models of Post helps us to understand how to handle the process of tagging. The examples given in this article gives us the answer for the confusions while translating a sentence from English to other languages.

References

1. http://languagehelper.wikispaces.com/file/view/StatNLP+E-Book.pdf
2. http://www.revolvy.com/main/index.php?s=Tag%20questions&item_type=topic
3. http://www.revolvy.com/main/index.php?s=Lemmatisation
4. http://dictionary.sensagent.com/part%20of%20speech%20tagging/en-en/
5. http://www.slideshare.net/jaslinep/transformational-grammar
6. https://en.wikipedia.org/wiki/Part-of-speech_tagging
7. http://taggedwiki.zubiaga.org/new_content/70a447e3679e6df6280ab2f158eb1521