言語処理学会第１３回年次大会（NLP-2007）
2007/03/20-22

漢字情報を利用した評価表現辞書の拡張法

東京工業大学 統合研究院
イノベーションシステム研究センター

乾 孝司　村上 浩司　橋本 泰一

背景と目的

- 評価表現辞書
  - 単語が 肯定/否定 のどちらの評価を表しやすいか
    - 例: 良い – 肯定，悪い – 否定，おいしい – 肯定

- 特定の評価が導かれやすい事柄も含まれる
  - 例: 合格 – 肯定，渋滞 – 否定

背景と目的

- 社会課題抽出への利用
  - 新聞記事から
    社会課題に関連するキーワードを見つけたい

例: 問題，医療過誤　→ 課題発生に関連：否定
例: 改善，再発防止策 → 課題解決に関連：肯定

背景と目的

- 欲しいキーワードは漢字列
  - 評判分析からみれば特殊
  - 既存手法で構築された評価表現辞書ではカバーが難しい

例: 医療過誤(否定)
再発防止策(肯定)

- 本研究の目的
  - 漢字列からなる語の評価極性判定
  - ３値分類
  - 肯定(p)/否定(n)/中立(e)

前提

- リソース
  - 生コーパス(新聞記事)
    - ここから社会課題の関連キーワードを見つけたい
  - 中小規模の評価表現辞書
    - 既存手法によって構築
    - ２値，肯定/否定

漢字列の評価極性判定: 基本的なアイデア

- 核要素
  - 語の極性に影響を与える一部の漢字系列
  - 核要素は一定の極性を導く(例: 助が核要素)

人命救助
援助活動

- 核要素 と 語の極性 の規則性を学習する

＊助＊ → 肯定

基本法(1/6)

- 決定リスト(decision lists，DL)
  - 評価表現辞書を教師情報とする
  - 評価表現辞書に現れない漢字を考慮できない
- ブートストラップ
  - コーパスから漢字を補う

基本法(2/6)

- 決定リスト(decision lists，DL)
  - 核要素→評価極性クラス (信頼度)

助 → 肯定 (0.7)

  - 証拠(核要素)候補の生成
    - 漢字ユニグラム
    - 漢字バイグラム

語	証拠候補
災害	災，害，災害

基本法(3/6)

- 決定リスト(decision lists，DL)
  - P(クラス｜証拠)に従う
  - *k* 番目の証拠に対するクラスと信頼度は…

事例生起を表す

$$f_j = \sum_{m \in M_j^k} occur(m)$$

occur(m) = 1

$$class_k = \arg\max_j \frac{f_j + \delta}{\sum f_j + \delta|C|}$$

$$conf_k = \max_j \frac{f_j + \delta}{\sum f_j + \delta|C|}$$

M_j^k ： ***k*** 番目の証拠と照合し クラスが ***j*** となる事例の集合

$j \in \{p, n, e\} = C$

δ ：スムージング・パラメータ (＝ 0.5で実験)

基本法(4/6)

- 決定リスト(decision lists，DL)
  - デフォルト規則
    - 中立を割り当てる

$$true \quad \rightarrow \text{中立}\,(\lambda)$$

λ ：信頼度の下限

λ ＝0.5で実験

基本法(5/6)

ブートストラップ(一般的な手続き)

1．学習：評価表現辞書を利用して決定リストを学習する

2．未知データへ適用：学習された決定リストに基づいて，
コーパスから抽出した漢字列の評価極性を判定する

3．再学習：判定結果を擬似教師ありデータとみなし，
これを教師ありデータに加えて，決定リストを再学習する

4．終了判定：ラウンドの前後で同じ決定リストが学習されれば
手続きを終了する．そうでなければ2．へ戻る

基本法(6/6)

- ブートストラップ(操作変更)
  - 擬似教師ありデータ
    - 誤りが含まれる → 誤りの影響を減らしたい
    - 各事例の信頼度は適用された規則の信頼度に従う
    - 規則の信頼度に従って事例の頻度をディスカウントする

$$occur(m) = \begin{cases} 1 & m \in \text{評価表現辞書} \\ 0.1 \times conf & \cancel{m \in \text{コーパス}} \quad \text{otherwise} \end{cases}$$

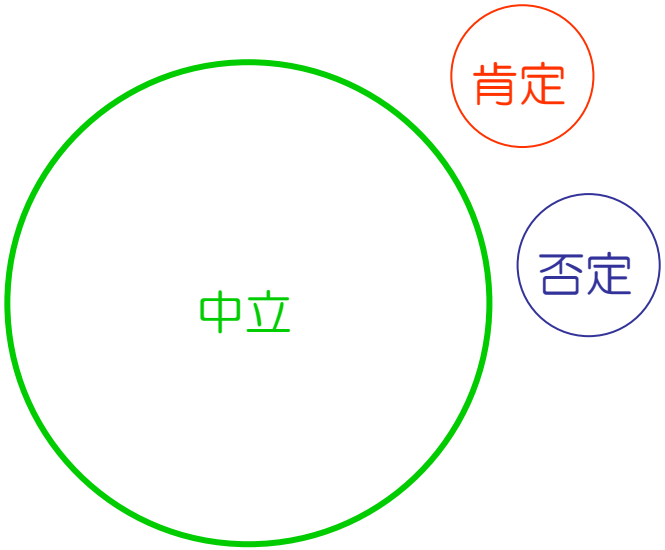
定率減

事例mのクラスを判定する際に
適用された規則の信頼度

データの特性

- クラス不均質性
  - 肯定/否定となる漢字列は一部
  - 大多数は中立である

  - 学習が中立に偏る
  - 基本法を改良



基本法→改良法

- 改良Ａ：
  - 中立クラスを割り当てる規則を削除する
    - 中立事例の数を減らす
    - デフォルト規則は削除しない
- 改良Ｂ：
  - 中立事例は頻度をディスカウント
    - 中立事例の影響を減らす

$$occur(m) = 0.1 \times \beta \times conf \qquad 0 \leq \beta \leq 1$$

実験

- コーパス
  - 新聞記事から漢字列を得る(２０万件)
    - 以下の条件を満たす漢字列を無作為に抽出
      - 条件１．すべての文字が常用漢字である
      - 条件２．６文字以下である
      - 条件３．ChaSenで解析した際，すべての形態素の品詞が「固有名詞」等以外の名詞か，未知語である

- 評価用の漢字列
  - ２０万件の一部1,000件 (p:220/n:332/e:448) 平均文字数: 4.5
  - 被験者２名間のκ値　0.67

- 評価表現辞書
  - [高村ら 2005]の後，人手で修正
  - 573語 (p:318/n:255/e:0) 平均文字数: 2.0

実験

- ベースライン(BL)
  - BL1：評価表現辞書内の語と照合

    満足 → 肯定
  - BL2：初期辞書内の語を一文字づつ漢字にばらして，それとの照合

    満 → 肯定

    足 → 肯定
- 評価尺度
  - 正解率
    (正しいクラスを出力した割合)

実験結果

Baseline1	0.598
Baseline2	0.577
DL	0.654
DL+bootstrapping（基本法）	0.752
DL+bootstrapping（改良法）	**0.798** ($\beta = 0.2$)

まとめ

- 評価表現辞書の拡張
  - 漢字列の評価極性判定 (例：医療過誤)
    - 肯定/否定/中立の３値分類
  - 基本法
    - 決定リスト + ブートストラップ
  - データのクラス不均質性 (class imbalanced)
    - 基本法を改良
    - 正解率を改善 (0.752→0.798)

今後の課題

- 学習データ：仮名漢字混じり語の利用
- 規則の証拠：文字位置，語境界情報の利用
- 漢字列の外部情報の利用[那須川ら 2004]
- Transductive SVMの検討
- 先行研究との比較

Thank you!