ABSTRACT

Named Entity Recognition (NER) is a sub-problem of Information Retrieval and Information Extraction. NER involves processing of unstructured and structured documents and it also identifies those entities that refer to people, location, organization and companies, and soon. Named Entity Recognition is intuitively simple for humans. Most of the named entities are proper names and have initial capital letters(for English language) and can easily be recognized. In natural languages there are large amount of ambiguity problems which makes it difficult for human named entity recognition process. For handling these ambiguities memorizing(lists of) names can assist the human recognition process.

In our thesis, Named Entity Recognition (NER) system is built with computational/unsupervised approach. Few examples of named entity types can be listed under human supervision. This approach will give better results for Named Entity Recognition (NER) especially for inflectional languages. We in our thesis investigate the process of improvement in performance of Named Entity Recognition for Nepali texts: what are the best statistics of the named entity type and what techniques maximize the generalization performance? To investigate the best approaches to Named Entity Recognition, we have selected two popular machine learning approaches, Gazetteer lists approach(Linguistic or Rule based approach) and Hidden Markov Model approach(Machine learning (ML) based approach).

Following contribution are made for this research work: First, stemming of text is done for finding out the proper root words from the sentence. Then, part of speech tagging is done for identification of nouns followed by chunking which identifies proper nouns from the sentence. Finally, NEs(Named Entities) are detected from the sentence using n-gram technique.

For machine learning system this semi-supervised approaches are about to break new grounds. In our thesis, we show that the hybrid approach can build a complete NER systems. On standard evaluation corpora, we can achieve that level of performances that compare to supervised systems in the task of annotating NEs(Named Entities) in texts.