混合ディレクレ分布を用いた文脈のモデル化と言語モデルへの応用

山本幹雄　貞光九月　三品拓也
筑波大学

1. 手法の概要と結果（11枚）
2. 他手法との関係（9枚）
   - Cache (1990〜), Trigger (1993〜)
   - Topic (Aspect) models
     - Dynamic adaptation
     - LSA, PLSA, Unigram Mixtures, LDA
     - DMA (Dirichlet Mixtures Allocation) ⇦

長距離情報による改善例

- Xに当てはまる単語を入れよ。

A (1) X
(2) の X
(3) マリナーズのX
(4) 大リーグで激しい打率争いをしている マリナーズのX

B (1) 不利益を被るのは個人X
(2) 株式市場に明らかな復調の兆しは見られない。... 持ち合い株の解消売りも継続しており... 不利益を被るのは個人 X

ベイズ学習

統計的言語モデル ⇨ $p(w|h)$

- *n*gramモデル＋Trigger

$p(w|h)$　　$p(w_i|w_{i-2}, w_{i-1}, t_1, t_2, ...)$

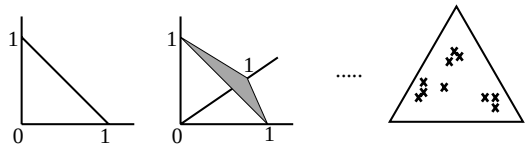
- ベイズ学習
  - 事後確率∝尤度×事前確率　　$\mathbf{P} = \{p(w)\}_{w=1,...,V}$

$$p(\mathbf{P}|h) \propto p(h|\mathbf{P})p(\mathbf{P})$$
$$= p_{Mul}(h|\mathbf{P})\underline{p(\mathbf{P})}$$

Pの事前分布

事前分布：$p(\mathbf{P})$

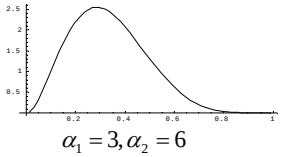
- Pは単体上のベクトル
  - 単体：$\sum_w p(w) = 1$



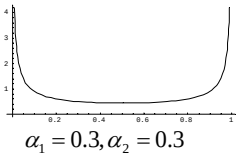
- 単体上の確率分布 ⇨ Dirichlet分布
  - ベータ分布の多次元化
  - 多項分布の共役事前分布　　$p(\mathbf{P}) = \frac{\Gamma(\sum_w \alpha_w)}{\prod_w \Gamma(\alpha_w)} \prod_w p(w)^{\alpha_i - 1}$

Dirichlet分布

2変数（ベータ分布）：

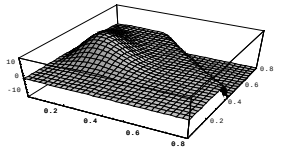


$\alpha_1 = 3, \alpha_2 = 6$

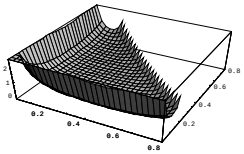


$\alpha_1 = 0.3, \alpha_2 = 0.3$

3変数：



$\alpha_1 = 3, \alpha_2 = 6, \alpha_3 = 6$



$\alpha_1 = 0.3, \alpha_2 = 0.3, \alpha_3 = 0.3$

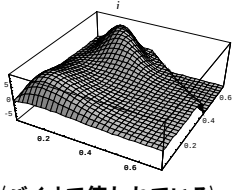
Dirichlet Mixture

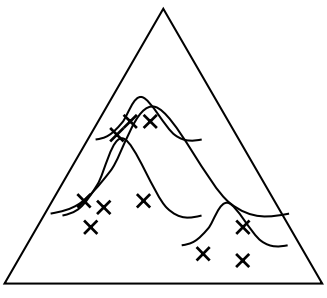
- Dirichlet分布では共分散がモデル化できない

⇩

- Dirichlet分布の混合

$$p(\mathbf{P}) = \sum_i \lambda_i p_i(\mathbf{P})$$





（バイオで使われている）

パラメータ推定

- 最尤推定 (Empirical Bayes)
  - 事前分布を考慮に入れたデータの尤度:

$$p(D \mid \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \prod_i \int p_{Mul}(d_i \mid \mathbf{P})\, p_{DM}(\mathbf{P} \mid \boldsymbol{\lambda}, \boldsymbol{\alpha}) d\mathbf{P}$$

(混合多変量負の超幾何分布、混合Dirichlet-Multinomial、混合Polya分布)

- アルゴリズム
  - ニュートン法＋EM
    - 収束しない、計算時間が膨大
  - ニュートン法の改良(Minka1998)＋EM
    - 収束する、計算時間が膨大
  - Leaving-one-out(Minka98)＋ニュートン法の改良＋EM
    - 収束する、高速

予測分布 (適応)

- 事後分布

$$p(\mathbf{P} \mid h) \propto p(h \mid \mathbf{P}) p(\mathbf{P})$$

- 予測分布 (事後分布の期待値)

$$p(w^* \mid h) = \int p(w)\, p(\mathbf{P} \mid h) d\mathbf{P}$$

  - 閉じた式を導出可能
  - LDA等は閉じた式を導出できない (積分ができない)
    - 変分近似、EP法、MCMC

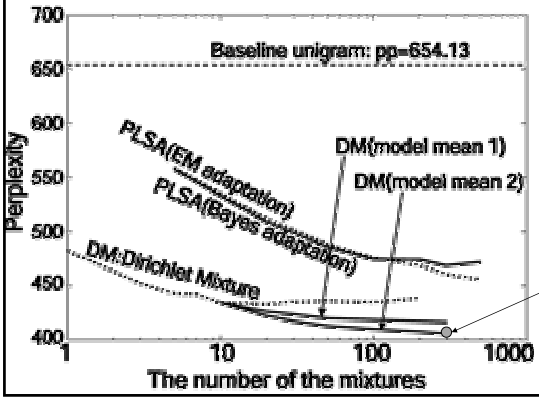
モデル平均

- 過適応が大きい
  - 実験結果
    - 10混合で性能が飽和してしまう

⇩

- モデル平均
  - パラメータ数 (混合数) の異なる複数のモデルを平均する
    - 300混合まで性能が上がる

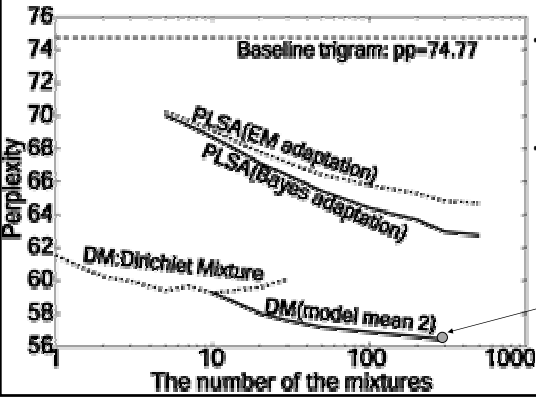
実験結果
(ユニグラム・パープレキシティ)



・学習データ
毎日新聞
1999年版,
・テスト
1998年版
・語彙
20000単語
・適応
20単語毎
に動的適応

38%削減

実験結果
(トライグラム・パープレキシティ)



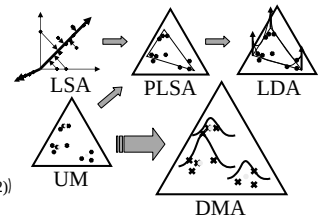
・トライグラム
学習データ
毎日新聞
1999年版,

・トライグラムとの融合方法
Unigram rescaling

25%削減

他のモデルとの関係

- Topic (Aspect) Models
  - LSA : Bellegarda(1998)
  - PLSA : Gildea&Hofmann(1999)
  - Unigram Mixtures : Nigam et al.(2000)
  - LDA: Blei et al.(2001), Minka(2002), (三品&山本(2002))
  - DMA



Aspect (Topic) Models

- *h*は離散的なTopic, *t*, を経由して*w*に影響を与える

Simplicial Mixture

$$p(w|h) = \sum_t p(w|t)\,p(t|h)$$

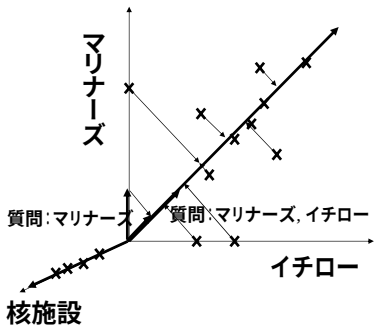
各topicの重み

$t=1$	$t=2$	$t=3$	
言語モデル1 (スポーツ分野)	言語モデル2 (政治分野)	言語モデル3 (経済分野)	...

- Soft decision: 上記
- Hard decision: $p(t'|h)=1$, 他は0.

LSA: Latent Semantic Analysis

Vector Spaceを次元圧縮した空間 (特異値分解)



ドキュメント クラスター

$$p(w|h) = p(w|\tilde{d})$$
$$= \sum_t p(w|D_t)\,p(D_t|\tilde{d})$$

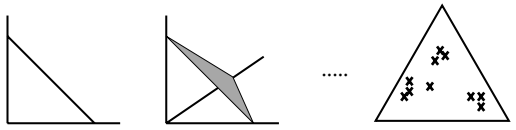
角度を使う

J.R. Bellegarda (1998)

Probabilistic LSA : PLSA

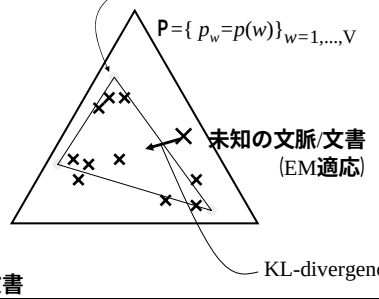
Thomas Hofmann, 1999, Probabilistic Latent Semantic Indexing, Proc. of SIGIR'99, pages 50-57

- LSA → PLSA
  - 空間: ベクタースペース → 確率空間 (単体)
  - 圧縮基準: 2乗誤差最小基準 → 最尤推定
  - 距離: ユークリッド距離,角度 → KL-distance
  - 手法: 特異値分解 → EMアルゴリズム
- 単体(Simplex) $\Sigma\, p(w) = 1$



PLSA

$$p(w|d) = \sum_t p(w|t)\,p(t|d)$$



× = *d* : 学習用文書

PLSAの次元圧縮結果:10次元

1999年毎日新聞 98211記事, 20k単語, unigram

普通のunigram確率から変化の大きかった上位単語

次元1：アメリカンフットボール, 新庄, ラグビー, 準々, 終盤, オールスター, 競技,
次元2：申し込み, ホール, 遺志, はがき , 消印, 会館, 告別, 喪主, ワッハ, 公益社
次元3：此花, 失跡, 課, 供述, 罪, 不定, 検察, 起訴, 無罪, 同署
次元4：摘出, 体内, 感染, 物質, アルツハイマー, 耐性, 脳波, 精巣, 胚, 卵子
次元5：あんた, 冷たい, ママ, 祖母, ええ, 先日, なあ, 蒙, 通有, あたし
次元6：覆わ, 両側, 運行, シャトル, ヘクタール, ウオーク, 水中, 海中, レジャー
次元7：エレクトロニクス, 三和, 年度末, 会計, 既存, 顧客, 住友銀行, 住友, 還付
次元8：エンターテインメント, 主題歌, スタジオ, 堪能, 美し, 旬, コンセプト, いやす
次元9：閣議, 正副, 再選, 選, 自民党, 官房, 党内, 自由党, 民輔, 府連
次元10：米兵, 朝鮮民主主義人民共和国, 平和, 任務, 断固たる, 紛争, ミサイル

PLSAの性能・問題

- 性能
  - 情報検索：LSIよりも高性能
  - 言語モデルへの応用 (D.Gildea and T.Hofmann. 1999.)
    - unigram-rescaling法
    - TrigramPPを約20%削減できる
- 問題 $p(w|d) = \sum_t p(w|t)\,p(t|d)$
  - パラメータが多い：$p(w|t)$→ V×M, $p(t|d)$→M×D
  - generative aspect modelsでない
    - 未知文脈/文書Wの確率を出せない

⇒ ベイズ学習

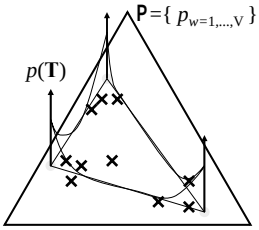
LDA: Latent Dirichlet Allocation

(Blei et al. 2001)

$$p(w|d) = \sum_t p(w|t)\underline{p(t|d)}$$

単体上のベクトル

$\mathbf{T}=\{ p_t=p(t) \}_{t=1,...,M}$



- **Tの事後分布**

$$p(\mathbf{T}|h) \propto p(h|\mathbf{T})P_{Dir}(\mathbf{T})$$

- **予測分布**

$$p(w|h) = \int p(w|\mathbf{T})p(\mathbf{T}|h)d\mathbf{T}$$

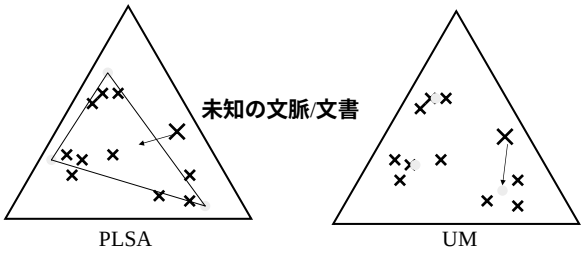
※変分近似: 2001
EP: 2002

UM: Unigram Mixtures

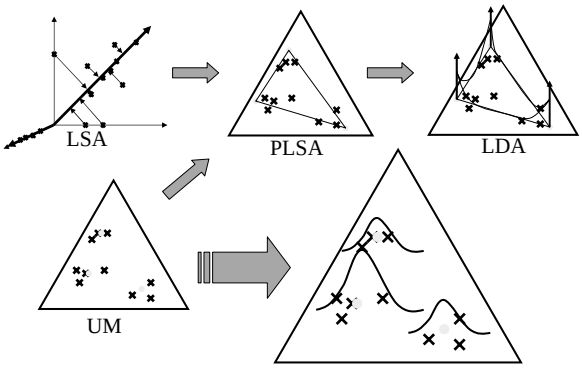
(Nigam et al. 2000)

- Hard decision: $p(t'|h)=1$, **他は**0.

$$p(w|h) = \sum_t p(w|t)p(t|h) = p(w|t')$$



DMA: Dirichlet Mixture Allocation



まとめ

- **文脈/文書モデル**
  - **単語出現確率の事前分布 → 混合ディレクレ分布**
    (PLSA**とは別の系列**)
- **技術的な詳細**
  - **パラメータ推定方法**
    - **問題**：**高次元、高混合数の場合**（**パラメータ数**200**万**）、**ニュートン法＋ＥＭでは収束しない、学習時間が膨大**
    - Minka**のニュートン法＋ＥＭで収束する、高速**
  - **モデル平均**（**過適応が問題**）
- **性能**
  - Trigger**なみ**

LDA**系と比べての優位性**
- **ベイズ学習で近似が不必要**
- **全確率領域のモデル化**