Vol.2013-NL-213 No.7
2013/9/12

Web 掲示板における皮肉の分類および自動検出

磯野 史弥[1,a)]　松吉 俊[2,b)]　福本 文代[2,c)]

概要：本研究では，Web 掲示板に存在する皮肉や誹謗中傷などの不適切な表現を自動的に検出する手法を提案する．我々は，Web 掲示板における皮肉を人手で体系的に分類し，8 つの分類クラス (疑問，推測，諦め，不相応，誇張，驚き，形容，対比) を構築した．それぞれの分類クラスに対して，対象の文とその前後文の評価極性を考慮する構文パターンを設計した．提案する皮肉検出システムは，構文パターンの集合を利用することにより，入力された文が皮肉文であるかどうかを判定する．提案する誹謗中傷検出システムは，Support Vector Machine (SVM) を用いて，入力された文が誹謗中傷文であるかどうかを判定する．ここでは，素性として，独自に構築した辞書に存在する誹謗中傷語の出現頻度と，対象の文とその前後文の評価極性を利用した．評価実験の結果，提案するシステムは，F 値においてベースラインを上回った．

キーワード：自動分類，フィルタリング，皮肉，誹謗中傷，Web 掲示板

Automatic Detection of Sarcasm in BBS Posts Based on Sarcasm Classification

Fumiya Isono[1,a)]　Suguru Matsuyoshi[2,b)]　Fumiyo Fukumoto[2,c)]

Abstract: We propose two detection systems that identify sarcasm and slander in posts on bulletin board system (BBS). We made a corpus of sarcasm in BBS, and classified sarcasm instances into eight classes: interrogative, guess, give-up, unbalance, exaggeration, shock, metaphor, and contrast. For each sarcasm class, we constructed syntactic patterns for detection of sarcasm that include sentence structures and polarity conditions of the target sentence, the previous sentence and the next sentence. Our first system detects sarcasm using a database of the syntactic patterns. We made a corpus of slander in BBS and a list of slander expressions extracted from the corpus. Our second system detects slander using Support Vector Machine (SVM), where as features, we use frequencies of words in the list, and positive expressions and negative expressions in the target sentence, the previous sentence and the next sentence. In the experiment, the proposed systems can achieve superior F-measures compared with baseline systems.

Keywords: classification, filtering, sarcasm, slander, bulletin board system

1. はじめに

世の中には，様々なテキストが溢れている．情報の受け手にとって有益なテキストが多数存在する一方で，他人を貶めることを目的とした誹謗中傷文や，対象を面白おかしく非難することを目的とした皮肉なども存在する．次の (1) と (2) に誹謗中傷文の例を，(3) と (4) に皮肉の例を挙げる．

(1) アイツはマジでキモい。
(2) あの世に行って二度と帰ってくんな
(3) A「国土をそして海を汚染して、さらに公的資金を何兆円も (もちろん税金) 投入してボーナス出すのか?」
　B「さすがは優良企業様やで!」

[1] 山梨大学大学院 医学工学総合教育部
Department of Education Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi
[2] 山梨大学大学院 医学工学総合研究部
Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi
[a)] g13mk002@yamanashi.ac.jp
[b)] sugurum@yamanashi.ac.jp
[c)] fukumoto@yamanashi.ac.jp

(4) 建物が古いうえ、それを補う工夫が一切されていないようでした。夏がメインのお宿なのでしょうか？この時期の利用としては、若干閑散とした部分がありますので、古い建物が淋しさ、肌寒さを増徴させていました。

例文 (1) と (2) は，典型的な誹謗中傷であり，「キモい」や「あの世に行って」などの直接的な罵り表現が使用されている．

例文 (3) では，ある企業の行動に対して B が皮肉を使用している (B の発話に下線を引いた)．「さすが～だ」という表現は，対象を褒める時に使用するものであるが，否定的な文脈でこれを使用することにより，B はその企業を非難している．

例文 (4) は，ある宿泊施設に泊まった顧客が述べた感想と苦情である．下線を引いた 2 文目に皮肉が使用されている．利用時期を宿泊施設に問いかける言語形式を使用することにより，「隙間風が入って肌寒かった」ということを間接的に述べている．言語形式としては疑問文であるが，顧客がこの疑問文の回答 (「はい」もしくは「いいえ」) を宿泊施設に求めていないのは，明白である．

いわゆる誹謗中傷文は，その中に特定の語句を含むため，自動認識において，その文に存在する表層的な手がかりが有効的に利用可能であると思われる．一方，対象を非難することを目的とした皮肉文を，その文のみの情報を用いて自動認識することは非常に困難であると思われる．なぜならば，皮肉の解釈は，文脈に大きく左右されるからである．例えば，以下の例文の下線部分には皮肉は感じられない．

(5) A「今年度の顧客満足度でも上位をキープしたそうだ」
B「さすがは優良企業様やで!」

(6) A「夏場は良い旅行プランが多いと感じます。夏がメインのお宿なのでしょうか？」
B「はい。夏に当ホテル周辺でイベントが多くありますので、それをもとにプランを提供させていただいております。」

我々は，自動情報フィルタリングによって，他人を貶めることを目的とした誹謗中傷文や，対象を面白おかしく非難することを目的とした皮肉を排除する機構を構築したいと考えている．この情報フィルタリングは，小学生や中学生など，書き言葉を利用して他者とうまくコミュニケーションする能力がまだ十分発達していない子どもを支援することができると考えている．誹謗中傷文を排除することは当然必要であるが，文脈によって皮肉と解釈されるかもしれない表現も検出することは，誤解のおそれをできる限り少なくするために必要であると思われる．

Web 掲示板の投稿記事やメールに対して，現在運営されているフィルタリングサービス [*1] のほとんどは，1 つの投稿記事全体や 1 通のメール全体を排除する．この仕様は，情報の受け手にとっては適切なものであると思われる．その一方で，書き言葉を用いたコミュニケーション能力がまだ十分発達していない情報発信者を支援することを考えると，文単位や節単位など，もう少し細かい単位で，不適切な箇所を検出できると良いと考える．発信者が書いた文章の中から不適切な文を検出し，不適切であることの理由を提示しながら，その部分を書き改めるように勧めるシステムは，発信者のコミュニケーション能力の向上に貢献すると思われる．

本研究では，不適切であることの理由を提示しながら，その部分を書き改めるように勧めるシステムの第 1 歩として，Web 掲示板の投稿記事から皮肉や誹謗中傷が含まれる文を検出するシステムを構築する．皮肉と誹謗中傷に関してそれぞれコーパスを構築し，どのような語句や言語形式が関係しているか分析する．そして，その分析結果を基に，皮肉文と誹謗中傷文のそれぞれを検出するシステムを独立に作成する．

本論文は，以下のように構成される．2 章では関連研究について述べる．3 章では，宿泊施設のレビューデータと Web 掲示板の投稿記事を対象としたコーパス構築について説明する．4 章と 5 章では，それぞれ，皮肉検出システムと誹謗中傷検出システムを提案する．続く 6 章では，これらのシステムの実験について述べる．7 章で全体をまとめる．

2. 関連研究

本章では，情報の受け手にとって不適切である表現の検出に関する関連研究として，皮肉の自動検出と誹謗中傷の自動検出に関する先行研究を述べる．

2.1 皮肉自動検出

滝澤ら [1] は，言外の意味を含む表現の一つである皮肉や反語などのアイロニーを検出する具体的な手法を提案し，機械によるアイロニー検出の可能性を示した．この手法では，アイロニーの標識と見なせる表現の有無などから，素性ベクトルを構築し，そのベクトルを基にアイロニーの度合いを判定する．彼らは，提案手法の「アイロニー度」と，心理実験によって求める「アイロニー度」との定量的な比較実験が必要であると述べている．

Mihalcea ら [2] は，単純ベイズと Support Vector Machine (SVM) を用いてユーモアのある短文を認識するシステムを提案し，認識に重要な素性を調査した．この調査において，認識に最も頻繁に利用された意味素性は否定表現であることが分かった．

Burfoot ら [3] は，SVM を用いて報道記事から皮肉を検出するシステムを提案した．このシステムは，対象とする皮肉文を報道記事に出現するものと限定することで，高い

[*1] 例えば，http://www.yahoo-help.jp/app/home/p/622/

表 1 構築したコーパス

		全文	皮肉	誹謗中傷
訓練データ	楽天	2,452	37	73
	掲示板	5,141	336	1,247
テストデータ	楽天	2,726	30	95
	掲示板	4,278	234	703
合計	楽天	5,178	67	168
	掲示板	9,419	570	1,950

精度で皮肉文を検出することができた．この手法には，独特な固有表現が含まれている場合に精度が低いという問題があることが述べられている．

Muh ら [4] は，Twitter と Amazon ユーザレビューの 2 つの異なるデータセットに対して，半教師付きのアルゴリズムである SASI[5] を用いて皮肉的な文を自動認識する手法を提案した．彼らは，データセットから抽出した特徴的な構文パターンを主な素性として利用し，k-近傍法により皮肉かどうかの判定を行う．彼らは，構文パターンの他に，出現した内容語や記号 (“!” や “?” や引用符など) の数を素性として利用している．評価実験において，彼らの手法は，Amazon ユーザレビューに対しては高い精度を出したが，Twitter に対しては，利用できる文脈の情報が非常に限られるため，再現率が非常に低い結果となった．

2.2 誹謗中傷自動認識

松葉ら [6] は，学校非公式サイトの投稿記事から，SVM を用いて有害情報を自動的に検出するシステムを提案した．このシステムは，事前に構築した有害単語辞書と，有害単語間の共起関係を利用して，有害な文であるかどうかを判定する．実際のデータでは，一度しか出現しない有害単語の組が全体の 8 割を占めていたため，共起関係の情報は有効ではなかったと彼らは述べている．

Adler ら [7] は，機械学習を利用して，Wikipedia の記事から荒らし記事を検出する手法を提案した．この手法では，メタデータ，テキスト本体，言語的特徴，評判の 4 種類の情報と，その組み合わせを素性として用いる．

メタデータ 記事が最後に編集されてからの経過時間，改訂コメントの長さなど

テキスト本体 記事に対する大文字の比率，加筆された単語数など

言語的特徴 代名詞の頻度，不適切語の頻度など

評判 編集を行ったユーザの評判，編集が行われた地域など

実験によって，精度に一番貢献した素性は，言語的特徴であり，4 種類の素性を組み合わせることで精度が大きく向上することが確認された．

3. コーパス構築

テキストデータに対して，皮肉であるかどうかと誹謗中傷であるかどうかを，それぞれ人手でラベル付けした．本研究では，次の 2 種類のテキストデータを用いた．

(1) 楽天トラベル: レビューデータ

(2) Web 掲示板

3.1 楽天トラベル: レビューデータ

本研究では，まず，楽天データの楽天トラベル: レビューデータ [*2] を用いて，皮肉と誹謗中傷のラベルを付けたコーパス (以下，「楽天」コーパス) を構築した．

対象としたレビュー集合は，小池ら [8] が使用したものと同じである．90% 以上の宿泊施設はレビュー数が 1 から 58 の範囲にあるという調査結果に基づき，レビュー数が 10 から 58 の範囲の宿泊施設の全体から，無作為に 40 の宿泊施設を抽出し，ラベル付けの対象とした．独自の文分割規則により半自動的に文分割を行い，5,178 文のテキストデータを得た．

主に作業者 1 人によって，皮肉と誹謗中傷のラベル付けを行った．本研究で扱う皮肉は，誹謗中傷の 1 種と考えることができるので，両方のラベルに該当する文については，皮肉のラベルのみを付与した．ラベル付けの判断が難しい文に関しては，新たにもう 1 人の作業者を加え，作業者 2 人によってラベルを定めた．ラベル付与の結果を表 1 の「楽天」の行に示す．このコーパスに存在する皮肉は 67 文，誹謗中傷は 168 文であった．

3.2 Web 掲示板

次に，インターネット上に存在する 5 つの Web 掲示板 [*3] からテキストデータを収集し，これを対象として，皮肉と誹謗中傷のラベルを付けたコーパス (以下，「掲示板」コーパス) を構築した．

くだけたテキストであることを考慮し，次の前処理を行った．

- 句点と “!” と “?” での文分割に加え，改行の位置でも文分割
- 複数行に渡ったアスキーアートを除外

これらの前処理により，9,419 文のテキストデータを得た．

前節と同様に，主に作業者 1 人によって，皮肉と誹謗中傷のラベル付けを行った．ラベル付与の結果を表 1 の「掲示板」の行に示す．このコーパスに存在する皮肉は 570 文，誹謗中傷は 1,950 文であった．

[*2] http://rit.rakuten.co.jp/rdr/index.html
[*3] http://blog.livedoor.jp/dqnplus/archives/1736747.html
http://blog.livedoor.jp/dqnplus/archives/1736731.html
http://blog.livedoor.jp/dqnplus/archives/1735211.html
http://hamusoku.com/archives/7126094.html
http://hamusoku.com/archives/7430403.html
(いずれも 2012 年 12 月 13 日にアクセス)

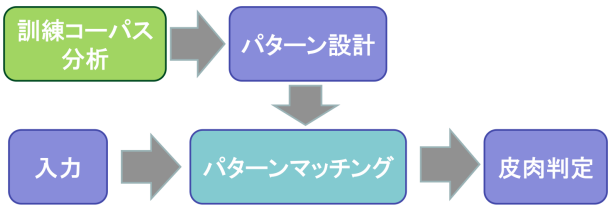

図 1 提案する皮肉検出システム

3.3 訓練データとテストデータへの分割

構築した「楽天」コーパスと「掲示板」コーパスを，それぞれ訓練データとテストデータに分割した．

「楽天」コーパスにおいては，40 の宿泊施設のレビューのうち，ランダムに選択した 20 の宿泊施設のレビューデータ 2,452 文を訓練データとし，残り 20 の宿泊施設のレビューデータ 2,726 文をテストデータとした．「掲示板」コーパスにおいては，3 つの Web 掲示板のテキストデータ 5,141 文を訓練データとし，残り 2 つの Web 掲示板のテキストデータ 4,278 文をテストデータとした．

上記の分割と，分割後のラベルの数をまとめたものを表 1 に示す．

4. 皮肉検出システム

本研究では，皮肉文と誹謗中傷文を独立の検出システムでそれぞれ検出する．この章では，提案する皮肉検出システムについて述べる．皮肉検出システムの概要を図 1 に示す．

4.1 皮肉の分類

1 章で述べたように，「特定の語句が出現した」という狭い範囲の情報のみでは，高い精度での皮肉検出は困難である．そこで，我々は，コーパス内の皮肉を人手で分類し，検出に必要な情報について分析した．

まず，我々は，皮肉の分類に関して言語学の文献の調査を行った．レビューデータや Web 掲示板における皮肉にも有効であるような分類体系を求めたが，そのような分類は存在しないことが分かった．そこで，2 つのコーパスの訓練データを観察し，皮肉を主に次の 8 つに分類した．

疑問，推測，諦め，不相応，誇張，驚き，形容，対比

それぞれの分類クラスにおける，コーパスの訓練データ中の事例数を表 2 に示す．各分類クラスの説明は，構文パターンの説明とともに次節で行う．

4.2 構文パターンの設計

皮肉の解釈は，文脈に大きく左右されるため，皮肉の検出時には文脈を考慮する必要がある．文脈利用の第 1 歩として，我々は，対象の文の直前と直後の文を考慮に入れて皮肉かどうかを判定する．

前節で述べた分類クラスごとに，そのクラスを特徴付ける構文パターンを設計した．この構文パターンは，対象文か前後文に否定的な語句が存在するかどうかの情報を利用する．本研究では，否定的な語句のリストとして，日本語評価極性辞書 [*4] を使用した．構文パターンを設計するにあたり，検出の F 値とともに，再現率を優先した．これは，情報フィルタリングの目的上，不適切な表現の誤検出が少ないことよりも，検出漏れが少ないことの方が重要であると考えたためである．

表 3 に，我々が設計した構文パターンの一覧を示す．構文パターンは，全部で 35 個ある．「対象文の形式」の欄に，対象文に対して適用される構文を示す．表では略記しているが，実際には係り受けの情報も利用する．“Neg +” は，対象文内に否定的な語句が存在し，その後ろの表現に対して否定的な語句が係っていることを表す．「直前」の欄に “Neg” と記載されている場合，対象文の直前の文内に否定的な語句が存在することを表す．「直後」の欄に “Neg” と記載されている場合も同様である．形態素解析用辞書 UniDic[*5] の語彙素を利用することにより，ある程度の表記の揺れに対応している．

以下，皮肉の各分類クラスとその構文パターンについて説明する．

4.2.1 疑問

言語形式の上では疑問文であるが，実際には回答を求めるのではなく，非難を述べることを目的とする．

以下に例を挙げる (下線が皮肉文)．

- 建物が古いうえ、それを補う工夫が一切されていないようでした。夏がメインのお宿なのでしょうか？
- 無銭飲食を疑われたようで、大変不快でした。朝食券を渡してあることはすぐに伝わりましたが、客に聞く前にまず店員間で確認をすべきではないでしょうか。

対象の文が疑問文であり，主に直前の文に否定的な語句が含まれるという構文パターンにより，このクラスの皮肉を検出する．

4.2.2 推測

言語形式の上では推測を述べているが，実際には誹謗中傷ととれる内容を述べることを目的とする．

以下に例を挙げる (下線が皮肉文)．

- 買い物依存の逆版みたいなセツヤクブームにのってるだけなんだろうな
- ついでに浮気してるか、ガソスタのお兄さんがイケメンかとかそんなことじゃね

対象の文に推測の言語形式が含まれており，対象文か直前の文に否定的な語句が含まれるという構文パターンによ

[*4] http://www.cl.ecei.tohoku.ac.jp/resources/sent_lex/wago.121808.pn
[*5] http://sourceforge.jp/projects/unidic/releases/57618

表 2 皮肉の分類と事例数

分類	楽天	掲示板	合計	例文 (複数文の場合，下線部が皮肉文)
疑問	15	104	119	原発に関わってるとネジ飛んじゃうのかな…？
推測	0	44	44	節約が趣味なんだろｗ
諦め	0	68	68	こういうのは言ってもきかないんだよ、それが楽しみのひとつなんだから。
不相応	8	48	56	こちらにと誘導され駐車したのにもかかわらず翌朝には遠い場所にいれかえられていました。高いお部屋の方やいい車などと入れ替えてるようで少し嫌な気持ちになります。
誇張	0	51	51	すげー会社だな ww
驚き	3	0	3	今までに階上がこんなにもうるさいってことが無かったからで、かなりびっくりしました。
形容	5	0	5	しかし今回ばかりは、感心するくらい狭かった…。 品のいい刑務所のようでした。
対比	6	0	6	部屋のソファーは、汚れていて座りたい気分ではなかった。 立地条件や食事はとてもよかったです。
その他	0	21	21	社員が高級外車乗ってるのも納得

り，このクラスの皮肉を検出する．

4.2.3 諦め

言語表現の上では，何かを諦めたようなことを表明しているが，実際には，対象を非難することを目的とする．

以下に例を挙げる (下線が皮肉文)．

- こういうのは言ってもきかないんだよ、それが楽しみのひとつなんだから。
- 動物の習性って思って諦めるしかないな。

対象の文に，諦めを表明する言語形式と否定的な語句が含まれるという構文パターンにより，このクラスの皮肉を検出する．

4.2.4 不相応

想定していたことと釣り合わない，もしくは，条件や環境に相応しくないことに対して非難することを目的とする．

以下に例を挙げる (下線が皮肉文)．

- こちらにと誘導され駐車したのにもかかわらず翌朝には遠い場所にいれかえられていました。高いお部屋の方やいい車などと入れ替えてるようで少し嫌な気持ちになります。
- コミュニケイションってやつは相手の馬鹿さ加減も受け入れないと成立しない
  知能に問題がある

対象の文に，逆接の表現か，想定したほど存在しないことを表明する表現が含まれ，直後の文に否定的な語句が含まれるという構文パターンにより，このクラスの皮肉を検出する．

4.2.5 誇張

対象を強く褒める言語形式により，誹謗中傷することを目的とする．

以下に例を挙げる (下線が皮肉文)．

- 国土をそして海を汚染して、さらに公的資金を何兆円も（もちろん税金）投入してボーナス出すのか？さすがは優良企業様やで！
- すげー会社だな ww

対象の文に「さすが」や「すごい」が含まれ，直前の文に

表 3 皮肉検出のための構文パターンの一覧

分類	直前	対象文の形式	直後
疑問	Neg	～でしょうか	–
	–	Neg ＋ ～でしょうか	–
	Neg	～なんですか	–
	Neg	～?	–
	–	Neg ＋ ～?	–
	Neg	～なの	–
	Neg	～ないの	–
	Neg	～ものか	–
	Neg	～では	–
推測	Neg	～なんだろう	–
	–	Neg ＋ ～じゃね	–
諦め	–	Neg ＋ ～なんだから	–
	–	Neg ＋ ～しかない	–
	–	Neg ＋ ～してもむだ	–
不相応	–	～のに ··· された	Neg
	–	～だったが ··· された	Neg
	–	～ないと ··· ない	Neg
	–	～なければ ··· ない	Neg
	–	～にもかかわらず ··· ない	Neg
	–	～ほど ··· ない	Neg
	–	なかなか～ません	Neg
	–	かなり～ない	Neg
	–	～しそうになる	Neg
	–	～どうでもよい	Neg
誇張	Neg	さすが～	–
	Neg	すごい～だな	–
	–	すごい～w	–
驚き	–	Neg ＋ ～おどろいた	–
	–	Neg ＋ ～びっくりした	–
	Neg	～はじめてのけいけん	–
形容	–	Neg ＋ ～みたい	–
	Neg	～のよう	–
対比	Neg	～はよい	–
	–	Neg ＋ ～なければよい	–
	–	Neg ＋ ～なければすばらしい	–

否定的な語句が含まれるという構文パターンにより，このクラスの皮肉を検出する．

4.2.6 驚き

特別驚いたことを表明することにより，その内容を非難することを目的とする．

以下に例を挙げる (下線が皮肉文)．

- 今までに階上がこんなにもうるさいってことが無かったからで、かなりびっくりしました。
- 今日は相部屋・・・・恐る恐る部屋に入って「こんばんわ、一晩よろしく」・・・・そんな訳ないよなぁとよく見るとベットメイクされていませんでした。長いサラリーマン出張生活で始めての経験。[*6]

対象の文に，驚きや初めてを表明する言語形式が含まれ，対象文か直前の文に否定的な語句が含まれるという構文パターンにより，このクラスの皮肉を検出する．

4.2.7 形容

明喩を用いて対象を誹謗中傷することを目的とする．

以下に例を挙げる (下線が皮肉文)．

- 風呂やトイレもドアが全部なくていなかの韓国のホテルみたいでした。
- しかし今回ばかりは、感心するくらい狭かった…。品のいい刑務所のようでした。

対象の文に明喩が含まれ，対象文か直前の文に否定的な語句が含まれるという構文パターンにより，このクラスの皮肉を検出する．

4.2.8 対比

悪い所と対比させて良い所も述べることを目的とする．他の分類クラスに比べ，皮肉の度合いはかなり低い．

以下に例を挙げる (下線が皮肉文)．

- 部屋のソファーは、汚れていて座りたい気分ではなかった。立地条件や食事はとてもよかったです。
- 浴槽が窮屈でなければ素晴らしいお宿でした。

対象の文に，良い所を褒める表現が含まれ，対象文か直前の文に否定的な語句が含まれるという構文パターンにより，このクラスの皮肉を検出する．

4.3 検出処理

本手法では，前節で説明した構文パターンを単純なパターンマッチングに利用することにより，皮肉文を検出する．

まず，入力された文章を文分割し，形態素解析器 MeCab[*7] と構文解析器 CaboCha[*8] により解析する．入力の構文解析結果と，構文パターンを比較することにより，その文が皮肉文であるかどうかを判定する．比較する構文パターンの順番は，訓練データを用いて事例数を確認しながら試行錯誤することにより定めた．

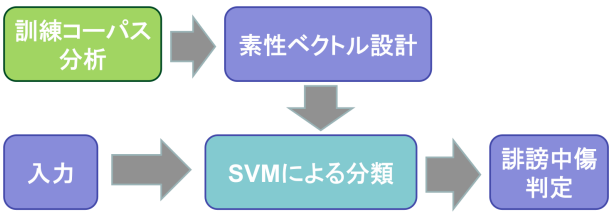

図 2 提案する誹謗中傷検出システム

日本語評価極性辞書の否定的な語句が存在する場合のみでなく，助動詞「ない」などの否定と係り受けの関係にある肯定的な語句が存在する場合も，「否定的な語句が含まれている」状況として扱った．同様の考え方で，否定的な語句が存在する場合でも，それが否定と係り受けの関係にある場合は，「否定的な語句が含まれている」状況と見なさなかった．

5. 誹謗中傷検出システム

この章では，提案する誹謗中傷検出システムについて述べる．誹謗中傷検出システムの概要を図 2 に示す．

5.1 誹謗中傷語辞書

ある文が誹謗中傷の内容を含むかどうかは，表層的な語句の手がかりにより判定することが可能であると思われる．本手法では，特徴的な語を素性とした機械学習手法を用いて誹謗中傷文を検出する．

まず，我々は，訓練データ内の誹謗中傷文から特徴的な語を抽出し，独自の誹謗中傷語辞書を編纂した．次の値が 3 以上の語を誹謗中傷語と認定した．

- 「その語が出現する誹謗中傷文の数」− 「その語が出現する誹謗中傷文でない文の数」

構築した辞書には，以下のような語が 112 語収録されている．

アホ，ウザい，くたばる，くさい，馬鹿，ハゲ，悪魔，鬼，キチガイ，殺し，消え去る，害，害悪，悲報，犯，欠ける，低能，爆発，爆ぜる，不味い，怪しい，腐る，裸，塵，屑，醜い，辞任，不愉快，尋常，やれやれ，軽視，地獄，亡者，情けない，年寄り，老害，土下座，括り，諸共，死人，死体，死骸，劣る，劣等，同類，同種，強奪，擦る，狂う，発狂，無恥，放棄，末路，欠陥，W，WWW

5.2 素性

本手法では，以下の項目を素性として利用した．4.3 節と同様に，肯定的な語句と否定的な語句の判定には，否定との係り受け関係を考慮する．

対象の文内

[*6] 「始めて」は，コーパス内の実事例における誤字．
[*7] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[*8] http://code.google.com/p/cabocha

- すべての語の出現頻度
- 前節の誹謗中傷語の出現頻度
- 肯定的な語句の合計出現頻度
- 否定的な語句の合計出現頻度

直前の文内

- 肯定的な語句の合計出現頻度
- 否定的な語句の合計出現頻度

直後の文内

- 肯定的な語句の合計出現頻度
- 否定的な語句の合計出現頻度

5.3 検出処理

本手法では，機械学習手法として SVM を用いる．

まず，4.3 節と同様に，入力された文章を文分割し，構文解析する．そして，対象文ごとに，前節の素性を抽出し，素性ベクトルを構築する．訓練データのうち，誹謗中傷文を正例，そうでない文を負例として SVM の学習を行った．

6. 実験

6.1 実験設定

提案手法である皮肉検出システムと誹謗中傷検出システムの評価実験を行った．実験には，コーパスの訓練データとテストデータを用い，クローズドテストとオープンテストを行った．誹謗中傷検出システムにおいては，SVM ツールとして SVM-light[*9] を利用し，5 分割交差検定を実施した．

比較対象として，単純なキーワードマッチングで検出を行うベースラインシステムを構築した．訓練データを利用して，皮肉と誹謗中傷のそれぞれに対して，次のようにしてキーワードのリストを定めた．

皮肉検出 「その語が出現する皮肉文の数」− 「その語が出現する皮肉文でない文の数」 ≥ 2

誹謗中傷検出 「その語が出現する誹謗中傷文の数」−「その語が出現する誹謗中傷文でない文の数」 ≥ 2

評価尺度として，適合率 P と再現率 R と F 値を用いた．

$$\text{適合率 } P = \frac{\text{正例と分類された正例事例数}}{\text{正例と分類された事例数}}$$

$$\text{再現率 } R = \frac{\text{正例と分類された正例事例数}}{\text{データ中の正例事例数}}$$

$$\text{F 値} = \frac{2PR}{P+R}$$

6.2 実験結果と考察

皮肉検出と誹謗中傷検出に関する，訓練データを用いたクローズドテストの結果を，それぞれ表 4 と表 5 に示す．

皮肉検出では，提案手法は，再現率を優先して構文パターンの設計を行ったため，表 4 の「再現率」の列を見ると，適合率に比べ非常に高い再現率が得られたことが分かる．その一方で，適合率はかなり低いことも分かる．

[*9] http://svmlight.joachims.org

表 4 皮肉検出のクローズドテスト

		適合率	再現率	F 値
ベース	楽天	0.04 (35/921)	0.95 (35/37)	0.07
	掲示板	0.08 (326/4,075)	0.97 (326/336)	0.15
提案	楽天	0.20 (37/185)	1.00 (37/37)	0.34
	掲示板	0.21 (211/994)	0.63 (211/336)	0.32

表 5 誹謗中傷検出のクローズドテスト

		適合率	再現率	F 値
ベース	楽天	0.04 (72/1,782)	0.99 (72/73)	0.08
	掲示板	0.25 (1,234/4,981)	0.99 (1,234/1,247)	0.40
提案	楽天	0.33 (71/212)	0.97 (71/73)	0.50
	掲示板	0.51 (560/1,104)	0.45 (560/1,247)	0.48

表 6 皮肉検出のオープンテスト

		適合率	再現率	F 値
ベース	楽天	0.01 (6/907)	0.20 (6/30)	0.01
	掲示板	0.06 (147/2,435)	0.63 (147/234)	0.11
提案	楽天	0.09 (14/150)	0.47 (14/30)	0.16
	掲示板	0.09 (102/1,150)	0.44 (102/234)	0.15

表 7 誹謗中傷検出のオープンテスト

		適合率	再現率	F 値
ベース	楽天	0.04 (93/2,408)	0.97 (93/95)	0.07
	掲示板	0.17 (685/4,045)	0.97 (685/703)	0.29
提案	楽天	0.13 (60/452)	0.63 (60/95)	0.22
	掲示板	0.38 (449/1,176)	0.64 (449/703)	0.48

表 5 を見ると，誹謗中傷検出では，提案手法は，「楽天」コーパスにおいて 0.97 という高い再現率を得られた．その一方で，適合率は 0.33 と，決して高い値ではなかった．「掲示板」コーパスにおいては，再現率が適合率を下回るという現象が起き，どちらも 0.50 付近であり，高い値ではなかった．

次に，皮肉検出と誹謗中傷検出に関する，テストデータを用いたオープンテストの結果を，それぞれ表 6 と表 7 に示す．

皮肉検出では，表 6 から読み取れる通り，「掲示板」コーパスにおける再現率は，提案手法がベースラインを下回ったが，それ以外の値はベースラインを上回った．提案手法の適合率は，ベースラインを上回ったものの，両コーパスにおいていずれも 0.09 と非常に低い．これは，再現率を優先して構文パターンを設計した結果，構文パターンで捉えるべき範囲の制約が緩くなってしまい，誤検出が多くなったためであると考えられる．従って，誤検出された皮肉文でない文を調査し，構文パターンを見直し，厳密な構文パターンを再設計する必要がある．テストデータを観察すると，8 つの分類クラスのいずれにも分類しがたい事例が存在した．これらの事例のために適切に新しい分類クラスを構築し，そのクラスに対する新たな構文パターンを設計する必要がある．

誹謗中傷検出では，表 7 に示されるように，両コーパスにおいて，提案手法は適合率と F 値でベースラインより高い結果が得られた．その一方で，再現率はベースラインより低い結果となった．これは，素性ベクトルの設計が十分でなかったためと考えられる．従って，誹謗中傷文の特徴をさらに調査することにより，辞書の改訂と素性ベクトルの設計に注力する必要がある．

7. まとめ

本研究では，Web 掲示板の投稿記事やレビューデータから皮肉や誹謗中傷が含まれる文を検出するシステムを構築した．このシステムは，構文パターンや辞書と，前後文の文脈情報を利用することにより，皮肉と誹謗中傷を検出する．

本研究で提案するシステムの精度は高くなかったため，さらなる改善が必要である．構築したコーパスを分析することで，皮肉の分類を精緻化し，厳密な構文パターンを収集する必要がある．誹謗中傷に関しては，適切な素性の集合を設計することが今後の課題である．

謝辞 本研究の一部は，科研費若手研究 (B)「否定焦点コーパス構築と焦点自動解析に関する研究」(課題番号: 25870278，代表: 松吉俊) の支援を受けている．

参考文献

[1] 滝澤修, 伊藤昭：アイロニー表現検出の一手法, 人工知能学会誌, Vol. 9, No. 6, pp. 875–881 (1993).
[2] Mihalcea, R. and Pulman, S. G.: Characterizing humour: An exploration of features in humorous texts, in *CICLing*, pp. 337–347 (2007).
[3] Burfoot, C. and Baldwin, T.: Automatic satire detection: Are you having a laugh?, in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 161–164 (2009).
[4] Muh, M., Tsur, O. and AriRappoport, : Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon, in *Proceedings od the Fourteenth Conference on Computational Natural Language Learning*, pp. 107–116 (2010).
[5] Tsur, O., Davidiv, D. and Rappoport, A.: Icwsm - A Great Catchy Name: Semi-supervised Recognition of Sarcastic Sentences in Product Reviews, in *International AAAI Conference on Weblogs and Social Media*, pp. 162–169 (2010).
[6] 松葉達明, 里見尚宏, 桝井文人, 河合敦夫, 井須尚紀：学校非公式サイトにおける有害情報検出, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, pp. 93–98 (2009).
[7] Adler, B., Alfaro L., de , Mola-Velasco, S., Rosso, P. and West, A.: Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features, in *ICLing '11: Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics, LNCS 6609*, pp. 277–288 (2011).
[8] 小池惇爾, 松吉俊, 福本文代：評価視点別レビュー要約のための重要文候補抽出, 言語処理学会第 18 回年次大会 発表論文集, pp. 1188–1191 (2012).