SLR 3-3

Particle Filter による文脈のベイズ推定
Bayesian Context Estimation via a Particle Filter

ATR
Spoken Language Communication Research Laboratories, Kyoto Japan

概要

文脈の時間的な変化に適応する言語モデル
話題の変化をオンラインで推定
– 超高次元離散系列の時系列推定

アプローチ

・隠れた変化点をもつ、確率的生成モデル
・テキストの確率的トピックモデル+逐次モンテカルロ法によるオンライン推定

結果

・「文脈の変化」を記述する最初の確率モデル
– 通常のHMMでは分布の遷移を扱えない
・現在、平均予測確率最大の文脈言語モデル

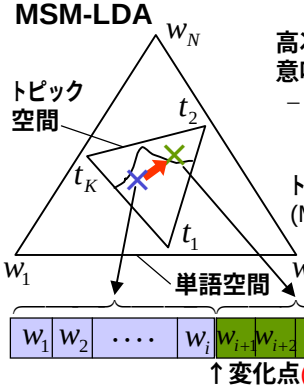
連絡先

ATR音声言語コミュニケーション研究所
音声言語処理研究室
担当者: 持橋大地
<daichi.mochihashi@atr.jp>

提案手法

文脈の時間的な変化を記述する確率モデル
■ 単語の出現頻度から、話題を分布として推定
問題：自然言語の単語空間は超高次元 (数万～数十万)
→ 二つのアプローチ：MSM-LDA、MSM-DM

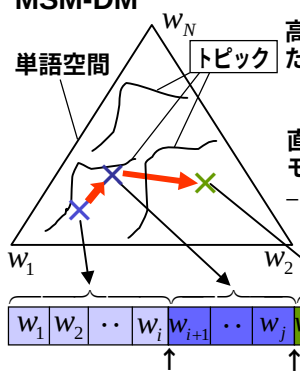
MSM-LDA



高次元の単語空間を、低次元の意味的なトピック空間に圧縮 (LDA)
– 変分ベイズEMアルゴリズム

トピック空間上の状態遷移モデル (Mean Shift Model, MSM)

MSM-DM



高次元の単語空間をカバーするため、混合分布を利用 (DM)

直接、単語空間上の状態遷移モデル(MSM)
– 情報の圧縮がなく、ロスがない

問題：時刻tでの変化点確率を求めること。

話題の変化点確率

■ 現在の文脈からみて「変な」語 (=確率の低い単語)
→ 話題が変化した可能性が高い

$$\begin{cases} p = \text{文脈をリセットした時の単語の予測確率} \times \rho \\ q = \text{現在の文脈での単語の予測確率} \times (1-\rho) \end{cases}$$

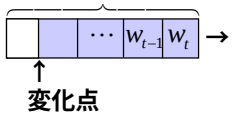
– ρ：文脈変化の事前確率 (例えば、0.01)

$$\text{話題の変化確率} = \frac{p}{p+q} \quad \text{(時刻 t=1,2,3,…毎に計算)}$$

■ ρ も自動的に推定できる (ベータ事後分布の期待値)

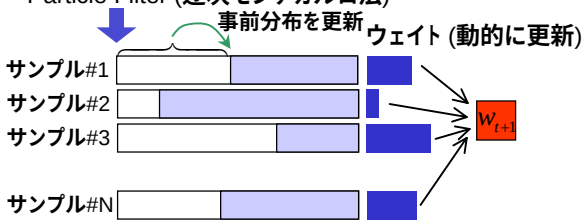
$$\langle \rho_t \rangle = \frac{\alpha + (\text{これまでの変化点数})}{\alpha + \beta + t} \qquad \alpha, \beta：\text{ハイパーパラメータ}$$

文脈言語モデルと Particle Filter



直前の話題の変化点を見つけ、そこからの履歴を用いて次の語を予測
– 文脈長は、場合によって変化

■ 変化点は、実際には複数シミュレーション
→ Particle Filter (逐次モンテカルロ法)



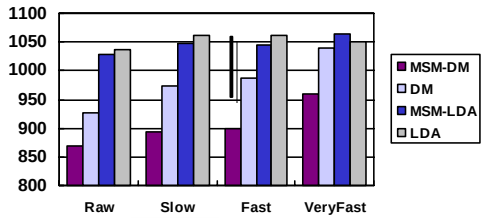
● 複数の変化点をシミュレーションし、重みづけして予測
● 直前の変化点以前の情報も、事前分布として取り込む

実験条件

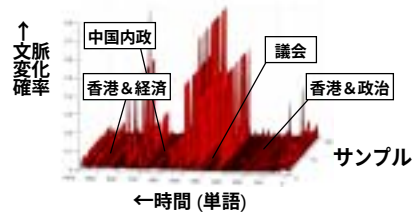
- British National Corpus (幅広い話題)
- 11,032,233語、語彙数＝52,846
- テストデータ：100文書x100文
  - Raw: 連続した100文を抽出
  - Slow～VeryFast: ランダムなスキップあり (Slow: 小、Fast: 中、VeryFast: 大)

実験結果

パープレキシティ=1/平均予測確率



実際のテキストの文脈変化確率



結論

話題の変化を記述する確率的モデルを定義し、オンラインで話題の変化点を推定することで、文脈に動的に適応する長距離言語モデルが可能となった。