



**The Hong Kong University of Science & Technology  
Human Language Technology Center  
Department of Information & Systems Management  
Department of Mathematics  
Department of Computer Science and Engineering**

**JOINT STATISTICS SEMINAR**

**“Nonparametric Bayesian Methods in Language Modeling”**

by

**Dr. Daichi MOCHIHASHI  
NTT Communication Science Laboratories**

**ABSTRACT**

In this talk, I will introduce some nonparametric Bayesian approaches recently grown in natural language processing, using such as Dirichlet processes, Pitman-Yor processes and their hierarchical extensions.

In the first part of the talk, I will first present what natural language processing is and why language modeling is a quite interesting and important problem. Nonparametric Bayesian priors will prove very useful there: it allows to automatically infer latent “categories” (syntactic and semantic) without human intervention, which needed enormous effort and are often inaccurate to describe actual phenomena in natural language. Among many natural language processing techniques, “ $n$ -gram” language models, i.e.  $(n-1)$  order Markov models over words, are very fundamental and heavily used in speech recognition and statistical machine translation.

In the second part of the talk, I will present my latest work on “infinite-gram” language model or “infinite Markov model” in NIPS 2007, where Markov order  $n$  is integrated out nonparametrically. This amounts to introducing a very simple prior over stochastic infinite trees, other than the Kingman's coalescents: it might have a close relationship to tailfree processes. I will present experimental results on large texts using a Gibbs sampler, and discuss about exchangeability and relationship to information theory.

Date: Friday, 23 May 2008  
Time: 4:00pm – 5:00pm  
Venue: Lecture Theatre H (Chen Kuan Cheng Forum, near lifts 27/28)  
The Hong Kong University of Science & Technology

For enquiries, please call 2358 7008 or visit our website at  
<http://www.cs.ust.hk/~hltc/seminars.html>  
\*\*\*\* ALL are Welcome \*\*\*\*

*Biography:*

*Daichi Mochihashi is a postdoctoral researcher in NTT Communication Science Laboratories, Kyoto, Japan (Japanese equivalent of AT&T Labs Research). He obtained his BS and PhD from University of Tokyo and Nara Institute of Science and Technology, respectively, in 1998 and 2005. His main interest is natural language processing, especially from Bayesian point of view. After graduation, he was a researcher at ATR Spoken Language Communication Research Laboratories and conducted research on language modeling. He joined NTT in 2007, as a member of machine learning group.*