評判情報の検索における隠語的造語法の応用*

木村友秋（学籍番号　200721528）
研究指導教員：藤井敦

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，評判情報を効率良く検索する手法について研究されている．

水口ら[1]は，「対象物，属性，評価表現」の 3 つ組に基づいて評判情報を抽出した．対象物とは，評価対象の名称，属性とは対象の構成要素，評価表現とは属性に対する書き手の考えや思いである．例えば，「近所の電気屋は商品が高い」という文では，「近所の電気屋」が対象物，「商品」が属性，「高い」が評価表現である．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，「商品が高い」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．

本研究は，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． Web では，評価対象の企業名を隠語で表記することがあり，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，どうもソフ〇バンクに嫌悪感を抱いています」というように，「ソフ〇バンク」という隠語を用いて「ソフトバンク」を批判する．以上から本研究は，隠語を用いて批判が書かかれたページを検索する手法を提案し，システムを実装した．

２．提案する批判検索手法

本研究の手法は，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 を検索することで，当該企業の批判である可能性が高いページを収集する．

2.1 隠語生成

国語学における隠語の研究では，使用者，使用目的，造語法などに基づいて隠語を分類している．このうち，本研究の隠語生成と関連するのは造語法に基づく分類である．しかし，既存の造語法では Web で使われる隠語すべてを分類することができない．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，批判の検索に有効な造語法８種類を特定した．さらに各造語法について，企業名とその読みを入力として，対象の企業を表す隠語を出力する隠語生成器を実装した．以下，造語法ごとに隠語生成の手法について説明する．また，「ソフトバンク」から生成される隠語の例を示す．

伏せ字 企業名中の1文字を〇に置き換える．「ソ〇トバンク」，「ソフトバ〇ク」

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．「Sフトバンク」，「ソフTバンク」

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．「sofutobannnku」，「そftばんk」

字種の変換 企業名の読みを平仮名で表記し，さらに読みの一部をカタカナに変換する．「ソフトばんく」，「そフトバンク」

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」，「ソフトパンク」

変換誤り 企業名を任意の数に分割し，各ブロックを読みが同じ漢字や仮名に置き換える．「祖父と万苦」，「祖ふとばんく」

意味の類似 企業名を任意の数に分割し，一部のブロックを意味が類似する別の言葉に置き換える．意味が類似する言葉はCyclone [2]を用いて取得する．「やわらか銀行」，「ソフト金庫」

発音の類似 ある隠語で検索したページには，同じ対象に対する別の隠語も存在する可能性がある．上記 7 種類の造語法で生成された隠語

* “An Application of Jargon-style Word Formation to Retrieving Reputations” by Tomoaki KIMURA

で検索したページから，元の企業名と発音が似た文字列をDPマッチングで特定し，隠語として抽出する．「損フトバンク」，「ソフトバン糞」

2.2 批判検索

2.1 節の手法で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． Yahoo!は，フレーズ検索を用いた場合でも，検索質問が含まれないページを検索することがある．そこで，検索されたページのうち，隠語を含まないページを削除する．

３．評価実験

「ソフトバンク」，「アマゾン」，「不二家」という企業名から生成された隠語を用いて Web を検索し，検索されたページが批判かどうかを人手で判定した．ただし，実装した 8 種類の造語法のうち「意味の類似」と「発音の類似」は予備実験の結果，隠語生成の精度が低かったため，実験には用いなかった．また，各企業の英語表記である「SOFTBANK」，「Amazon」，「FUJIYA」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．比較対象として，本文に元の企業名(非隠語)を含むページを検索し，批判かどうか人手で判定した．

3.1 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 1 に示す．括弧内は，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」と「アマゾン」では，隠語を用いた方が批判検索の精度が高かった．「不二家」では，隠語と非隠語で批判検索の精度に差が無かった．

表 1: 企業ごとの批判検索精度

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

3.2 誤り分析

提案手法で検索されたページのうち，批判ではなかったページを分析した．その結果，生成した隠語が「隠語を意図していない別の言葉」と偶然一致することが主な原因だった．例えば，「不二家」から生成された隠語である「フジ矢」で Web を検索すると，工具製造業者である「フジ矢」について書かれたページが検索された．すなわち，この例では隠語が元の企業名とは別の実体に偶然一致した．

「隠語を意図していない別の言葉」には，「別の実体」，「企業名の誤字」，「ハンドルネーム(Web 上のニックネーム)」，「中国語」，「語句の一部」，「誘導目的」があった．「語句の一部」とは，隠語がある文に書かれた語句と偶然一致する場合で，例えば「不二家」の隠語である「藤や」が，「この公園では藤や芭蕉がきれい」という文に一致する場合である．「誘導目的」とは，企業名を入力誤りした Web ユーザを自分のページに誘導する目的で，意図的に企業名の入力誤りが書かれる場合である．

「不二家」の隠語が非隠語と比べて精度に差が無かった原因は，「不二家」から生成される隠語の多くが「別の実体」に一致してしまい，批判が検索されなかった為である．この問題を解決するため，検索質問分類手法[4]を応用し，「別の実体」に一致する隠語の一部を分類することに成功した．別の実体に分類された隠語を消去して「不二家」で再検索を行ったところ，隠語での精度は 7.4%に向上し，非隠語の精度 6.0%と比べて高精度で批判を検索することができた．

４．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 を検索することにより，企業に対する批判を効率良く検索した．残された課題は，商品名などの固有名詞にも提案手法を応用し，評価することである．

文献

[1]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．DEWS2008 論文集，2008.

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

[3]Yahoo!，http://www.yahoo.co.jp/

[4] 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.