

トピック情報を用いた 評価文書分類

貞光九月 山本幹雄

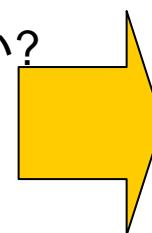
(筑波大学)

内野寛治 松井くにお

(Fujitsu Laboratories of America, Inc.)

研究の目的と背景

- Web上の膨大なテキストデータから、ある対象に対する意見を定量的に示したい
e.g. **肯定的意見** x% **否定的意見** y%
- 評価文書分類
 - ある対象に対する評価を含む文書(評価文書)を肯定評価文書・否定評価文書の2ラベルに分類する
 - 単純なNaïve Bayesで80%の精度
※ベイズ識別： $p(\omega_{\text{posi}}|\mathbf{d})$ と $p(\omega_{\text{nega}}|\mathbf{d})$ のいずれが大きいか?
- 評価表現辞書構築
 - ある対象に対する評価文書からキーワードとなる評価表現を提示する
 - 単純にはNaïve Bayesモデル中の素性の確率比をとれば良い



トピック

トピックによる評価表現の違い

この映画のラストシーンは何度見ても泣ける

ファンの回転音が大きすぎて泣ける

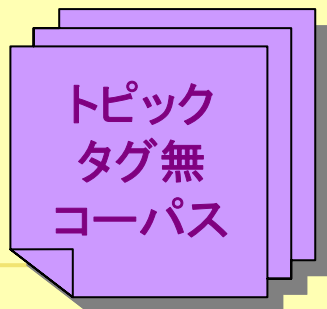
極性が異なる

■ 同じ表現でも意味が異なる場合があるため、トピックを考慮することで分類性能の向上が期待される

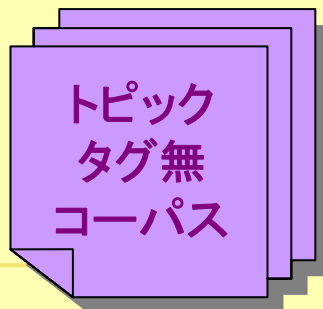
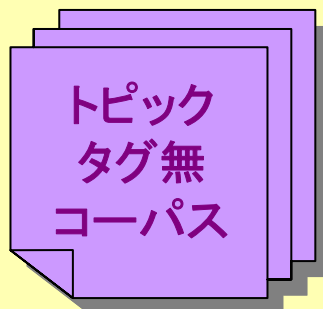
■ 極性が変わらないまでも、個々のトピックに特有の表現が存在するため、トピック依存の辞書が欲しい

トピックタグの有無とモデルの適用

訓練データ



テストデータ



トピック別ナイーブベイズ法



トピックモデル(initialized版)



トピックモデル

トピック別ナイーブベイズ法(トピックタグ有り)

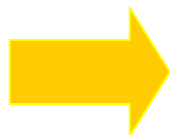
従来手法



コーパス全体の
ナイーブベイズモデル

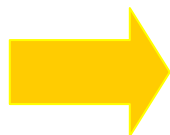
$$P_{all}(d | \omega) = \prod_{w \in d} p(w | \omega)$$

提案手法



DVDタグ内のナイーブベイズモデル

$$P_{seg-DVD}(d | \omega, DVD) = \prod_w p(w | \omega, DVD)$$



電化製品タグ内のナイーブベイズモデル

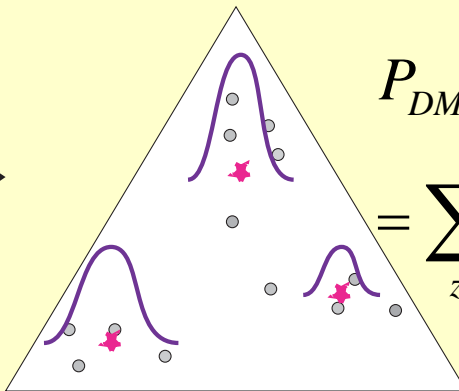
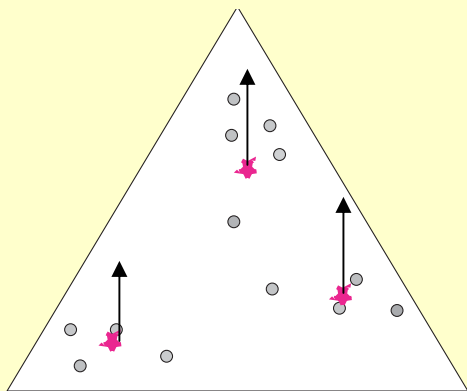
$$P_{seg-elec}(d | \omega, elec) = \prod_w p(w | \omega, elec)$$



※コーパス量の少ないトピックに対しては全体モデルとの線形補間

トピックモデルの適用 (トピックタグ無し)

Unigram Mixtures(UM) Dirichlet Mixtures(DM)



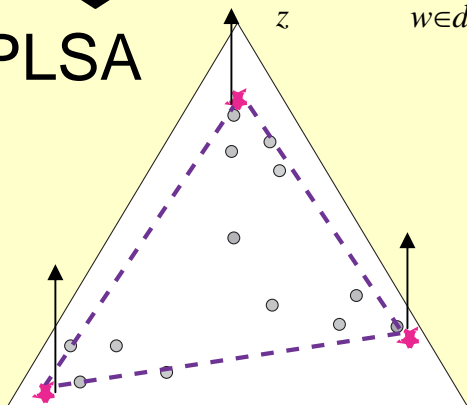
$$P_{DM}(d | \omega)$$

$$= \sum_z \lambda_z^{(\omega)} \frac{\Gamma(\alpha_z^{(\omega)})}{\prod_v \Gamma(\alpha_{zv}^{(\omega)})} \prod_v \theta_v^{\alpha_{zv}^{(\omega)} - 1}$$

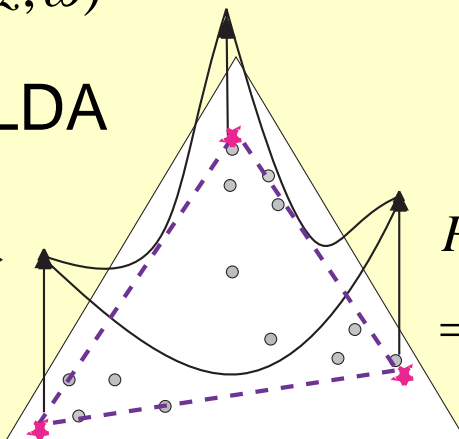
$$P_{UM}(d | \omega)$$

$$= \sum_z \lambda_z^{(\omega)} \prod_{w \in d} p(w | z, \omega)$$

PLSA



LDA

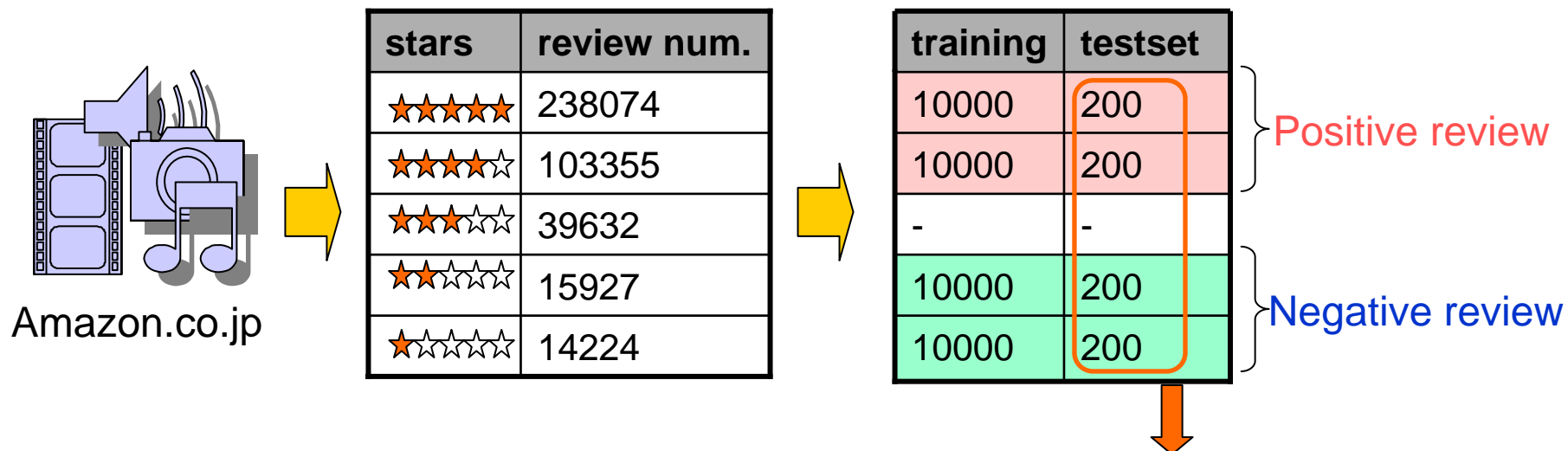


$$P_{LDA}(d | \omega)$$

$$= \int P_D(\lambda^\omega | \alpha) \prod_{w \in d} \sum_z \lambda_z^\omega p(w | z, \omega) d\lambda$$

- △ 語彙次元シンプレックス
- △ (虚線) トピック次元サブシンプレックス
- 1文書
- ★ トピックを強く表す点

実験条件



- development test: 各評点1000レビューずつ抽出
- トピックタグ: 全12トピック、非平衡
- 形態素解析: Chasen2.4.0(品詞付き)
- 素性: 1gram 素性
- 語彙: 全学習データを通して出現回数10回以上の計20,522単語

※レビューのタイトル、レビュアー名、その他メタ情報は実験データに含まない

トピック別ナイーブベイズ法と線形補間

トピック別ナイーブベイズ法による 評価文書分類精度

Model	Base line	divided	LI (dev)	LI (MAX)
Acc.	83.75	83.89	84.29	85.77

Baseline: 非分割ナイーブベイズ

divided : トピック別ナイーブベイズ

LI(dev): development testによって補間重みを最適化

LI(MAX): testsetを見て補間重みを最適化

線形補間 (LI):

$$P(d | \omega) = (1 - \lambda_{all}) P_{seg}(d | \omega) + \lambda_{all} P_{all}(d | \omega)$$

トピック別ナイーブベイズ法と線形補間 ～データサイズと線形重み～

category	DVD	electronics	music	book
train	5823	1121	5832	21918
dev. test	603	93	606	2178
testset	97	26	109	452
$\lambda_{all}^{dev.}$ (LI dev.)	0.6~1.0	0.2~0.4	0.0~0.2	0.7
λ_{all}^{MAX} (LI MAX)	0.6, 0.8, 1.0	1.0	0.4	0.2~0.4

※重み(λ)は0.0~1.0まで0.1刻みで11点

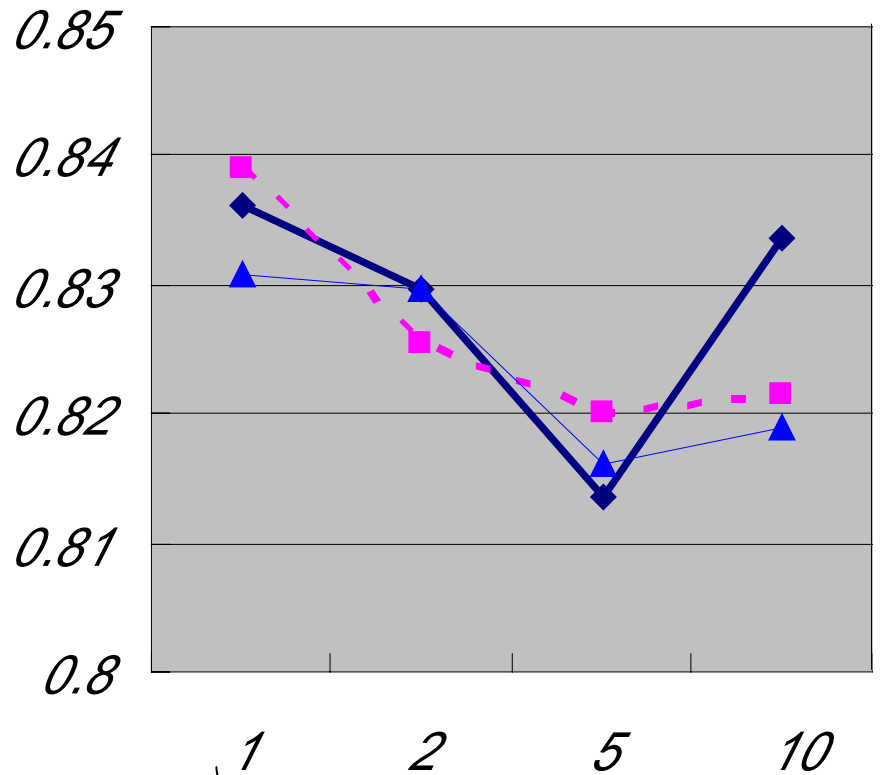
考察:

- データ数の少ないトピックで、 λ_{all} が大きくなることが望ましいと考えられるが、developmentテストの結果では必ずしもそうになっていない
- しかし、テストセットでの最適な重みにはそのような傾向が見られる
⇒ developmentテストセットのノイズが原因か？

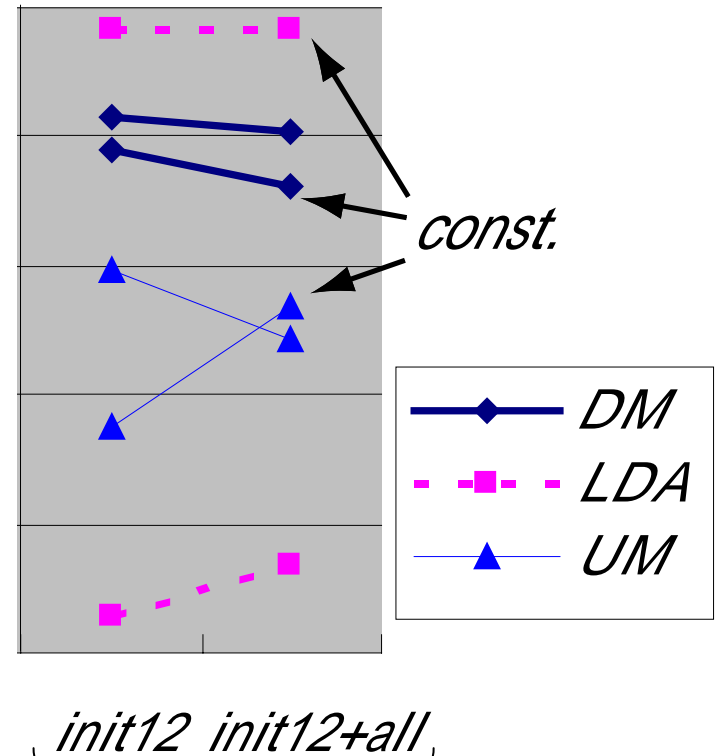
トピック別評価表現辞書

	全体	エレクトロニクス	音楽	本
Negative	アーシェ	不良	凡庸	まがい
	腹立たしい	返品	落胆	水増し
	水増し	原因	稚拙	侮辱
	粗末	なぜ	単品	しかるに
	ナメ	発生	無神経	駄本
Positive	ファンキー	一眼	風景	交差
	あたたか	楽しい	情景	なかでも
	愛しく	嬉しい	引き込ま	いつしか
	コンクール	疲れ	優し	解き明かし
	しっとり	驚き	ぴったり	待ち遠しい

トピックモデルによる評価文書分類



train: unsupervised
test: unsupervised



train: supervised
test: unsupervised

まとめと今後の課題

■ まとめ

- supervised/unsupervised双方の場合でのトピック情報を用いた評価文書分類を行った
- 線形補間を行うことで単純なトピック別ナイーブベイズ法でも精度向上
- トピックモデルをunsupervisedデータに対して適用した場合、効果はなかったが、supervisedデータを利用した場合、トピック別NBよりも良い結果を示した

■ 今後の課題

- discriminative trainingの適用
- トピック別辞書のうち、似たトピックについてはクラスタリング
- unsupervised dataからのトピック別評価表現辞書の構築