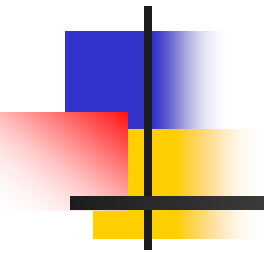


誹謗中傷を表す文の自動検出

長岡技術科学大学

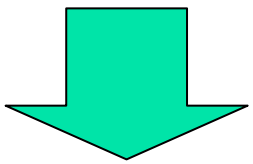
石坂 達也, 山本 和英

背景と目的

- Web上には他者を誹謗中傷する書き込みがある
  - ネットいじめと呼ばれる社会問題となっている
  - 最悪の場合，自殺を引き起こしている

現状

人手による巡回　負担が大きい



目的

Web上の誹謗中傷の自動検出

誹謗中傷(悪口文)の定義

- 他の情報を必要としない侮辱や誹謗中傷している単語,句を含む文

皮肉などは対象外
(例) お前天才じゃね？

(例)

・あの政治家死ね

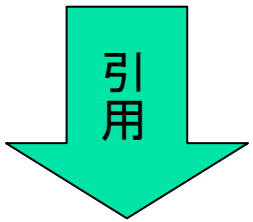
・奴らはバカな暇人野郎

手法の方針

- 誹謗中傷文は悪口単語の影響が大きい
- 誹謗中傷は人への評価ともいえる

評価情報を分析するための手法を引用する

- ある単語が好評表現/不評表現かの判定する手法



単語が悪口単語/非悪口単語かを判定して文分類

単語の悪口度計算

- SO-PMI Algorithmを使用
  [Wang and Araki, 2007]

単語wとpositiveの検索ヒット数

negativeの検索ヒット数

$$C(w) = \frac{hits(w, positive) * hits(negative)}{hits(w, negative) * hits(positive)}$$

重み

positiveの検索ヒット数

$$f(\alpha) = \alpha * \log_2 \frac{hits(positive)}{hits(negative)}$$

$$SO(w) = \log_2[C] + f(\alpha)$$

SO-PMIの概要

- 共起情報を利用して単語wの極性を判定
  - positiveと共起しやすいなら悪口単語
  - negativeと共起しやすいなら非悪口単語
  - 検索ヒット数の差を補正するための$f(\alpha)$

- WangらはSO-PMIを用いて好評文/不評文に分類
  - 好評文が78%, 不評文が72%の精度で分類できた

positive,negative単語の選択

- 評価表現を対象とする場合
  - positive＝素晴らしい，好き，楽しい，満足
  - negative＝不良，悪い，欠点，最悪

    極性が逆の単語を使用(好評⇔不評)
- 悪口単語を対象とする場合
  - positive＝悪口単語(ウザい，死ね，キモい)
  - negative＝悪口単語と極性が逆の単語

    悪口単語の逆とは…？
    - 褒め言葉？非悪口単語？

単語の極性計算

悪口単語にはウザいを使用

- 褒め言葉を使用
  - 可愛い, 素敵, イケメン

愚民	-3.688
派閥	-3.413
売ら	-3.250
兆	-3.190
廃止	-3.162

- 非悪口単語を使用
  - 机, チューリップ

消えろ	-6.697
失せろ	-6.667
ジャニヲタ	-6.371
死ねよ	-6.370
メンヘラー	-6.364

- negativeには非悪口単語を使用する
  - 今回はとりあえず「机」で行う。

SO-PMIの結果の例

SO-PMIが小さい単語

- 消えろ -6.697
- 失せろ -6.667
- ジャニヲタ -6.371
- 死ねよ -6.370
- メンヘラー -6.364
- 鼻糞 -6.175
- イラネ -6.172
- ツマラン -6.143
- カワイソス -6.108

SO-PMIが大きい単語

- 我 7.702
- 充分 7.744
- 媒体 7.801
- 台北 7.841
- 能 7.863
- 招か 7.942
- 保有 8.026
- 有意 8.311
- 威 8.649

SO-PMIが小さいとき，悪口単語が多い
さらに，悪口生起単語も多い

SVMを用いた分類実験

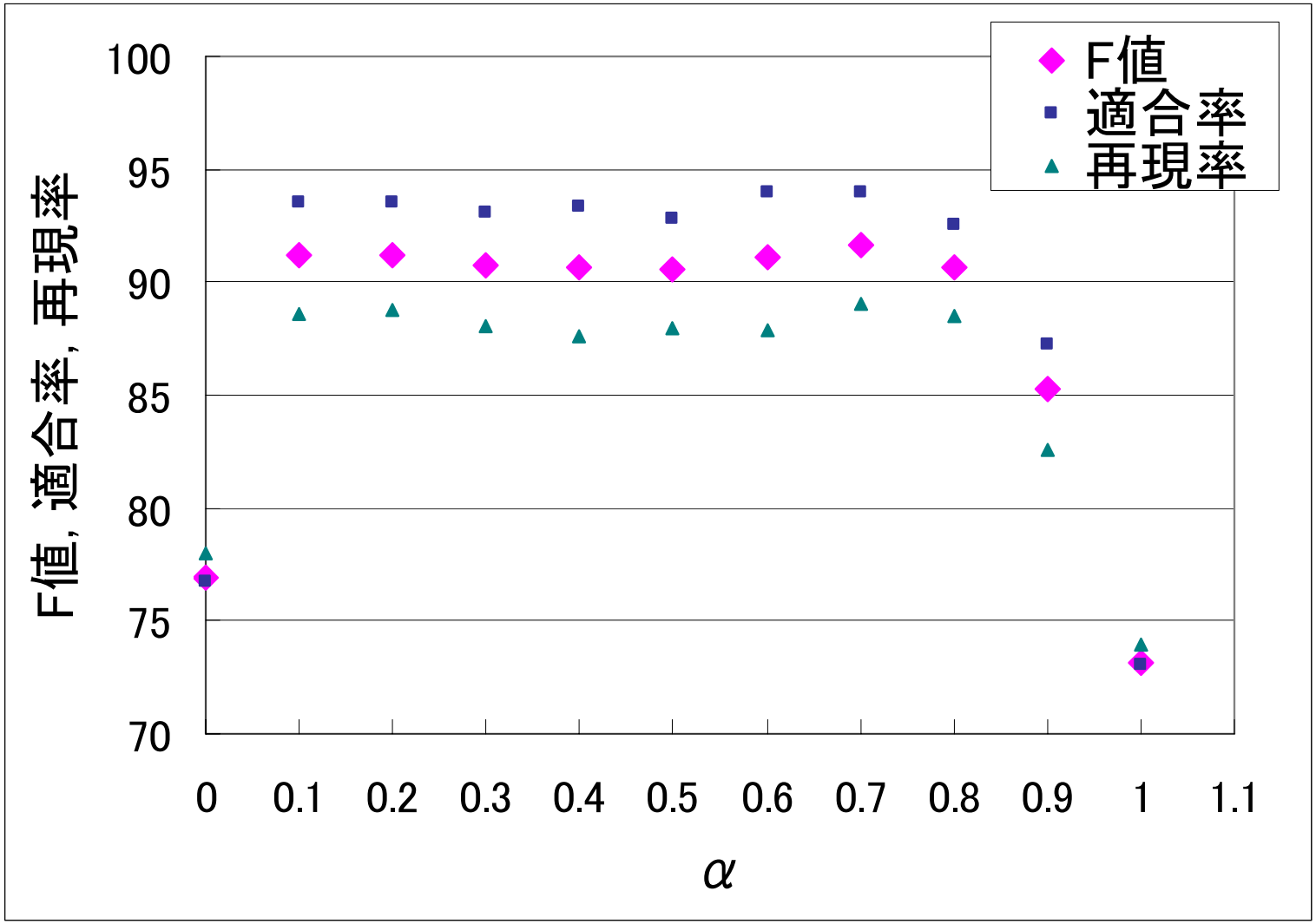
- 入力文が悪口文/非悪口文を判定
- TinySVMを使用
- 学習データ＆評価データ
  - 「2ちゃんねる」から収集
  - 被験者３人により作成
  - ２人以上一致した評価を使用
  - 5分割交差検定
    - 悪口文1400文, 非悪口文1400文
      - 各380文は評価データとして使用

素性と特徴量とα

- 素性
  - 文に含まれる形態素(記号除く)
- 特徴量
  - 各単語のSO-PMI
- α
  - SO-PMIに使われるαによって精度が変動
  - αは0～1.0(0.1刻み)

実験結果1-1

F値，適合率，再現率はそれぞれ交差検定における平均値



実験結果1-2

- αが0.7の時、F値が91.64で最高
- 悪口単語を含む文は分類精度 高
  - （例）お前みたいな認識の馬鹿は死ねば良いと思う。
- 悪口単語が悪口として使われない文は…
  - 状況によって分類精度が異なる
  - （例）糞かっこいい―
  - （例）あのパンはバカうまいな
- 比喩のような表現の悪口文は分類精度 低
  - （例）お前はサル以下の脳みその持ち主だな

実験結果2 -αと精度の関係-

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
testset A	90.9	91.1	90.2	89.6	89.3	91.1	92.1	92.3	88.6	72.0
testset B	90.4	90.9	90.5	91.3	91.3	90.7	90.5	89.1	84.3	74.5
testset C	92.3	92.1	91.4	90.9	90.0	90.9	91.6	89.5	85.4	69.6
testset D	92.3	92.5	92.1	92.0	92.3	92.7	93.6	92.5	86.6	75.2
testset E	90.2	89.6	89.5	89.5	89.8	90.0	90.4	89.8	81.4	74.6

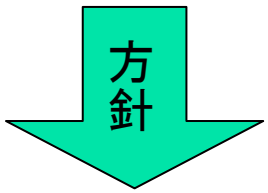
- テストセットによって最適なαが異なる
- 入力に応じてαを変化するようにするべきか

今後 取り組みたい事

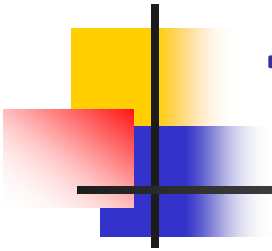
- 比喩表現を使用している悪口文の検出
  - 単語単体への着目は効果が小さい

単語同士の繋がりで変化する意味の同定

- 入力文に応じたαの設定



入力文の特徴の同定



まとめ

- 誹謗中傷を表す文の検出することが目的
- 今回は評価表現分類の手法を使用
- 単語が悪口単語かどうかを計算した
- SVMの分類結果は最高でF値91.64
- 比喩を使う悪口文は抽出できない
  - 単語の繋がりによる意味の変化を同定したい