ABSTRACT

The process of assigning or labelling part of speech for every word in a given sentence according to the context is called as part of speech tagging. This is one of the useful tasks in Natural Language Processing (NLP). It plays an important role in Speech and NLP such as Speech Recognition, Speech Synthesis, Information Retrieval, Word sense disambiguation and Machine Translation. But the main problem of part of speech tagging is that in a natural language, a huge percentage of words are ambiguous i.e., one word may present more than one lexical category. Therefore ambiguity resolution is the main challenge of part of speech tagging. There are a many approaches to implement part of speech tagger, i.e. Rule-Based approach, Statistical approach and Hybrid approach [1]. Hybrid models are combination of rules based and statistical models and make new methods using strongest points from each method [2]. It makes use of essential feature from ML approaches and uses the rules to make it more efficient. Our objective of this thesis is to develop an effective POS tagger for Nepali language using hybrid based approach, .i.e. Hidden Markov Model (HMM) integrated with Rule-Based method.

In this thesis, we present a Nepali part-of-speech tagger which is a morphologically rich language. Nepali (नेपाली) language originally belongs to the Indo-Aryan branch of Indo-European family [5]. This language takes its roots from Sanskrit, which is the classical language of India. Nepali language was known as Gurkha, Gurkhali or Khas Kura[6]. In the 11th century Nepali Language was initially developed from the Brahmi script and eventually Nepali Language got written in the Devanagari script. Nepali is spoken by more than 40 million people, mostly in Nepal, Bhutan, Myanmar, West Bengal and in other parts of India.

In this thesis, we have also presented a tagset for Nepali that has been developed as a part of the work which consists of 43 tags. The tagset covers almost all the grammatical categories in the Nepali language and is based on BIS[1] (Bureau of Indian Standards) framework.

The research work has been initiated with a study on part of speech tagging on its paradigms and various approaches, namely, Rule-based Model, Hidden Markov Model, Maximum Entropy Model, Memory based Learning Technique, Conditional Random Field Model and Transformation based part of speech tagging model. At this initial stage of POS Tagging for Nepali Language, we have very limited resource of annotated corpus. We tried to see which technique maximizes the performance with these limited resources. By experiments, the best configuration is investigated using HMM with rule based approach (hybrid approach). Initially, a very limited lexicon and rules were present in POS tagger which resulted in a low accuracy. However, when our lexicon reached more than 20,000 words, the system performed with an enhanced accuracy of 93.50%. We believe that further error analysis and increase in the size of tagged corpus would improve the system performance.

[1] The Bureau of Indian Standards (BIS) is the national Standards Body of India working under the aegis of Ministry of Consumer Affairs, Food & Public Distribution, Government of India.