Abstract

Named Entity Recognition which is a subfield of information extraction is one of the most important topics of Natural Language Processing. It is a process through which the machines understand the proper nouns in text and associates them with proper tags. NER has made significant progress in European languages, but in Indian languages due to the lack of proper resources, it is a challenging task. As natural language is a polysemous, ambiguity exists among the name references. Recognizing the ambiguity and assigning a proper tag to the names is the goal of NER. Thus NER is a two stage process i.e., identification of the proper nouns and the classification of these nouns into different classes such as person, location, organization and miscellaneous which includes date, time, year, etc. The main aim of our work is to develop a computational system that can perform NER in text in the Assamese language which is a resource poor Indo-Aryan language.

Our thesis discusses the different issues related to NER in general and in Indian languages, along with the different approaches to NER. We discuss the different works carried out by different researchers in different Indian languages along with the datasets and the tagsets used by them. We focus on rule-based approach first which involves the identification of the root word from an inflected form, which is known as stemming, and further we derive some hand coded rules for different classes of NEs. We also discuss the tagging of NEs using gazetteer list. Then we experiment with two machine learning approaches, namely CRF and HMM, and find that in *our language the system performs reasonably well. Lastly we implement hybrid approach which involves both rule-based approach and machine learning approach and find that compared to single approach, the combined approach improves the performance of the system. Finally we conclude our work and also suggest some future work in this line.*