二項事後分布に基づくN-gram 記号連鎖確率の推定

川端　豪

[要旨]　音声認識の言語モデルとしてよく用いられる n-gram 記号連鎖確率の高精度推定は、統計的言語処理における重要な課題の一つである。言語モデルとしての能力を高めるために、グラム数を大きくすると n-gram コンテキストの種類が語彙数の n 乗オーダになるため、推定の信頼性を高めるために非現実的な量の言語コーパスが必要になる。本論文では、(n-1)-gram の二項事後確定分布（BPD）を n-gram の先験確率として継承することによって少数のデータから n-gram 確率を推定する理論について述べる。また、理論上恣意的なパラメータとなる継承係数の決定法を示す。加えて、n-gram 確率推定値および継承係数を見通しよく計算するためのデータ構造について述べる。