評判情報の検索における隠語の生成と順位付け

太田 裕貴[1]　　藤井 敦[2]

[1] 東京工業大学 工学部 情報工学科
[2] 東京工業大学 大学院情報理工学研究科 計算工学専攻

1　はじめに

Web上にある評判情報は，企業においては自社に対するイメージや商品に対する感想を知る際の参考情報として，消費者においては商品の購入を検討する際の参考情報として重要である．そのため評判情報の検索に関する研究が行われている．

評判情報を検索するためには，レビュー記事や掲示板などの評判が高密度で書かれたページだけでなく，評判が書かれていることが保障されない様々なページからも評判を検索する必要がある．水口ら[1]は，blog記事から「対象物，属性，評判」の組を評価表現として利用することで，評判情報を検索している．

木村ら[2]は，評価表現が明記されていない評判情報を検索するために「隠語」に着目した．隠語とは,「ソフ○バンク」のように正式名称である「ソフトバンク」があえて隠された言葉である．木村らは，正式名称が隠語で書かれたページには，その対象の評判や批判（悪い評判）が書かれやすいという性質に着目し，Web上における隠語の造語法として以下の13種類を特定した．

> 伏字，英字化，入力誤り，字種の変換，表記の類似，変換誤り，転置，逆さ読み，省略，イニシャル化，反意

木村らは，上記の造語法を用いて正式名称から隠語候補を生成し，生成した隠語候補を検索質問として検索することにより，評判文書や批判文書の検索精度を向上させた．

しかし，木村らの手法で生成された隠語候補には，隠語とそうではない文字列が混在しているため，本来の対象とは無関係な文書が検索されてしまう場合がある．

そこで本研究は，木村らの手法で生成された隠語候補に順位付けを行う手法を提案する．隠語候補に順位を付けることにより，本来の対象とは関連しにくい不適切な隠語候補を除くことができ，評判情報の検索精度を向上させることが可能になる．

2　隠語の性質

2.1　概要

本研究は，Web上の隠語が持つ2つの特徴に着目する．まず，正式名称とその隠語が同じ文書に出現しやすいという性質（共起性）である．もう1つは，正式名称とその隠語は似たような文脈に出現しやすいという性質（文脈類似性）である．

2.2　共起性

QAサイトなど複数のユーザが投稿する場合，全員が同一の隠語だけを用いて文書を書くことは少なく，正式名称や他の隠語も一緒に出現しやすい．以下に例文をあげる．

> **投稿者1**：ぼくの携帯はソフ○バンクです
> **投稿者2**：ソフトバンクなので送れますね

なお，以下の例文に示すように，この傾向は投稿者が1人の場合でも見られる．

> ソＦトバンクに物色に行ったものの …
> Ｓフトバンクのカタログを見ながら …

2.3　文脈類似性

正式名称と隠語の文脈が類似している例文を以下にあげる．

> **正式名称で書かれた文書**:
> 携帯ソフトバンク使ってますが，メール受信に料金はかかりますか？
> **その隠語で書かれた文書**:
> Ｓフトバンク携帯を使っています．受信メールを全部消してしまいました

上記の例文では，両方の文書に共通して「携帯，使う，ます，メール，受信」という語が含まれている．

3　提案する順位付け手法

3.1　用語の定義

提案手法中に現れる語を以下のように定義する．

「正式名称」とは，企業等の対象が持つ正確な名前である.「隠語」とは,「正式名称」をあえて隠した言葉である.「隠語候補」とは，木村ら[2]の造語法により自動生成された文字列であり,「隠語」とは限らない．また，正式名称を x とした時に，その名称に対する隠語候補の1つを x' と表記する．

3.2 共起性に基づく手法

正式名称 *x* とその隠語候補 *x'* の自己相互情報量を式 (1) に基づいて計算する [3]．分子と分母の文書数には，Yahoo![1]で検索した際に表示される検索文書件数を使用する．

$$PMI(x, x') = \log \frac{x \text{ と } x' \text{の両方を含む文書数}}{(x \text{ を含む文書数}) \quad (x' \text{を含む文書数})} \quad (1)$$

式 (1) の値に基づいて複数の隠語候補に順位を付ける．さらに，上位K件の隠語候補のみを検索質問として用いる．

3.3 文脈類似性に基づく手法

文脈類似度に基づく手法では，文書に出現する語を用いて対象語をベクトルで表現し，2 つのベクトルを内積等によって比較する．しかし，正式名称は企業のホームページやニュースで用いられる文語表現であり，一方隠語は blog 記事や QA サイトで用いられる口語表現で書かれることが多い．そのため，正式名称の周辺語と隠語候補の周辺語が異なり，類似度の計算が上手く行われない傾向にある．具体例を表 1 にあげる．「ソフトバンク」と隠語候補である「Sフトバンク」の周辺語のうち，出現頻度上位 20 語を載せる．20 語のうち共通しているのは 8 語だけである．

表 1:「ソフトバンク」と「Sフトバンク」の周辺語上位 20 件

ソフトバンク	する，日，月，年，モバイル， 電話，いる，情報，サイト，こと， 株式会社，発表，iphone，の，softbank， れる，サービス，福岡，円，ショップ
隠語候補 S フトバンク	する，日，月，年，の， いる，ん，ある，なる，こと， 携帯，middot，てる，cm，よう 会社，れる，人，くる，電話

本手法は次の手順で行う．最初に，隠語候補の周辺語を獲得するため，生成した各隠語候補を検索質問として，Yahoo!を用いて Web 検索する．各隠語候補について検索されたページのうち上位 N 件を用いる．これらのページからストップワードを除去し，MeCab[2]を用いて形態素解析を行う．その後，名詞，動詞，形容詞を周辺語として抽出する．ただし，数字は正式名称との関連性が低いことが多いため抽出しない．

次に，式 (2) を用いて，隠語候補 *x'* の周辺語 *t'* ごとに，周辺語 *t'* が正式名称 *x* と同じ文書に出現する確率を計算する．分子と分母の文書数には，Yahoo!で検索した際に表示される検索文書件数を使用する．

$$P(x|t') = \frac{x \text{ と } t' \text{の両方を含む文書数}}{t' \text{を含む文書数}} \quad (2)$$

重複した分だけ周辺語 *t'* に重みを付けるため，周辺語 *t'* の重み *W*(*t'*) を式 (3) を用いて以下の式で定義する．log を使うことで，出現頻度の影響を弱くする．

$$W(t') = 1 + \log(t' \text{の出現頻度}) \quad (3)$$

式 (2) と式 (3) の 2 つの数値を用いて，各隠語候補 *x'* と正式名称 *x* との類似度を式 (4) で計算し，各隠語候補のスコアとする．

$$Sim(x, x') = \frac{1}{M} \sum_{i=1}^{M} P(x|t'_i) \quad W(t'_i) \quad (4)$$

式 (4) の値に基づいて複数の隠語候補に順位を付ける．さらに，上位K件の隠語候補のみを検索質問として用いる．

表 2:「アマゾン」の隠語候補に対する順位付けの実行例

順位	共起性		文脈類似性	
	隠語候補	隠語かどうか	隠語候補	隠語かどうか
1	海士損	×	亜マゾン	○
2	尼ぞん	○	海士損	×
3	あマゾン	○	アマ存	○
4	アマゾソ	△	海人損	○
5	ア M ゾン	○	あま損	○
6	あま存	○	尼ぞん	○
7	アマ損	○	アマぞん	○
8	アマソン	×	ア M ゾン	○
9	アマゾ N	×	あまゾン	△
10	尼損	○	アマゾ N	×

3.4 順位付けの実行例

正式名称「アマゾン」から隠語候補を生成し，「共起性に基づく手法」と「文脈類似性に基づく手法」の 2 つ手法を用いて順位を付けた結果のうち，上位 10 件を表 2 に示す．2 つの手法で隠語候補への順位の付け方が異なっている．「隠語かどうか」は，隠語候補が隠語として，適切ならば○，不適切ならば×，適切でも不適切でもないならば△で表す．

4 評価実験

4.1 概要

「ソフトバンク」，「アマゾン」，「不二家」という 3 つの正式名称を対象として，木村らの手法を用いて生

[1] http://www.yahoo.co.jp/
[2] http://mecab.sourceforge.net/

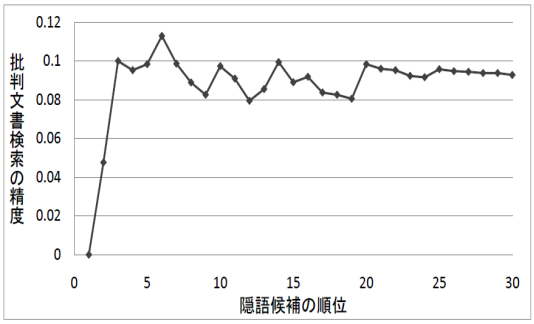

図 1: 「共起性に基づく手法」の評価（ソフトバンク）

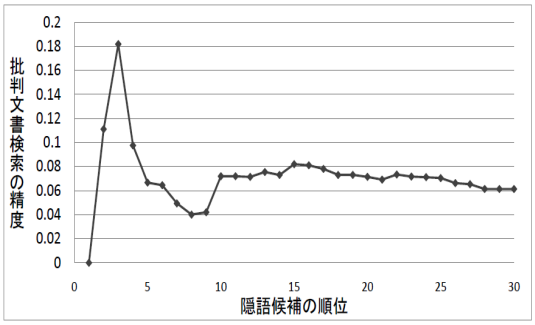

図 2: 「共起性に基づく手法」の評価（アマゾン）

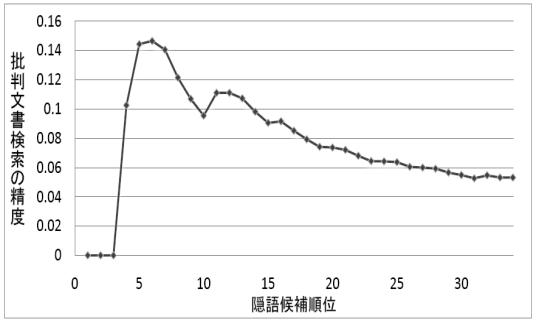

図 3: 「共起性に基づく手法」の評価（不二家）

成された隠語候補に順位を付けた．使用した隠語候補は,「ソフトバンク」が 30 種類,「アマゾン」が 30 種類,「不二家」が 34 種類である．順位付けされた各隠語候補に対して，批判（悪い評判）文書の検索精度を評価する．

4.2 「共起性に基づく手法」の評価

評価の方法を図 1 を使って説明する．隠語候補の上位 1 件を検索質問として用いた時の精度は，図の一番左の点の高さにあたる．上位 2 件までを用いた時の精度は，左から 2 番目の点の高さであり，上位K件までを用いた時の精度は，図の左からK番目の点の高さである．隠語候補全てを使った精度である一番右の点の高さよりも，左からK番目の点の高さが高いとき，上位K件の隠語候補の精度は，隠語候補全てを使った精度より高くなる．

従って，隠語候補の順位付けが成功したかどうかを，左からK番目の点が一番右側の点より高いかどうかで判断する．

表 3: 不二家の隠語候補の周辺語上位 15 件

	不二家と一緒に書かれやすい周辺語
英字	hujiya，fujiya，mognavi，fuziya，fujiya，kddi，jugem，tabearuki，blogram，softbank，jtb，ufj，tsutaya，skywork，smap
動詞	ささえる，支える，没す，かつ，飲む，飲む，貼る，もつ，帰る，いう，合う，もらう，もらえる，とろける，残る
形容詞	甘い，白い，うまい，懐かしい，赤い，若い，濃い，珍しい，厳しい，青い，やさしい，辛い，イイ，ええ，深い

図 1 の「ソフトバンク」では，上位 6 件を用いた時に，隠語候補全てを使った精度を上回っていた．しかし，隠語候補全てを使った精度とほぼ同じ精度であるため，順位付けは成功していない．図 2 の「アマゾン」では，上位 2 件を用いた時に，精度が最大となった．さらに，ほとんどの点において，隠語候補全てを使った精度を上回っていたため，順位付けに成功している．図 3 の「不二家」では，上位 1〜3 位の精度が 0 であった．しかし，上位 4 位以降では隠語候補全てを使った精度を上回っていたため，順位付けに成功している．

4.3 「文脈類似性に基づく手法」の評価

「文脈類似性に基づく手法」により隠語候補の順位付けを行うのに際して，予備調査を 2 つ行った．

名詞以外に英字，動詞，形容詞を周辺語に使用することが，隠語候補の順位付けに適当であるかを調査した．

正式名称「不二家」を例に，生成された隠語候補の周辺語において，式 (2) における正式名称「不二家」と同じ文書に出現する確率が高い周辺語上位 15 件を表 3 に示す．

動詞と形容詞では正式名称と関連性が高い語が多く見られた．そのため，周辺語に動詞と形容詞を用いるのは妥当である．

英字では正式名称の英語による異表記が見られた．しかし，それ以外の周辺語は正式名称「不二家」と関連性の低い語が多くあった．さらに，アルファベットを羅列しただけの本来は存在しない語を含むため，周辺語から英字を除外する．

隠語候補の周辺語を獲得する際に，各隠語候補を検索質問として検索し，獲得したページのスニペットを使用した．スニペットの上位何件までを使用するかを 50 件ずつ変化させることにより，隠語候補の順位を付けるのに適当なスニペットの数を調査した．その結果，使用するスニペットの数が 50 件の時に，精度が高くなることが分かった．

以上 2 つの調査により，周辺語に名詞, 動詞, 形容詞を使用し，周辺語を獲得するために使用するスニペッ

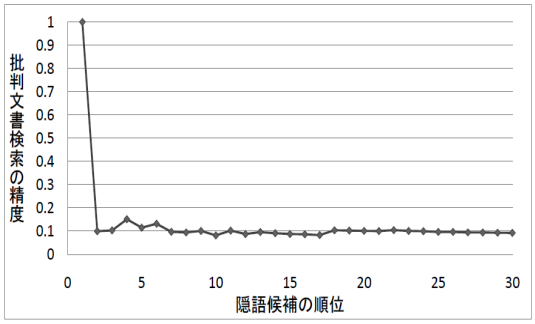

図 4: 「文脈類似性に基づく手法」の評価（ソフトバンク）

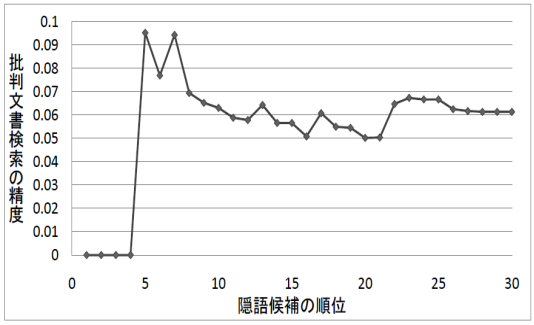

図 5: 「文脈類似性に基づく手法」の評価（アマゾン）

トを上位 50 件とした．

図 4 に「ソフトバンク」，図 5 に「アマゾン」，図 6 に「不二家」の結果を示す．

4.3.1 2 手法の比較

「共起性に基づく手法」と「文脈類似性に基づく手法」の結果を比較する．「ソフトバンク」と「アマゾン」では，2 つの手法に差が見られなかった．一方，「不二家」では，「文脈類似性に基づく手法」の精度が「共起性に基づく手法」の精度より高かった．

以上の結果より，「文脈類似性に基づく手法」が「共起性に基づく手法」よりも隠語候補の順位付けに効果的である．

4.4 誤り分析

提案手法で，上位に順位付けされた隠語候補のうち，正しい隠語として書かれていない文書について分析した．

2 つの手法で上位に順位付けされた隠語候補のうち，精度を下げている隠語候補に以下がある．

> sofutobanku，ソフトパンク，ソフトハンク，ソフトバンく，そｆｔばんｋ，アマソン，アマゾN，アマゾん

上記の隠語候補が 2 つの理由で精度を下げている．

1 つは，上記の隠語候補は，正式名称以外に誤字や表記が似ている語をあえて書くことで，アクセス数を増やそうとしている文書で使われることが多くあった．正式名称以外に誤字や表記の似ている語を同じ文書に

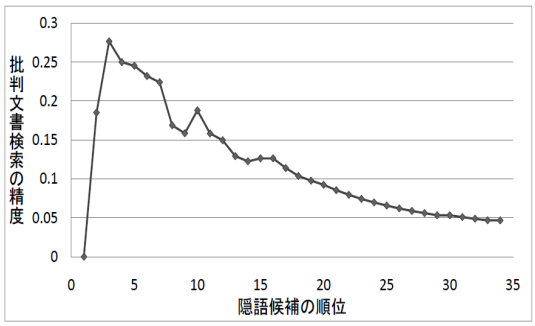

図 6: 「文脈類似性に基づく手法」の評価（不二家）

一緒に書くため，順位付けでは上位になりやすい．しかし，このような文書に出現する隠語候補は隠語ではない．

もう 1 つは，上記の隠語候補は，意図しない誤字として文書で使われることが多くあった．意図しない誤字の場合，文書を書いている人は誤って隠語候補を書いているため，この文書の文脈は正式名称の文書と類似する．しかし，隠語として使われていない．

「不二家」のように，誘導目的や誤字が少なく，別の実態を表す隠語候補が多いときに，「文脈類似性に基づく手法」によって，文脈類似度の低い別の実態を表す隠語候補を下位に順位付けることができる．

5 おわりに

本研究は，正式名称から隠語候補を生成して評判情報検索をする際に，生成した隠語候補に順位付けをすることにより，隠語ではない隠語候補を除き，評判文書を効率的に検索した．今後は，周辺語の獲得をスニペットだけでなく，本文からも獲得することが課題である．また，誘導目的や誤字を含む文書を区別して，評判文書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，文部科学省科学研究費補助金特定領域研究「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 21013003）によって実施された．

参考文献

[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．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

[2] 木村友秋，藤井敦．評判情報の検索における隠語的造語法の応用．言語処理学会第 15 回年次大会論文集，pp. 284–287，2009

[3] Peter. D. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. *In Proceedings of the Twelfth European Conference on Machine Learning*，pp. 491–502，2001