

Department of Chinese and Bilingual Studies
Department of Computing
Joint Research Seminar

Date: 23 March 2015 (Monday)
Time: 4:30 – 6:00 pm
Venue: GH803, PolyU

“Chinese Word Segmentation from a View of Word Formation”

By

Dr Teresa Wu

Visiting Scholar, Noah’s Ark Lab, Huawei

Abstract

Word segmentation is considered an important first step for natural language processing tasks. However, it is challenging to appropriately identify Chinese words in any given text as the language is composed of a string of continuous characters without any demarcated boundaries or explicit morphology. The absence of any explicit indicators and delimiters results in ambiguity or disagreement in Chinese word segmentation.

In order to unambiguously signal boundaries in a sequence of characters, Chinese word segmentation rules need to reflect the fundamental principles of word formation, that is, morphology. The idea of segmenting Chinese characters into small chunks and working on individual characters (fine-grained analysis) is suggested to assist the segmentation process. Morphological restrictions and analyses of Chinese compounding, derivational complex words, affixation and phrasal structure are taken into consideration to develop Chinese word segmentation rules that can be used by a computer in processing the language.