評判情報の検索における隠語的造語法の応用

筑波大学大学院図書館情報メディア研究科
木村友秋　藤井敦

06T4073R 三上健太

はじめに

- Web上の評判情報
  - → ・企業側:自社商品改善のための情報
  - ・消費者側:商品の良し悪しを判断するための情報
- 評判情報の検索に関する研究
  - → ・レビュー記事や掲示板など評判が高密度なテキストを対象
  - ・一般的な情報も対象

しかし

- 評価表現が明記されていない評判も存在

例:「近所の電気屋は未だに定価で商品を売っている」

→「近所の電気屋は高い」

はじめに(2)

- Web上の評判
  - → 隠語で表すことがある
    - 例：ソフ〇バンク→ソフトバンク

- 評判情報のうち批判的な情報に焦点をあてる
- 批判がかかれたページの検索手法を提案
- 隠語を用いる

隠語の分類

- ❑ 音節省略　：　警察→サツ
- ❑ 音節転換　：　声→エコ
- ❑ 形状類似　：　針→松葉
- ❑ 色彩類似　：　カラス→墨
- ❑ 連想　　：　はさみ→カニ
- ❑ 動作　　：　入浴→ザブン
- ❑ 比喩　　：　犬→おまわりさん
- ❑ 禁忌　　：　スルメ→アタリメ
- ❑ 音の疎通　：　聞く→菊
- ❑ 字謎　　：　酒→サンズイ
- ❑ 符牒　　：　トイレ→突き当たり
- ❑ 類推　　：　包丁→バラス→バラシ

提案する批判検索手法の提案

- 対象企業名を表す隠語を生成
  - → 隠語を検索質問としてWeb検索
- 前述の隠語分類：Web登場以前のもの
  - → Webにおける隠語の造語法を特定
    - 各造語法について隠語生成器を実装
- 実装した隠語生成器
  - → 企業名とその読みを入力
  - → その企業名に対応する隠語を出力

隠語造語法とその生成手法

- 伏せ字：企業名の１文字を〇に置き換える
  例：ソ〇トバンク、SOFTB〇NK
- 英字化：企業名中の１文字をアルファベット１文字に置き換える
  例：Sフトバンク、ソフトバNク
  手法：企業名中の1文字をローマ字の先頭1文字に置き換える
- 入力誤り：半角/全角モードを反転して企業名を入力する
  例：sofutobanku、そftばんk
  手法：企業名が日本語表記→読みをローマ字に変換
  企業名が英語表記→文字列をひらがなに変換
- 字種の変換：文字列の一部をカタカナやひらがなに置き換える
  例：ソフトバンく、そフトバんく
  手法：企業名の読みを分割し、片方をひらがな、片方をカタカナに変換

隠語造語法とその生成手法(2)

- 表記の類似：企業名の一部を見た目が似た文字に置き換える
  例：ンフトバソク、SOFTB@NK
- 変換誤り：意図的に漢字変換を誤る
  例：祖父と万苦、ソフトバン苦
  手法：企業名の読みをブロックに切り分けて漢字変換辞書を参照し、変換可能な組み合わせを出力
- 意味の類似：企業名の一部または全部を類義語に置き換える
  例：ソフト銀行、やわらか土手
  手法：企業名の読みをブロックに切り分け、Cycloneを用いて類義語を選出し、置き換え可能な組み合わせを出力
- 発音の類似：企業名の一部から発音が似た別の語句を連想し置き換える
  例：損フトバンク、孫フトバンク
  手法：DPマッチングを用いて企業名と読みが類似する文字列を抽出

批判検索

- 生成した隠語を検索質問としてWeb検索
- 検索にはYahoo!を用いている
- 検索されたページ集合から隠語のないページを削除

評価実験

- 隠語を含むページを検索し、批判が書かれているか判定
- 実験データ
  → 「ソフトバンク」、「不二家」、「アマゾン」
- 実験手法
  1. 企業名から隠語を生成
     → 「ソフトバンク」:57件、「不二家」:45件、
     「アマゾン」:49件

  2. 生成した隠語1件につき最大20件まで「本文中に隠語を含むページ」を収集し、重複を削除
     → 「ソフトバンク」:522件、「不二家」:498件、
     「アマゾン」:474件

実験結果

- 各企業について隠語、非隠語で検索したページの精度を比較
- 隠語を用いたほうが高精度で批判を検索可能

	ソフトバンク	不二家	アマゾン	全体
隠語	12.3%	6.0%	7.4%	8.6%
	(64/522)	(32/530)	(35/474)	(131/1526)
非隠語	3.1%	6.0%	0.7%	3.3%
	(15/488)	(27/449)	(3/428)	(45/1365)

実験結果(2)

- 造語法ごとに比較すると・・・
  - → 伏せ字の精度が高い
  - →「伏せ字」は批判を検索する精度が高い

造語法	批判文書数	検索文書件数	精度
伏せ字	65	395	16.5%
変換誤り	34	435	7.8%
英字化	11	149	7.4%
字種	11	168	6.5%
入力誤り	2	36	5.6%
表記の類似	8	219	3.7%
全体	131	1526	8.6%

誤り分析

- 検索したページのうち批判でなかったものを分析
  → 検索誤りの原因について分析
- 原因は生成した隠語が「隠語を意図していない別の言葉」と偶然一致したから
  例：ハンドルネーム、実在する別の企業名
  →別の実体を指す文字列を特定、削除できれば批判検索の精度があがる！
- 隠語での検索：感情的な批判がヒット
  非隠語での検索：理性的、感情的両方の批判がヒット
  → 理性的な批判は隠語では説得力がダウン

隠語生成の改善

- Web検索エンジンに入力された検索質問を調査型と誘導型に分類する手法を用いる
- 調査型：Web上の情報を広く調査するための検索質問
  誘導型：ある事項に関する代表的なページを検索するための検索質問
- 「別の実体を指す文字列」→対象となる企業以外の事項
  → 誘導型に分類されやすい
  → 生成された隠語から誘導型に分類されたものを削除

↓

批判検索の精度を向上させることに成功！

おわりに

- Webで用いられる隠語の造語法を特定し、一部の造語法について隠語生成器を実装
- 自動生成した隠語を用いてWeb検索することにより、批判文書を効率的に検索

- 今後
  - 未実装の造語法を用いて批判検索を行う
  - 誘導目的、ハンドルネーム、中国語等を含むページを区別して検索精度を向上させる