CLE SEMINAR SERIES-III

Topic: Word Segmentation System for Urdu OCR
Presenter: Ms. Farah Adeeba
Presentation Date: 30th April, 2013.
Venue: Main Lab, CLE building.

Abstract:

This talk presents a technique for word segmentation system for Urdu OCR. Word segmentation is the problem of dividing a string of written language into words. Some languages such as English provide the clear indication for words. In such languages the words are separated using the space. In Urdu, letters are joined together to form units called ligatures and word can consist of one or more ligatures without explicit delimiter to indicate word boundaries. Task of Word segmentation in Urdu OCR is to convert a given sequence of ligatures into a sequence of words and resolve ambiguity among them. The system has been developed to be used in Urdu Nastalique OCR system but can be used to determine word boundaries for any given text. Using this technique, a word identification rate of 97.9% has been achieved.