Hidden CRFsを用いた極性反転構造のモデル化

貞光九月　山本幹雄
(筑波大学)

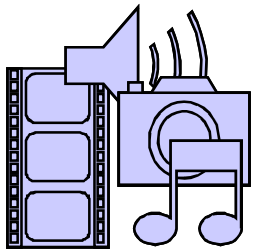
研究の目的･提案手法

- 目的1：評価文書分類(P/N文書分類)精度の改善
- 目的2：極性の反転している箇所の情報活用

- 提案手法: 極性反転構造を捉える
  - 単語レベル及び文レベルでの反転構造モデル化
  - シーケンス情報 ＋ 識別モデル ＋ 文書ラベル分類
    ＝Hidden CRFs

評価文書分類の概要

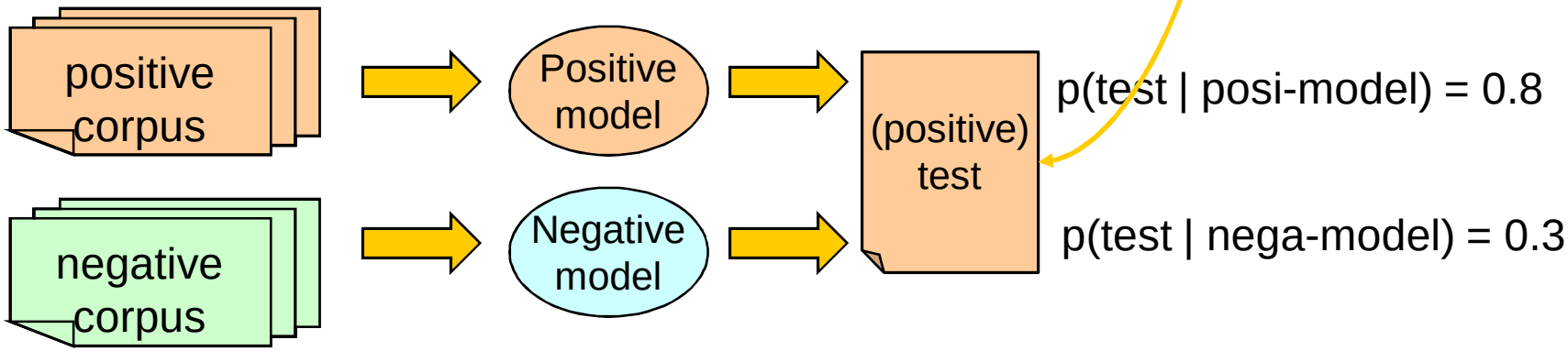
※10交差検定
※評点以外のメタ情報(トピック・著者等)
は全て削除



Amazon.com

stars	review num.
★★★★★	230000
★★★★☆	130000
★★★☆☆	39000
★★☆☆☆	15000
★☆☆☆☆	14000

training	testset
6120	680
6120	680
-	-
6120	680
6120	680



極性反転構造

ナイーブベイズによる分類誤り例
(”*razor*”**についてのレビュー** from Amazon.com)

I like this razor very much.
It gives a close shave without damaging the skin.
However, the only drawback is that you must constantly be holding in the “On” button.
My last razor had a button that when it was switched on, it stayed on.

極性が反転する箇所に現れやすい表現(企業からすれば有益な箇所)

極性反転の要因:
-*部分的な批判*
-*他商品への比較*
-*例示*
-*引用* *etc.*

Hidden CRFs

CRFsのHMMsに対する利点
- (系列の) 識別モデル → 尤度最大ではなく 条件付確率最大
- 素性設計が自由

$$P_{CRF}(q|d;\lambda) = \frac{\exp\{\lambda \cdot f(q,d)\}}{\sum_q \exp\{\lambda \cdot f(q,d)\}}$$

(Lafferty et al. '01)

$$P_{HCRF}(\phi|d;\lambda) = \frac{\sum_q \exp\{\lambda \cdot f(\phi,q,d)\}}{\sum_{\phi'}\sum_q \exp\{\lambda \cdot f(\phi',q,d)\}}$$

(Quattoni '04, Gunawardana '05)

q : ラベルシーケンス
λ: 素性重み
f : 素性関数
φ:極性ラベル

HCRFsのHMMsに対する利点:
- CRFsの利点を両方引き継ぐ

HCRFsのCRFsとの違い:
- ラベル系列 q を推定するのではなく文書全体のラベルφを推定する

先行研究

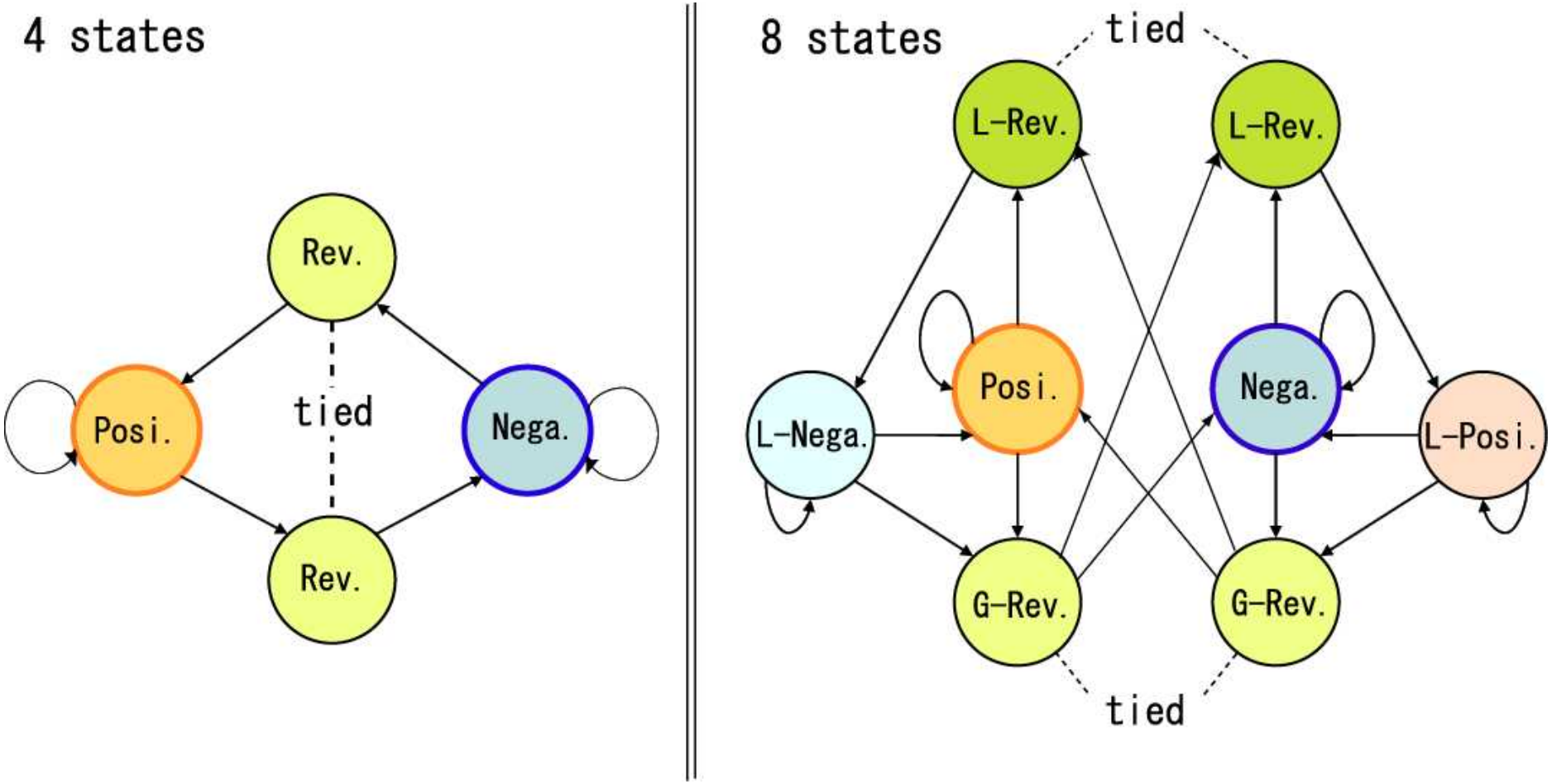
- Mao and Lebanon(2007)
  CRFsによって文ラベルを付与した後 regression からの評価文書分類
  ⇒文毎の極性ラベルが必要。Hidden CRFsでは不要

- McDonald et al.(2007)
  文書と文のように粒度の異なる対象を同時に極性分類
  シーケンシャルな情報も用いる
  ⇒文毎にラベルが必要。Hidden CRFsでは不要

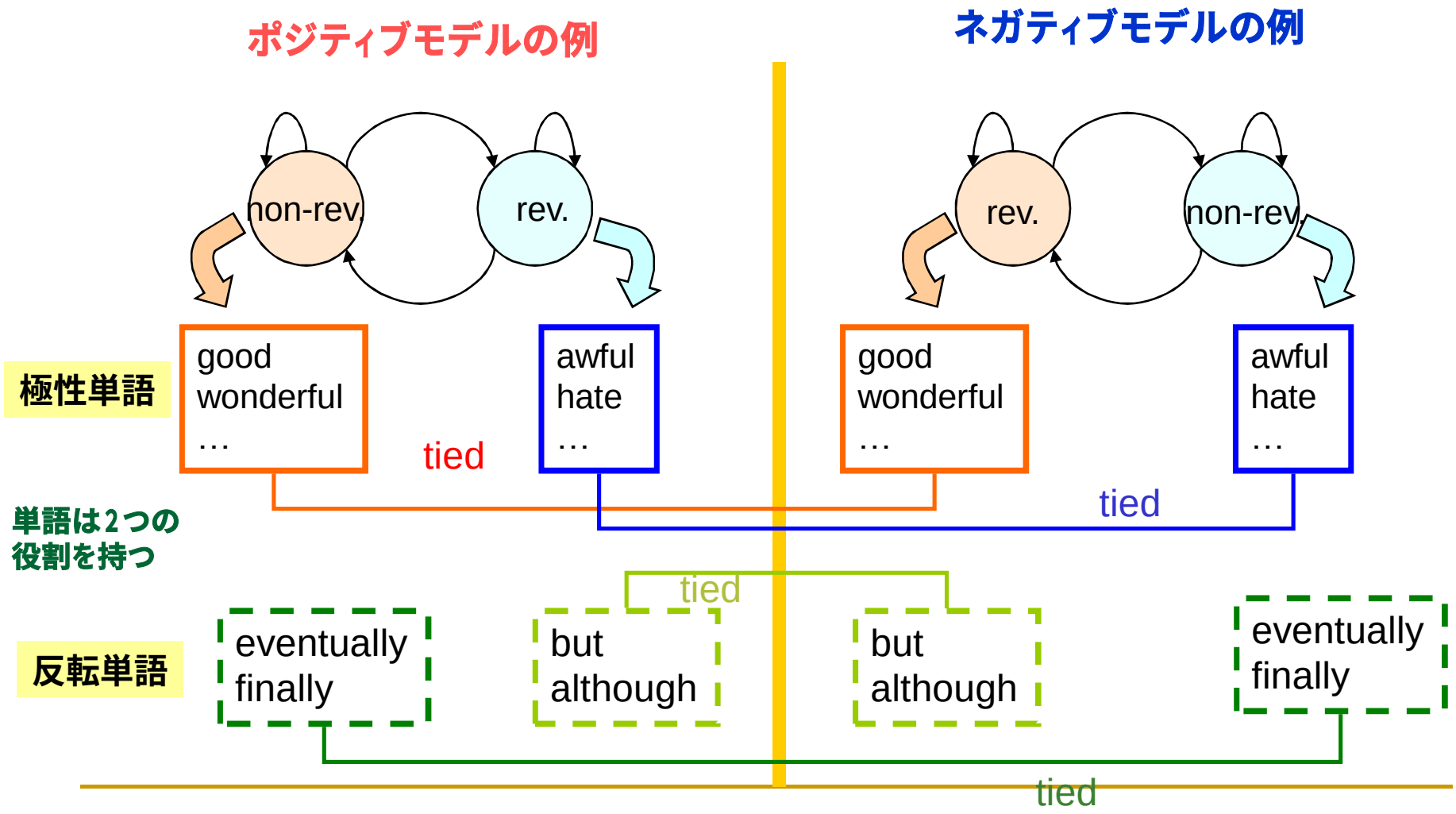
- Ikeda et al.(2008)
  単語レベル、文レベルにおける極性反転に注目した評価文分類
  ⇒シーケンシャルな情報や極性表現辞書外の単語を考慮しない

単語遷移モデル

This is not a bad book but I don't like the characters.



文遷移モデル

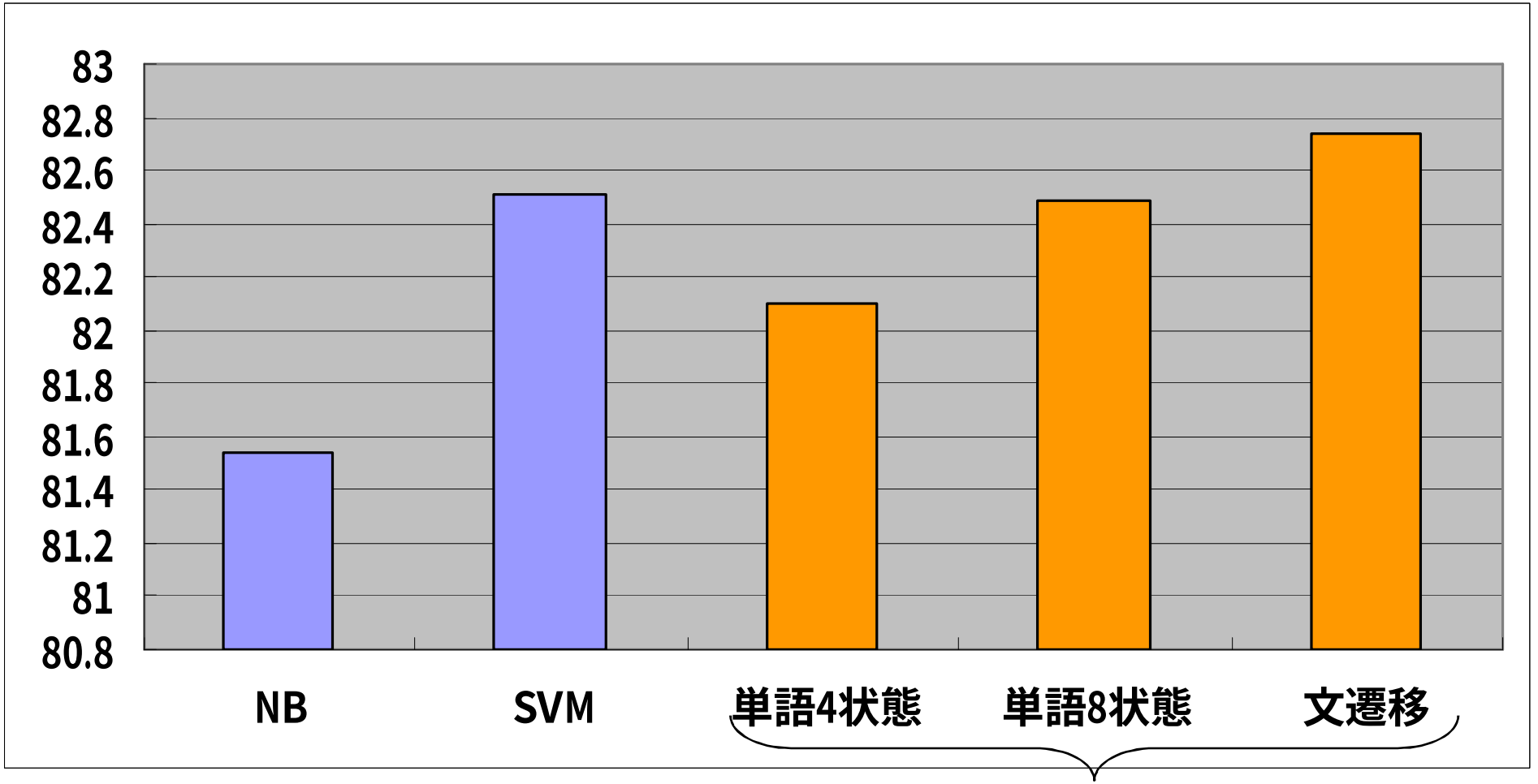

ポジティブモデルの例
ネガティブモデルの例
non-rev.
rev.
rev.
non-rev.
極性単語
good
wonderful
…
awful
hate
…
good
wonderful
…
awful
hate
…
tied
tied
単語は2つの
役割を持つ
tied
反転単語
eventually
finally
but
although
but
although
eventually
finally
tied

実験条件

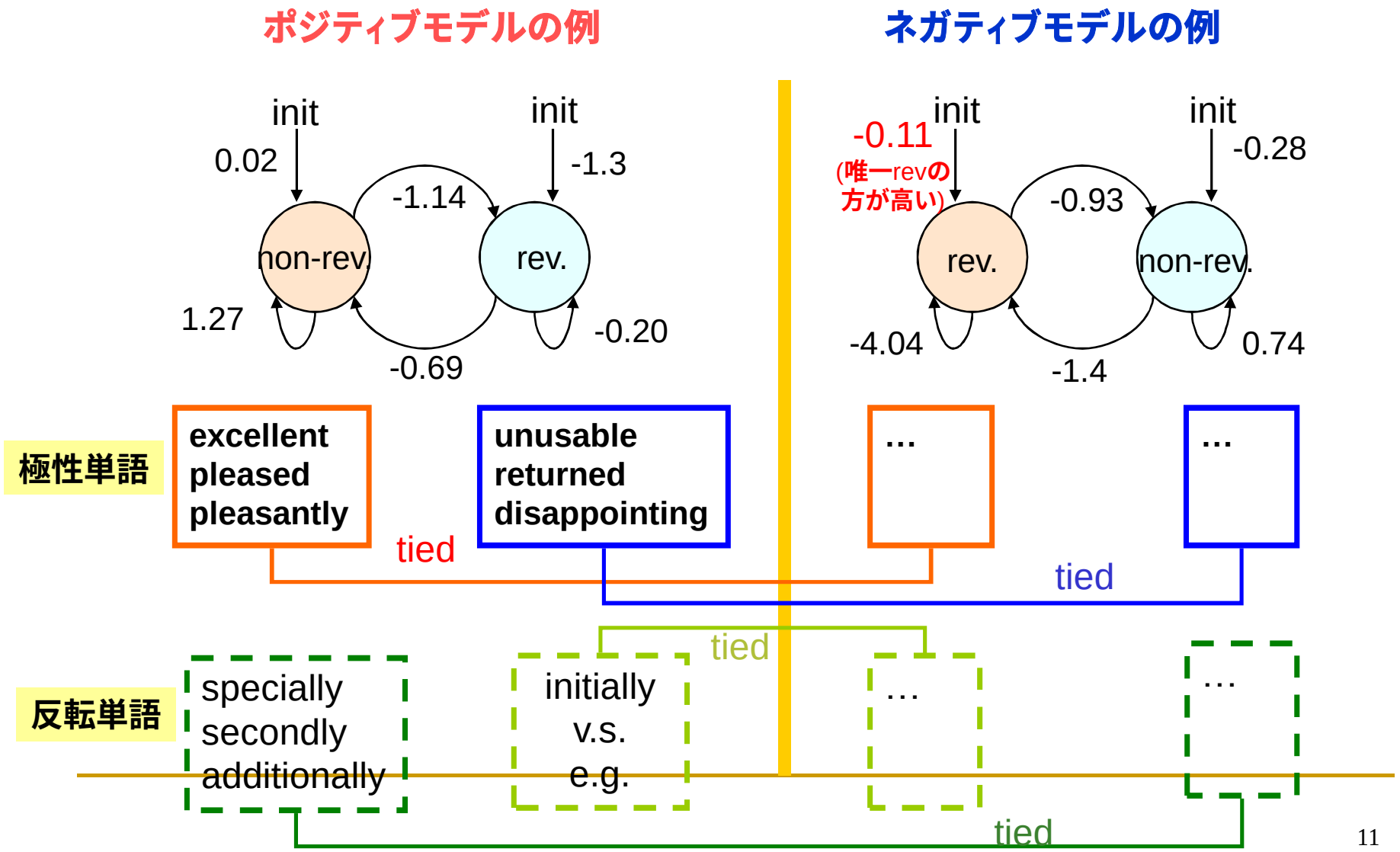
- Amazon.comよりPosi/Negaそれぞれ6800レビュー
- 語彙: df20以上の単語 4051単語

- ベースライン
  - NB: Naïve Bayes
  - SVM: SVM light (gaussian kernel)
- Hidden CRFs学習
  sigmoid損失関数＋最急降下法 (MCE)

評価文書分類結果


83
82.8
82.6
82.4
82.2
82
81.8
81.6
81.4
81.2
81
80.8
NB
SVM
単語4状態
単語8状態
文遷移
Hidden CRFs

文遷移モデルの学習結果



まとめと今後の課題

- **まとめ:**
  - **極性反転構造を捉えることには成功**
  - **単語レベル遷移･文レベル遷移いずれのモデルも SVMと同等程度の性能**

- **今後の課題:**
  - **リッチなモデル化が識別に結びついていない ⇒他の損失関数及び学習法導入**
  - **極性反転箇所についての情報抽出**