IndoWordNet and Multilingual Resource Conscious Word Sense Disambiguation

Pushpak Bhattacharyya
Computer Science and Engineering, IIT Bombay, India

Wordnets have become crucial resources for NLP. They are complex structures capturing various kinds of lexico semantic relations among words. The first wordnet in the world was built; for English at Princeton University. This was followed by wordnets of European languages forming the EuroWordnet. At IIT Bombay the first wordnet for Indian languages was constructed for Hindi. This was followed by many other languages including Marathi, Sanskrit, Bangla, Tamil, Telugu, Punjabi, Gujarathi, and North East languages. In the first part of the talk we describe the principles and methodolgies followed in multilingual wordnet construction. We close this part of the discussion with a brief description of the Pan-Indian multilingual dictionary standard that IndoWordnet has given rise to and is the essential resource for multilingual WSD.

Word Sense Disambiguation (WSD) is a fundamental problem in Natural Language Processing (NLP). Amongst various approaches to WSD, it is the supervised machine learning (ML) based approach that is the dominant paradigm today. However, ML based techniques need significant amount of resource in terms of sense annotated corpora which takes time, energy and manpower to create. Not all languages have this resource, and many of the languages cannot afford it.

In the second part of the presentation, we discuss ways of doing WSD under resource constraint. First we describe a novel scoring function and an iterative algorithm based on this function to do WSD. This function separates the influence of the annotated corpus (corpus parameters) from the influence of wordnet (wordnet parameters), in deciding the sense. Next we describe how the corpus of one language can help WSD of another language, i.e., LANGUAGE ADAPTATION. This is presented in three setting of "complete", "some" and "no" annotation. From this we move on to DOMAIN ADAPTATION where the notion of active learning and injection are pursued to do WSD in a domain with little or no annotated corpora. The extensive evaluation and good accuracy figures lend credence to the viability of our approach which points to the possibility of expanding from one language-domain combination to all language-domain combinations for WSD, i.e., multilingual general domain WSD, a long standing dream of NLP.

The talk is presented in a multilingual setting of Indian languages. There are 22 official languages in India with strong requirements of machine translation and cross lingual search. Our languages of focus in this talk are Hindi and Marathi along with English and the domains of focus are Tourism and Health which are important to India.

The presentation is based on work done with PhD and Masters students and researchers: Dipak Narayan, Nitin, Rajat, Deabsri, Mitesh, Salil, Saurabh, Anup, Sapan and Piyush, and published in fora like ACL, COLING, EMNLP, GWC, ICON and so on.