K. K. Wagh Institute of Engineering Education & Research, Nashik
Department of Computer Engineering

POS Tagging for the Hindi Language

Abstract

Part of Speech (POS) Tagger is an essential tool that is used to develop language translator and information extraction. The primary goal of Natural Language Processing (NLP) is to understand natural languages by parsing them.

Generally while analyzing natural languages there exist various sub-tasks. These sub-tasks depend on the inbuilt structure of language and do not require complete knowledge and understanding of language. Part-of-speech tagging is one of them. Part-of-speech tagging is a practice of assigning language-specific grammatical tags to each word of the language-specific input text, according to word's appearance in the text. These tags can be like a noun, adverb or a number.

There exist a variety of taggers for the most popular language in the world, i.e., English. However, such taggers cannot be used for morphologically rich Hindi language as a difference exists between structures of both languages. In order to perform POS tagging, we have used libraries which are provided by NLTK Python library. The tagging is performed in the following stages:

1. Getting dataset.
2. Pre-processing the dataset
3. Stemming and Lemmatising words
4. Training POS Tagger
5. Tagging new line

Project Members:

Devashish Katoriya

Ashutosh Bhawsar

Bhushan Shilawat

Ninad Kapadnis

Project Guide:

Prof. A. V. Kolapkar