CLE SEMINAR SERIES-III

Topic: Sense Tagged CLE Urdu Digest Corpus

Presenter: Ms. Saba Urooj

Presentation Date: 11th November, 2014

Venue: KICS Seminar Hall

Abstract:

This paper presents the construction of an Urdu Sense Tagged corpus using four main lexical resources; an Urdu wordlist consisting of 5000 high frequency content words, a 100K words corpus annotated with part of speech (POS) tags, an Urdu WordNet with approximately 5058 senses and Urdu morphological analyzer. The paper also briefly presents Urdu word-sense annotation tool, a software tool developed to provide an easy interface for sense tagging, ensuring tagging consistency and accelerating the annotation speed. In this version of the Urdu sense tagged corpus, 17,006 words have been sense tagged with 2285 unique senses. The final section discusses the linguistic and tool specific challenges in the construction of sense tagged corpus and describes future work in this context.