2009－NL－190 (8)
2009／3／25

ベイズ階層言語モデルによる教師なし形態素解析*

持橋 大地　山田 武士　上田 修功
NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町「けいはんな学研都市」光台 2–4
daichi@cslab.kecl.ntt.co.jp {yamada,ueda}@cslab.kecl.ntt.co.jp

概要

本論文では, 教師データや辞書を全く必要とせず, あらゆる言語に適用できる教師なし形態素解析器および言語モデルを提案する. 観測された文字列を, 文字 n グラム-単語 n グラムをノンパラメトリックベイズ法の枠組で統合した確率モデルからの出力とみなし, MCMC 法と動的計画法を用いて, 繰り返し「単語」を推定する. 提案法は, あらゆる言語の生文字列から直接, 高精度で未知語のない n グラム言語モデルを構築する方法ともみなすことができる.

キーワード: 形態素解析, 言語モデル, ノンパラメトリックベイズ法, MCMC

Bayesian Unsupervised Word Segmentation with Hierarchical Language Modeling

Daichi Mochihashi　Takeshi Yamada　Naonori Ueda
NTT Communication Science Laboratories
Hikaridai 2-4, Keihanna Science City, Kyoto Japan 619-0237
daichi@cslab.kecl.ntt.co.jp {yamada,ueda}@cslab.kecl.ntt.co.jp

Abstract

This paper proposes a novel unsupervised morphological analyzer of arbitrary language that does not need any supervised segmentation nor dictionary. Assuming a string as the output from a nonparametric Bayesian hierarchical n-gram language model of words and characters, "words" are iteratively estimated during inference by a combination of MCMC and an efficient dynamic programming. This model can also be considered as a method to learn an accurate n-gram language model directly from characters without any "word" information.

Keywords: Word segmentation, Language Modeling, Nonparametric Bayes, MCMC

[0]本文は, PDF 版を参照されたい.