情報科学研究科 重点プロジェクト

数学と諸分野の協働推進による学際的・総合的な新領域研究の開拓

MATHEMATICS × EXTENSIVE SCIENCE

第26回講演会 兼 第73回応用数学連携フォーラム

日時 2019年2月12日（火）14時00分～15時00分

会場 東北大学 情報科学研究科棟 中講義室

講演者 持橋 大地 氏（統計数理研究所）

タイトル 階層Pitman-Yor過程による半教師あり形態素解析

概要

自然言語の単語の頻度分布は一般に巾乗則に従い，それはPitman-Yor過程(Pitman and Yor 1997), あるいはそれを階層化した階層Pitman-Yor過程(Teh 2006)とよばれる確率過程によってよく近似できることが知られている。このことを利用し，文字と単語のMarkovモデルにネストされた階層Pitman-Yor言語モデルを考えることで，たとえ未知の言語でも，生の文字列だけから「単語」を完全に自動的に推定できる統計モデルを2009年に発表した(Mochihashi et al. 2009)。しかし，実際の応用では人手で構築した単語分割の教師データや辞書が利用できることが多く，そうした教師データの情報も利用する半教師あり学習の必要性がもっとも高いと考えられる。そこで，形態素解析の教師あり学習のための標準的なモデルであるCRF(条件付確率場)と上記の階層Pitman-Yor過程による言語モデルを接続し，互いに補い合うことで,教師データにない未知の言語表現からも適切に「単語」を認識することのできる半教師あり学習の枠組(Fujii et al. 2016)について解説する。また，近年のニューラル言語モデルの発展も踏まえ，現代的な立場からこうした研究をどう発展させていくかについての議論も同時に行いたい。

http://www.math.is.tohoku.ac.jp/~project/