言葉の理解メカニズムー不適切表現理解からー

フロンティア研究センター（青江研究室）

システム概要

掲示板データ

582 : 馬鹿が : 08/07/12 22:41:57 ID:〇〇
そんな批判するなら利用したことないんだよな？
利用したことあるのに批判なら，お前頭おかしいぞ

212 : 少年法により名無し : 08/07/12 22:46:45 ID:〇〇
徳島駅で〇〇（人名）を殺す
：

記事番号・投稿題目・本文の取得

記事番号　投稿題目　本文

582 : 馬鹿が : 08/07/12 22:41:57 ID:〇〇
そんな批判するなら利用したことないんだよな？
利用したことあるのに批判なら，お前頭おかしいぞ

212 : 少年法により名無し : 08/07/12 22:46:45 ID:〇〇
徳島駅で〇〇（人名）を殺す
：

不適切表現検査・犯罪表現検査

不適切表現　犯罪対象

582 : 馬鹿が : 08/07/12 22:41:57 ID:〇〇
そんな批判するなら利用したことないんだよな？
利用したことあるのに批判なら，お前頭おかしいぞ

212 : 少年法により名無し : 08/07/12 22:46:45 ID:〇〇
徳島駅で〇〇（人名）を殺す
：

荒れ度計算・危険度分類

【荒れ度計算】
総投稿数における
不適切投稿の
割合から算出

【危険度分類】
犯罪投稿における
犯罪対象の数から分類

判定結果

【荒れ度】
掲示板の状態

＜掲示板1＞
荒れ度：25
状態：少し荒れている
＜掲示板2＞
荒れ度：11
状態：荒れていない
＜掲示板3＞
荒れ度：56
状態：荒れている
：

【危険度】
犯罪の種類・対象

＜記事番号1＞
危険度：2
犯罪の種類：放火
犯罪対象：場所
＜記事番号45＞
危険度：3
犯罪の種類：殺人
犯罪対象：人物・場所
：

背景・目的

背景

2ちゃんねる　Yahoo! 掲示板

電子掲示板の利用

不適切（誹謗中傷や差別など）な書き込み
- プロバイダ責任制限法による処置

犯罪を予告（殺人や爆破など）する書き込み
- 逮捕や書類送検といった処分

管理者や警察組織の負担が増加

既存システム

掲示板の監視を支援するシステムの提供

既存システムの特徴
- 禁止単語を検出し，投稿を監視（サイバーコップス）
- ユーザからの通報による犯罪予告の監視（予告.in）

問題点
- 組み合わせで意味を持つ表現は検出できない
  例） お前頭悪い　→　（お前 ＋ 頭 ＋ 悪い）
- 優先的に監視すべき掲示板や投稿は分からない
  例） 悪質な投稿を含む掲示板：掲示板1，3，4，・・・
  犯罪予告を含む投稿：記事番号3，45，56，・・・

目的

組み合わせ表現を検出し，掲示板や投稿の状態を判定

特徴1：二種類の表現を検出
- 不適切表現（誹謗中傷や差別的な語句）
- 犯罪表現（殺人や爆破，対象を表す語句）

不適切表現
害人（当て字），禿（独自語）
犯罪表現
火事にする（組み合わせ）

特徴2：掲示板の荒れ判定
- 掲示板全体の荒れ具合を判定
  ⇨ 優先的に監視すべき掲示板の把握

掲示板1：荒れていない
掲示板2：荒れている
掲示板3：荒れていない

特徴3：投稿の危険判定
- 犯罪を予告する記事の危険性を判定
  ⇨ 犯行の具体性が高い投稿の把握

記事番号24：危険度1
記事番号73：危険度4
記事番号89：危険度1

不適切表現検査と荒れ判定

不適切表現

閲覧者にとって不適切な印象を与える表現や投稿

不快 ・・・誹謗中傷や差別的な言葉（5種類）
例） おまえ頭悪いだろ，DQNは喋らなくていいよ

非文 ・・・不自然な構成の文章や投稿
例） えgせsがが，＝｜＝（（‘）’＆（

重複 ・・・同一の投稿が複数回連続
584 : 〇〇 : 08/07/12 23:12:39 ID:〇〇 ＜1＞
あげ
585 : 〇〇 : 08/07/12 23:12:41 ID:〇〇 ＜2＞
あげ
586 : 〇〇 : 08/07/12 23:12:42 ID:〇〇 ＜3＞
あげ

検査に必要な情報を取得
例）文 「こいつバカ」
→ こいつ/バカ
読み：こいつ/ばか
品詞：人称代名詞/普通名詞
形態素数：2

不適切表現検査
形態素解析
不快検査　非文・重複検査
不適切投稿の集計

登録されている情報と照合する ⇔ 不快検出ルール

文：「こいつ（人称代名詞）/バカ（ばか）」

照合要素	意味情報
人称代名詞	人物
ばか，あほ，・・・	苛立ち用語

不快照合パターン	意味分類
＜苛立ち用語＞	【苛立ち】
＜人物＞＋＜苛立ち用語＞	【誹謗中傷】

照合要素：約450
照合パターン：約130

非文検査
- 文字数，文字構成，形態素の平均長から判定
  非文1「gs/ガ/ガ/g/ジ/t」形態素数：6
  文字数：7（閾値は6以上）
  文字構成：2種（文字が2種類以上）
  形態素の平均長：1.2（閾値は1.5以下）
- 1行の文字数，行の連続から判定
  非文2「あ（1）　文字数1（閾値1以下）
  ：
  が（8）」 行数8（閾値6行以上）

重複検査
- 同一投稿が3回以上連続した場合に
  ≪重複≫と判定
  A内容　A内容　A内容

全投稿記事に対して検査
不適切表現を含む投稿（不適切投稿）の数を算出
※重複はまとめて一つとする

荒れ判定部
荒れ度計算

荒れ度計算
掲示板の荒れ具合を表す判定値
総投稿数における不適切投稿数の割合から算出
例）掲示板1　総投稿数：251
不適切投稿数：90

$$荒れ度 = \frac{不適切投稿数}{総投稿数} \times 100 = \frac{90}{251} \times 100 \fallingdotseq 35.9$$

算出結果の範囲による掲示板の状態
0～9.3　：荒れていない
9.3～34.9：少し荒れている
34.9～　：荒れている
⇨ 閾値により範囲を設定

犯罪表現検査と危険判定

犯罪表現

犯罪を予告する表現と犯罪の対象を表す語句

殺人・暴力 ・・・殺意や暴力性が感じ取れる表現
例） 殺す，頃してやる（当て字），殴る

爆破・放火 ・・・爆破や放火が感じ取れる表現
例） 爆弾を仕掛けた，火事にしてやる

隠語 ・・・犯罪取引につながる特定の意味を持つ語句
例） レンコン（拳銃），ガンジャ（大麻）

犯罪対象 ・・・人物，場所，日付を表す表現
例） 〇〇（人名）を殺す，東京（地名）を爆破する

検査に必要な情報を取得
例）文 「徳島駅を爆破」
→ 徳島/駅/を/爆破
読み：とくしま/えき/を/ばくは
品詞：地名/地名接尾語/格助詞/普通名詞
形態素数：4

犯罪表現検査
形態素解析
殺人・暴力検査
爆破・放火検査
隠語検査
犯罪対象検査
犯罪投稿の集計

登録されている情報と照合する ⇔ 犯罪検出ルール

文：「徳島（地名）/駅（地名接尾語）/を/爆破（ばくは）」

照合要素	意味情報
ばくは，はかい・・・	爆破用語
ばくだん，だいなまいと・・・	爆破道具

犯罪照合パターン	意味分類
＜爆破用語＞	【爆破】
＜爆破道具＞＋＜動詞の語幹＞	【爆破】

照合要素：約200
照合パターン：約120

登録されている情報と照合する ⇔ 犯罪対象検出ルール

文：「徳島（地名）/駅（地名接尾語）/を/爆破（ばくは）」

照合要素	意味情報
地名	地名/地名
地名接尾語	地名/地名接尾語

犯罪対象照合パターン	意味分類
＜地名/地名＞＋＜地名/地名接尾語＞	【場所】
＜組織名/組織名＞	【場所】

照合要素：約1100
照合パターン：約200

全投稿記事に対して検査
犯罪表現を含む投稿
（犯罪投稿）の記事番号を保存

危険判定部
危険度分類

危険度分類
投稿の危険性を表す判定値
犯罪投稿における
犯罪対象の数（種類の多さ）から段階的に分類
例）記事番号47　対象数：1
犯罪対象　場所：徳島駅
→ 危険度2

犯罪対象	犯罪対象の組み合わせ	危険度
0種類	犯罪対象無し	1
1種類	人物or場所or日付	2
2種類	人物・場所or人物・日付or場所・日付	3
3種類	人物・場所・日付	4

評価

荒れ判定の精度評価実験

- 実験データ：2ちゃんねる，Yahoo!掲示板，学校裏サイト計60個
  - “荒れている”，“少し荒れている”，“荒れていない”
    掲示板を各20個ずつ使用
- 評価方法
  - 掲示板の荒れ具合を正しく判定できていれば正解とし，正解率で評価
    ⇨ 判定が正しいかどうかの2値判断であるため

荒れ度判定結果

内訳	荒れている	少し荒れている	荒れていない	総合
掲示板数	20	20	20	60
正解数	18	16	12	46
正解率(%)	90	80	60	76.7

- 総合的な荒れ判定精度は76.7%という結果
- 比較参考として，従来研究[一藤ら2006]では2ちゃんねるに対する判定精度が65%という結果

⇨ 1. 他の掲示板を含めても一定の判定精度を得られた
2. 荒れている掲示板に対する判定精度は良好であり，優先的に監視すべき掲示板を把握できる

危険判定の精度評価実験

- 実験データ：
  - クローズデータ：投稿記事数100 [犯行記事一覧]
    - 過去に逮捕や通報がされた投稿記事や文（約250文）
  - オープンデータ：投稿記事数200 [予告.in]
    - 予告.inに投稿された投稿記事や文（約540文）
- 評価方法
  - 犯罪表現と犯罪対象を全て適切に検出し，正しく判定できていれば正解とし，正解率で評価
    ⇨ 判定が正しいかどうかの2値判断であるため
  - 犯罪対象の未検出または過剰検出が存在する場合を準正解としている

危険判定結果

内訳	クローズデータ	オープンデータ
判定記事数	100	150
正解数	73	103
準正解数	17	25
正解率(%)	73.0	68.7
正解＋準正解(%)	90.0	85.3

- 正解率はクローズ，オープン共に70%前後の結果
- 準正解を含めると両データ共に80%以上の判定精度

⇨ 利用ユーザが少ない個人サイトなどに対しても，実用可能な精度を得られている