Named Entity Recognition, Classification and Transliteration in Bengali

Named Entity Recognition and Classification (NERC) is an important task in almost all Natural Language Processing (NLP) application areas. In most NLP application areas, there is also an increasing need to translate Out Of Vocabulary (OOV) words from one language to another. Technical terms and named entities (NEs) constitute the bulk of OOV words. Translation of NEs involves both translation and transliteration. The research in NERC and Transliteration involving Indian languages has started to emerge very recently. Bengali is the seventh language in the world in terms of number of native speakers, one of the popular languages in India (Rank 2) and the national language of Bangladesh.

In this talk, I will present a multi-engine approach for NERC in Bengali and a modified joint source-channel model for NE transliteration that I developed in my PhD thesis. The approaches to NERC are based on unsupervised machine learning, supervised machine learning algorithms like Hidden Markov Model (HMM), Maximum Entropy (ME), Conditional Random Field (CRF) and Support Vector Machine (SVM), a semi-supervised machine learning algorithm and voting. The proposed modified joint source-channel model is based on a regular expression based alignment technique. I will report a number of other lexical resources that we had to create ourselves as a basis for NERC and transliteration systems.