# 評判情報の検索における隠語的造語法の応用

木村 友秋　藤井 敦
筑波大学大学院図書館情報メディア研究科

## 1 はじめに

Web 上の評判情報は，企業にとっては，自社の商品やサービスを改善するための参考情報として，消費者にとっては，商品やサービスの良し悪しを判断するための参考情報として重要である．そこで，評判情報の検索に関する研究が行われている．

評判情報の検索は，レビュー記事や掲示板など評判が高密度で書かれたテキストを対象とした手法 [3] と一般的な情報も対象とした手法 [4] に分けることができる．Web から網羅的に評判情報を検索するには，後者のように評判が書かれていることが保障されない様々なテキストからも評判を検索する手法が必要である．後者の先行研究には，水口ら [4] の eHyouban がある．水口らは，blog 記事から「対象物，属性，評価表現」の 3 組みに基づいて評判情報を抽出する．

しかし，評価表現が明記されていない評判もある．例えば,「近所の電気屋はいまだに定価で商品を売っている」という文は,「近所の電気屋は高い」ことを意味する評判であるにも拘わらず,「高い」という評価表現はない．そこで，評価表現のない評判情報を検索する手がかりとして,「隠語」に着目した．隠語とは，本来の意味があえて隠された言葉である．

Web 上の評判では，評価対象の企業名を隠語で表すことがある．企業名が隠語で表されている場合，同時に企業への批判が書かれることがある．例えば,「うちの家族は，どうもソフ○バンクに嫌悪感を抱いています」というように,「ソフ○バンク」という隠語を用いて「ソフトバンク」を批判する．隠語を用いて批判する理由には，婉曲やことば遊びなど複数の理由が考えられる．

以上から，本研究では評判情報のうち批判的な情報に焦点をあて，隠語を用いて批判が書かれたページを検索する手法を提案する．本研究は，評判検索における新しいモデルを提案する点で意義がある．また，提案手法を既存の手法と組み合わせることで，より多くの評判を検索することができる．

## 2 隠語の分類

前田 [5] は，隠語を以下に示す 12 種類に分類した．括弧内は，通常の言葉とその隠語である．

**音節省略** 元の語を一部省略する（警察 ⇒ サツ）

**音節転換** 音節を倒置する（声 ⇒ エコ）

**形状類似** 形が似た事物に置き換える（針 ⇒ 松葉）

**色彩類似** 色が似た事物に置き換える（カラス ⇒ 墨）

**連想** 連想される事物に置き換える（はさみ ⇒ カニ）

**動作** 動作の擬音語に置き換える（入浴 ⇒ ザブン）

**比喩** 元の語を別の事物に例える（犬 ⇒ おまわりさん）

**禁忌** 縁起の悪い言葉を置き換える（スルメ ⇒ アタリメ：スルメはお金をスルから縁起が悪い）

**音の疎通** 同音の言葉に置き換える（聞く ⇒ 菊）

**字謎** 漢字のなぞかけをする（酒 ⇒ サンズイ）

**符牒** 客に意味がわからないように言葉を置き換える（トイレ ⇒ 突き当たり）

**類推** 既存の隠語から新しい隠語を類推する（包丁 ⇒ バラス ⇒ バラシ：人を殺すという意味のバラスから類推）

野村ら [6] は，以下に示すような分類も提案している．結果類似（盗む ⇒ 買う），近接（煙草 ⇒ モク），婉曲（質屋 ⇒ 一六銀行），特殊化（犯人 ⇒ ホシ），謎（無賃乗車 ⇒ サツマノカミ：薩摩守忠度はサツマノカミ**タダノリ**だから）．

## 3 提案する批判検索手法

### 3.1 概要

1 章で説明したように，ある企業名を隠語で書いたページには，その企業に対する批判が書かれやすい．そこで，対象の企業名を表す隠語を自動的に生成し，その隠語を検索質問として Web を検索する．

### 3.2 隠語生成

2 章で説明した隠語の分類は，Web が登場する以前の研究である．そこで，Web における隠語の造語法を特定するために,「ソフトバンク」に関する種々の隠語を人手で分析し，表 1 に示す 13 種類の造語法を特定した．2 章の造語法では，表 1 の上から 5 種類は分類されていない.「意味の類似」,「発音の類似」,「反意」は，既存の造語法が言葉全体を別の言葉に置き換えるのに対して，言葉の一部分を別の言葉に置き換える点が異なる．

表 1 の上から 8 種類の造語法について，隠語生成器を実装した．残り 4 種類の造語法は，以下の理由から対象外とした.「転置」と「イニシャル化」の隠語は，企業と関係ないページにヒットする.「省略」は，単なる省略語

表 1: Web ページの分析によって特定した隠語的造語法

| 造語法名 | 説明 | ソフトバンクを隠語化した例 |
|---|---|---|
| 伏せ字 | 企業名中の 1 文字を○に置き換える | ソ○トバンク　SOFTB ○ 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 | Ｓフトバンク　ソフトバＮク |
| 入力誤り | 半角/全角モードを反転して企業名を入力する | sofutobanku　そｆｔばんｋ |
| 字種の変換 | ひらがなの一部をカタカナに，カタカナの一部をひらがなに置き換える | ソフトバんく　そフトバんく |
| 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 | ンフトバソク　SOFTB@NK |
| 変換誤り | 意図的に漢字変換を誤る | 祖父と万苦　ソフトバン苦 |
| 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 | ソフト銀行　やわらか土手 |
| 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 | 損フトバンク　孫フトバンク |
| 転置 | 企業名を 2 つに区切り，前後を入れ換える | バンクソフト |
| 逆さ読み | 企業名を逆から読む | クンバトフソ　KNABTFOS |
| 省略 | 企業名の一部を省略する | ソフバン　ソフトバ |
| イニシャル化 | 企業名をイニシャルにする | SB　S |
| 反意 | 企業名の一部を意味が反対の言葉に置き換える | ハードバンク |

として用いられることが多い.「反意」は，企業名に反意語のある語句が含まれている必要があり，限られた企業名にしか適用することができない.

実装した隠語生成器は，企業名とその読みを入力すると，その企業名に対応する隠語を出力する．以下，造語法ごとに隠語生成の手法について説明する.

**伏せ字** 企業名中の 1 文字を○に置き換える．N 文字の企業名からは N 通りの隠語が生成される.

**英字化** 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．例えば,「不二家」の「不」をローマ字の「Fu」にして先頭 1 文字だけを残し,「Ｆ二家」を生成する.

**入力誤り** 企業名が日本語表記の場合は，企業名の読みをローマ字に変換する．企業名が英語表記の場合は，ローマ字読みできる文字列をひらがなに変換する．ただし，ローマ字読みできない部分はそのまま残す.

**字種の変換** 企業名の読みを 2 分割し，前のブロックをひらがなに，後ろのブロックをカタカナに変換する．または，前のブロックをカタカナに，後ろのブロックをひらがなに変換する．読みが N 文字の企業名からは，最大 2N 通りの隠語が生成される.

**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．置き換えは，次に示す４パターンを用意した.「ア ⇔ ァ」のように大文字と小文字を相互に入れ換える.「ソ ⇔ ン」のように見た目が似ている文字を相互に入れ換える.「バ ⇔ パ」のように濁音と半濁音を相互に入れ換える.「バ/パ ⇒ ハ」のように濁音か半濁音を清音に置き換える.

**変換誤り** 企業名の読みを N 個のブロックに切り分けて，漢字変換辞書[1]を参照して各ブロックの変換候補を最大 M 件挙げ，変換可能な組み合わせを全て出力する．本研究では，N=2，M=3 とした.

**意味の類似** 企業名の読みを N 個のブロックに切り分けて，Cyclone[2]を用いて各ブロックの類義語を最大M 件挙げ，置き換え可能な組み合わせを全て出力する．本研究では，N=2，M=3 とした.

**発音の類似** ある隠語で検索した Web ページには，同じ対象に対する別の隠語も書かれている可能性がある．この傾向は，掲示板のように複数の人間が同じ企業に関して投稿する場合によく現れる．そこで，上記 7 種類の造語法で生成した隠語で検索されたページの中から，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する．上記 7 種類の造語法は，企業名とその読みから隠語を「生成」するのに対して,「発音の類似」は Web ページから隠語の候補を「抽出」する．これは,「発音の類似」を生成することは人手でも難しく，生成するよりも Web から抽出する方が正確性が高いためである.

なお，各造語法から生成される隠語の数を制限する.「伏せ字」は，2 文字以上を○に置き換えると，企業と関係ないページにヒットしやすくなるので 2 文字以上の置き換えはしない.「英字化」も同様の理由で 2 文字以上の置き換えはしない.「字種の変換」では原理的には N 文字の企業名から最大 $N^2$ 種類の隠語が生成される．しかし,「ソふトばンく」のような複雑な変換パターンの隠語は入力の手間がかかるので，Web でほとんど用いられない．そこで，字種の境界を 1 箇所に限定する.

## 3.3 批判検索

3.2 節の手法を用いて生成した隠語を検索質問として Web を検索する．現在，検索には Yahoo![3]を用いている．さらに，検索されたページの集合から，実際には隠語が書かれていないページを削除する．Yahoo!は，フレーズ

[1] http://openlab.jp/skk/wiki/wiki.cgi

[2] http://cyclone.slis.tsukuba.ac.jp/

[3] http://www.yahoo.co.jp/

検索を用いた場合でも，検索質問が含まれていないページを検索することがある．例えば,「不二家」の隠語である「〇二家」を検索質問として用いると,「〇二家」を含まずに「環二家」を含むページが検索される．

# 4 評価実験

## 4.1 実験手法

提案手法を用いて隠語を含むページを検索し，各ページについて批判かどうか判定した．判定では，本文に書き手の批判的な態度が含まれているページを批判と判定した．企業の失敗や過失などの事実を記述しながらも，その事実に対して批判的な態度が示されていないページは批判と判定しなかった．根拠のない誹謗中傷は，書き手の批判的な態度が示されていることから批判と判定した．

## 4.2 実験データ

「ソフトバンク」,「不二家」,「アマゾン」という 3 つの企業名を対象に，提案手法を用いて批判検索を行った．

まず，各企業名から「伏せ字」,「英字化」,「入力誤り」,「字種の変換」,「表記の類似」,「変換誤り」の造語法を用いて隠語を生成した．予備実験の結果,「意味の類似」と「発音の類似」は隠語生成の精度が低かったため用いなかった．各企業の英語表記である「SOFTBANK」,「FUJIYA」,「Amazon」も正式名称なので，日本語表記と英語表記の両方に対して隠語を生成した．その結果,「ソフトバンク」からは 57 件,「不二家」からは 45 件,「アマゾン」からは 49 件の隠語が生成された．

次に，各企業名について，生成した隠語 1 件につき上位から最大 20 件まで「本文中に隠語を含むページ」を収集し，さらに重複を削除した．最終的に,「ソフトバンク」,「不二家」,「アマゾン」に対して，それぞれ 522 件，498 件，474 件のページが検索された．

比較対象として，本文中に元の企業名（非隠語）を含むページを収集し，重複を削除した．検索には日本語表記の非隠語のみ用いた．英語の表記を用いても企業の公式ページのような批判以外の情報が増えるだけであった．「ソフトバンク」,「不二家」,「アマゾン」に対して，それぞれ 488 件，449 件，428 件のページが検索された．

## 4.3 実験結果

各企業について，隠語で検索したページと，非隠語で検索したページの精度を比較した結果を表 2 に示す．括弧内の数字は，批判文書数と検索文書数である．精度は，検索文書数に対する批判文書数の割合である．

「ソフトバンク」では，隠語での精度 12.3%に対して，非隠語での精度は 3.1%,「アマゾン」では，隠語での精度 7.4%に対して，非隠語での精度は 0.7%であった.「ソフトバンク」と「アマゾン」は，隠語の方が批判検索の精度が高かった．しかし,「不二家」では隠語も非隠語も精度は 6.0%で同等だった．全体では，隠語での精度 8.6%に対して非隠語での精度は 3.3%と，隠語の方が精度が高かった．

表 2: 企業ごとの批判検索精度

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

以上より，提案手法は，隠語を用いない検索と比べて，同精度または高精度で批判を検索することができた．

造語法ごとの批判検索精度を表 3 に示す．全体における精度である 8.6%を上回ったのは,「伏せ字」の 16.5%のみであった.「伏せ字」の次に精度が高い「変換誤り」の 7.8%は,「伏せ字」の半分程度であった．すなわち,「伏せ字」は造語法の中でも特に批判を検索する精度が高かった．

表 3: 造語法ごとの批判検索精度

| 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% |

## 4.4 誤り分析

提案手法で検索したページのうち，批判でなかったページを分析し，検索誤りの原因について分析した．

批判でないページが検索された原因は，生成した隠語が「隠語を意図していない別の言葉」と偶然一致したことによる．生成した隠語で一致した文字列がどのような言葉であったのかを表 4 に示す．

表 4: 生成した隠語に一致した文字列の種類

| 企業 | 隠語 | 誤字 | 誘導目的 | HN | 中国語 | その他 |
|---|---|---|---|---|---|---|
| ソフトバンク | 453 | 35 | 19 | 6 | 0 | 9 |
| 不二家 | 138 | 1 | 0 | 0 | 5 | 386 |
| アマゾン | 381 | 10 | 37 | 17 | 1 | 28 |
| 全体 | 972 | 46 | 56 | 23 | 6 | 423 |
| 批判 | 130 | 1 | 0 | 0 | 0 | 0 |

表 4 の「隠語」は，一致した文字列が隠語であった場合であり，提案手法が適切に機能した場合である.「隠語」に一致したページ 972 件のうち，最下行の「批判」に示したように，批判は 130 件であった．それに対して,「隠語」以外に一致したページ 554 件のうち批判は 1 件しかなかった．すなわち,「隠語」は他の種類と比べて批判の目的で用いられやすい．

「誤字」は，一致した文字列が企業名の入力誤りであった場合である．文字列が隠語か入力誤りか判断が難しい場合は全て「隠語」に分類し，文脈から明らかに入力誤りとわかる文字列だけを「誤字」に分類した．

「誘導目的」は，一致した文字列が意図的にタイプミスされた場合であり，主に広告収入目的であった．例えば，ページ内に「ソフトバンク」のタイプミスである「そｆｔばんｋ」と「ソフトバンク」の広告を埋め込むことにより，タイプミスをしたユーザを自分のページに誘導し，広告をクリックさせて広告収入を得る狙いである．そこで,「誘導目的」は批判の目的では用いられない．

「HN」は，一致した文字列がハンドルネーム（Web上のニックネーム）として使用されていた場合であり，批判の目的では用いられない．

「中国語」は，一致した文字列が中国語であった場合である．漢字で構成された隠語は，中国語のページにヒットすることがある．本研究では，中国語で書かれたページは検索の対象外である．

「その他」は，一致した文字列が企業と関係ない文の一部であった場合である.「不二家」の隠語で収集したページは「その他」に分類されることが多かった．そこで,「その他」に分類されたページを調べたところ,「変換誤り」の造語法で生成された隠語の多くが別の実体を指す文字列であった．例えば,「不二家」から生成された隠語の「フジ矢」,「藤家」,「富士家」は，それぞれ実在する店舗等の名称である．生成した隠語のうち，別の実体を指す文字列を特定し，削除することができれば，批判検索の精度を向上させることができる．この点について，4.5 節で検討する．

最後に，隠語では検索されにくい批判について考察する．隠語を用いた検索では，感情的な批判だけが検索されたのに対し，非隠語を用いた検索では，理性的な批判と感情的な批判が検索された．隠語で理性的な批判を検索することができなかった理由は，理性的な批判に隠語を用いると批判の説得力が弱くなるからである．

### 4.5 隠語生成の改善

藤井 [1] は，Web 検索エンジンに入力された検索質問を調査型（informational）か誘導型（navigational）に分類する手法を提案した．調査型とは，ある話題について Web 上の情報を広く調査するために使用される検索質問である．誘導型とは，ある事項（人物，組織，商品，イベントなど）に関するトップページや代表的なページを検索するために使用される検索質問である．

「別の実体を指す文字列」は，対象となる企業以外の事項を指す語句である．すなわち誘導型に分類されやすい．そこで，藤井の手法を用いて，生成した隠語のうち，別の実体を指す文字列を誘導型に分類することができるかどうか実験した．分類対象は,「変換誤り」で生成された「不二家」の隠語 20 件である．分類のためのコーパスは，NTCIR-5 Web タスクの 1T バイトデータ [2] を使った．分類結果を表 5 に示す．実際に Web を検索し，別の実体を指すと判明した隠語の候補には下線を引いた．

表 5: 「不二家」に対して生成した隠語の分類結果

| 調査型 | 誘導型 |
|---|---|
| ふじ矢　不治屋　フジ屋　藤屋<br>藤や　ふじ家　藤矢　富士矢<br>負じや　藤ヤ　不治や　不治矢<br>不じや　不治家　不じや | ふじ屋　フジ矢<br>富士や　富士家<br>藤家　フジ家 |

誘導型として分類された 6 件の隠語候補は，全て別の実体を指す文字列すなわち「ノイズ」であった．そこで，これら 6 件を削除して「不二家」について批判検索を再度行った．その結果，435 ページが検索され，うち 32 件が批判であり，検索精度は 7.4%であった．ノイズを削除する前は，530 件中 32 件が批判で検索精度は 6.0%であった．すなわち，ノイズを削除することによって，批判検索の精度を向上させることができた．

## 5 おわりに

本研究は，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した．さらに，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した．今後は，未実装の造語法を用いて批判検索を行うことが課題である．また，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，批判検索の検索精度を向上させることが課題である．

## 謝辞

本研究の一部は，文部科学省科研費特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（課題番号: 19024007）によって実施された．

## 参考文献

[1] Atsushi Fujii. Modeling anchor text and classifying queries to enhance web document retrieval. *Proceedings of the 17th International World Wide Web Conference*, pp. 337–346, 2008.

[2] K. Oyama, M. Takaku, H. Ishikawa, A. Aizawa, and H. Yamana. Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2). *In Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp. 423–442, 2005.

[3] Peter. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417 – 424, 2002.

[4] 水口弘紀, 土田正明, 久寿居大. Weblog を対象にしたリアルタイム評判情報分析システム eHyouban. 電子情報通信学会 第 19 回データ工学ワークショップ論文集, 2008.

[5] 前田太郎. 外来語の研究. 岩波書店, 1922.

[6] 野村雅昭, 小池清治. 日本語事典. 東京堂, 1992.