The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023

皮肉文検出のための皮肉状況の検出

Detecting sarcastic expressions for extracting sarcasm in a text

畑　玲音 *1
Reon Hata
森野　穣 *1
Yutaka Morino
松下 光範 *1
Mitsunori Matsushita

*1関西大学
Kansai University

Sarcasm is a linguistic phenomenon in which a target's faults and weaknesses are criticized indirectly and is characterized by using positive expressions to convey negative meanings. For this reason, an expression's superficial meaning and intention don't match in sarcastic sentences. When using a computer to determine whether an affirmative expression appears in a sentence is sarcasm, it's necessary to extract the three elements of sarcasm together: the target of the sarcasm, the situation that is the premise of the sarcasm, and the sarcastic expression. Although previous studies have attempted to identify sarcastic sentences by extracting the target of sarcasm, there are many sarcastic sentences in which the target of sarcasm is not explicitly stated. Therefore, in this study, we create a classifier for detecting sarcastic situations from sentences expressing situations using BERT and attempt to extract sarcastic sentences in which the target of the sarcasm isn't explicitly stated.

1. はじめに

近年，Social Networking Services （以下 SNS と記す） 利用者の増加に伴い，SNS 上での誹謗中傷が問題となっている．誹謗中傷とは，他人へ悪口を言ったり罵ったりする行為である「誹謗」と，根拠のない嘘やでたらめを述べる行為である「中傷」の語を組み合わせたものであり，デマや揶揄，罵倒，皮肉，嫌がらせなど文字や言葉を用いた攻撃行為である． 誹謗中傷の対策の一つとして，それに気づいたユーザが SNS の運営者に通報する方法が挙げられる．この場合，運営者は投稿を確認し，必要に応じて投稿者への警告や削除，アカウント凍結などの対応を行う．しかし，ユーザによる通報や運営者による対処は人手により行われており，SNS 上からその誹謗中傷の投稿が削除されるまでに，被害者の目に入ってしまうことが大半である．特に Twitter の投稿数は 1 分間に約 34 万件 *1 と非常に多いため，人手で一つずつ誹謗中傷を確認することには限界がある．

そこで，膨大な SNS の投稿から誹謗中傷を自動検出するための研究が行われている [石坂 11, 伊藤 21]．これらの研究では，誹謗中傷に用いられる表現の検出や感情分析による誹謗中傷文の検出を試みている．

誹謗中傷には，ひと目で悪口とわかる表現を用いたものだけではなく，皮肉のように，表層的には悪口にはみえない場合も存在する．皮肉は対象の欠点や弱点を遠回しに批判する言語現象であり，肯定表現を用いて否定的な意味を伝えるという特徴を持つ．そのため，皮肉文では表現の表層的な意味とその意図が一致しない．例えば，ゴミが多い「家」を対象（以下，皮肉対象と記す）にした「とても綺麗だね」という肯定表現は皮肉を表現する文（以下，皮肉表現と記す）になり，対象が新築の「家」であるような場合には，同一文章であっても皮肉表現にはならない．

連絡先: 畑玲音，関西大学総合情報学部，〒 569-1095，大阪府高槻市霊山寺町 2-1-1，Tel:072-690-2437，Fax:072-690-2491，k223167@kansai-u.ac.jp

*1 https://www.inc.com/larry-kim/15-mind-blowing-statistics-reveal-what-happens-on-the-internet-in-a-minute.html

皮肉には肯定表現が使われる場合が多いため，書いた文章が意図せずに皮肉として解釈される状況がしばしば発生する．意図しない誹謗中傷を防ぐために，計算機を用いて皮肉を検出することは重要な課題であると言える．

これまでに，皮肉文の検出を試みた研究も存在する．肥合らは，皮肉を「否定的な意味を伝える肯定表現」と定義し，機械学習を用いて皮肉対象となる人やものを検出し，それを含む文章を皮肉文として抽出することを試みている [肥合 18]．この研究では皮肉対象が対象文中に明示されている必要があるため，皮肉対象が非明示である場合には皮肉文として抽出することができない．そこで本研究では，皮肉対象が明示されない場合でも皮肉文を抽出できるようにする手法について検討する．

2. 皮肉文の特徴

本研究においても，肥合らの定義に倣い，皮肉を「否定的な意味を伝える肯定表現」と捉え，皮肉を含む文の検出を試みる．皮肉文の判定基準を以下の文を例に説明する．

例 1 君は講義中に呑気にゲームができて楽しそうだね.

例 2 始発から遅延なんて幸先がいいね.

例 3 君はご飯を食べているとき楽しそうだね.

例 1 の文では,「君」という皮肉対象が明示され,「楽しそうだね」という肯定表現を用いているため，皮肉文と判断できる．例 2 の文は，電車やバスなどが対象となる皮肉であり，その対象が非明示な文である．人はこの文の「始発」や「遅延」という単語から電車やバスなどの対象を想像する．その対象が「始発から遅延している」ことに肯定表現を用いているため，皮肉であると判断できる．

例 3 の文では，例 1 と同様に「君」が対象であり,「楽しそうだね」という肯定表現を用いているものの皮肉文ではないと判断するのが妥当だろう．例 1 の「君」が皮肉対象と判断され，例 3 の「君」が皮肉対象ではないと判断されることは，皮肉対象の判定が文の対象ならびに肯定表現以外の情報を考慮する必要があることを示唆している． 本研究では，皮肉文における

肯定表現が潜在的に伝える “否定的な意味” は，その皮肉の前提となる状況に起因すると考え，これを皮肉状況と定義する．

例 1 では「君」の「講義中に呑気にゲームができる」状況は否定的な意味を持ち，皮肉状況である．例 2 では「電車やバス」が「始発から支援している」状況は否定的な意味を持ち，皮肉状況である．例 3 の「君」が「ご飯を食べている時」という状況は，否定的な意味を持たないため，皮肉状況とはならない．この皮肉状況を検出することが可能になれば，皮肉対象が明示されていない場合の皮肉文検出も可能になると考える．

文中から皮肉状況を検出するには，皮肉状況文とそれ以外の文を判断する必要がある．そこで本稿では，皮肉文を検出する端緒として，状況を表す文からの皮肉状況の検出を試みる．

3. 皮肉状況検出の提案手法

状況文には，皮肉状況とそれ以外の状況（以下，非皮肉状況と記す）が存在する．状況文から皮肉状況を検出する方法として，状況文を皮肉状況とそれ以外の状況に分類することで検出する．状況文を分類するために，分類器の作成を行う．分類器の作成には状況文を集めた，状況文コーパスが必要である．この分類器が，皮肉文に含まれる皮肉状況が検出可能か検証する．そのため，皮肉文に含まれる皮肉状況を表す文のみを集めた，皮肉状況文コーパスが必要である．

状況文コーパスでは，TwitterAPI を用いて，皮肉状況と非皮肉状況の文を収集し作成する．分類器の作成には，Bidirectional Encoder Representations from Transformers（BERT）[Jacob 18] を用いた．

BERT を用いて皮肉の検出を試みた研究としては Hankyol らの研究 [Hankyol 20] や諏訪らの研究 [諏訪 21] が挙げられる．これらの研究では，BERT に皮肉と非皮肉の文をデータセットとして学習しているため，誤判断された原因が不明である．そこで，皮肉文を皮肉対象・皮肉状況・皮肉表現などの要素に分解し学習することにより，その原因を解明することが必要であると考える．皮肉状況文コーパスでは，TwitterAPI を用いて皮肉文を収集し，皮肉文から皮肉状況を抽出する．

4. データセット

本研究では Twitter の投稿（以下，ツイートと記す）を収集し，状況文コーパスと皮肉状況文コーパスのデータセットの作成を行う．ツイートは TwitterAPI を用いることにより取得した．先行研究では,「#皮肉」,「(皮肉)」を採用している [肥合 18]．ハッシュタグが用いられている投稿は，そのキーワードの内容が含まれた投稿である可能性が高いため,「#皮肉」では皮肉の内容を示す文章を取得できる．また，文章が皮肉であることを意図して，投稿者は文末に「(皮肉)」をつける傾向があるため,「(皮肉)」をクエリとすることで，皮肉の内容を示す文章の取得が期待される．しかし,「#皮肉」や「(皮肉)」をクエリとすると皮肉文を含むツイートの収集は可能であるものの，その中に皮肉状況を含むツイートが含まれるとは限らない．そこで本研究では,「皮肉」をクエリとし，2022 年 5 月から 7 月に投稿されたツイートから 161,693 件のツイートを収集した．これらを用いてコーパスを作成する．なお，学習時のノイズ低減のために，収集したツイートからアカウント名と URL の削除を前処理として行った．

4.1 状況文コーパス

状況文コーパスの文が満たす条件として，まず，状況を表す文が含まれている必要がある．次に，皮肉状況と非皮肉状況を分類するため，その状況を表す文が皮肉状況が含まれている文と非皮肉状況が含まれている文である必要がある．最後に，文を学習し，皮肉状況を検出するため，その文だけで状況を判断できる必要がある．

皮肉状況が含まれている文は,「皮肉」のクエリで収集した文から,「と（い | ゆ | 言）う皮肉」が含まれている文を抽出した.「と（い | ゆ | 言）う」は，事柄を取り立てて強調するときに使う表現であり事柄とは，その物事がどのような様子を表しているかを示す.「と（い | 言）う皮肉」は，その前に書かれている様子を強調し皮肉であると表現したいときに用いられる．そのためこの表現は，皮肉だと投稿者が感じるときに用いられ，皮肉表現を含まない.「と（い | ゆ | 言）う皮肉」の前の文を抽出することにより，皮肉状況を表す文のみを収集することができる．Twitter の口語性を考慮し,「って（い | ゆ | 言）う皮肉」や「とか（い | ゆ | 言）う皮肉」などの類似した表現も収集する．そのため,「う皮肉」をクエリとして抽出した．以下に，皮肉状況が含まれている文の例を挙げる（下線部が皮肉状況）．

例 4 裁判官が法を犯すという皮肉.

「皮肉」のクエリで収集した文のうち,「う皮肉」は 7,966 件存在した．状況文コーパスの作成にあたり，文章内に「皮肉」と言う単語が含まれている場合，この単語が手がかり語となり学習されることが懸念される．それを避けるため,「皮肉」という単語を削除し皮肉状況のみを抽出する必要がある．例 5 であれば,「という皮肉」を削除することにより皮肉状況のみを抽出することができる.「う皮肉」で抽出した文には全て「皮肉」という単語が含まれており，文から「と（い | ゆ | 言）う皮肉」を削除する前処理を行い皮肉状況文とした．

非皮肉状況が含まれている文が満たすべき条件は，皮肉状況ではないことと，皮肉状況と条件を合わせるため皮肉状況と同様の方法により収集と前処理を行えることである．以上の条件を満たす文は,「皮肉」をクエリとし収集した文には存在しないため，TwitterAPI を用いて新たに収集を行った．クエリとして「奇跡」を選択し,「奇跡」という単語が入っている文を取得した．取得した文のうち，非皮肉状況では，皮肉状況を表すときには用いることがない「と（い | ゆ | 言）う奇跡」を採用し抽出した．皮肉状況と同じように Twitter の口語性を考慮し「う奇跡」を抽出し,「と（い | ゆ | 言）う奇跡」などの部分を削除しデータの整理を行った．これにより 2,514 件のデータを収集した．以下に例を示す (下線部が非皮肉状況).

例 5 推しと会話できたっていう奇跡の展開でした.

この文を状況コーパスにおける非皮肉状況を表す文として使用する．

これらの皮肉状況を表す文と，非皮肉状況を表す文章を合わせて状況文コーパスを作成した．学習時に件数による差が生じないように，それぞれランダムに 2,500 件ずつ抽出し，5,000 件を状況文コーパスとした．

4.2 皮肉状況文コーパス

皮肉状況文コーパスは，作成した分類器を用いて皮肉文に用いられている皮肉状況を検出できるかを検証するために必要である．皮肉文に含まれている皮肉状況を抽出することにより皮肉状況文とする．皮肉状況の抽出は，皮肉文から皮肉表現となる文を削除することにより行う．そのため，まず,「皮肉」のクエリで収集した文から皮肉文を抽出する．

先行研究に倣い,「#皮肉」や「(皮肉)」を含む文を皮肉文候補として抽出した．得られた 6,116 文に対して，クラウドソー

シングを利用し人手により皮肉文か否かを判断させた．この際，状況文コーパスの作成と同様に，文脈情報以外（アカウント名を示す@や URL を示す https など）の削除と手がかり語となる「皮肉」を削除する前処理を行った．クラウドソーシングには Yahoo!クラウドソーシングを用いた．協力者数は 647 人であった．各ツイートにつき 3 人を割り当て，1 つのツイートに対し 2 人以上が皮肉と回答したツイート文を皮肉文として採用した．これにより，3,587 件を皮肉文として収集した．

次に，皮肉文に含まれている皮肉表現の文の抽出について説明する．以下に，皮肉表現が含まれている皮肉文の例を示す (下線部が皮肉表現)．

例 6 日本ってつくづく平和ですね.

この文から下線部箇所を削除することにより，皮肉状況を抽出することができると考える．収集した皮肉文に対して，Yahoo! クラウドソーシングを用いて 390 人に皮肉表現の記述箇所にアノテーションを付与させ，皮肉文に含まれる皮肉表現を収集し，この皮肉表現を皮肉文から削除することで皮肉状況を抽出した．ここで，例 6 のように皮肉表現を削除してしまうと皮肉状況が残らない文は除外し，3,094 件を皮肉状況文コーパスとして採用した．

5. 皮肉状況と非皮肉状況の分類

状況文コーパスの文を皮肉状況と非皮肉状況に分類するための分類器は BERT を用いて作成した．実行環境には，Google Colabratory[*2]（13GB RAM，NVIDIA Tesla T4 GPU）を用いた．また，BERT の事前学習済モデルには東北大学乾研究室が公開している,「BERT の日本語事前学習モデル」[*3] を用いた．この事前学習モデルに対して，状況文コーパスを用いて皮肉状況と非皮肉状況の文の特徴量を学習させ，ファインチューニングを行うことにより，分類器を作成した．ファインチューニングにおけるドロップアウト率は 0.1，学習係数は 1.0×10^{-5}，バッチサイズは 16，エポック数は 10 とした．学習には，状況文コーパス 5,000 件のうち，4,000 件を用いた．この際，学習データを訓練データ 3,000 件と検証データ 1,000 件にランダムに分けて行った．

6. 精度検証実験

6.1 分類器の精度検証

本章では，皮肉状況の検出に対して精度の評価を行う．評価実験 1 では，BERT により作成した皮肉状況と非皮肉状況を分類する分類器の精度の評価を行う．この実験には，状況文コーパスのうち学習に用いなかった 1,000 件を用いて行う．

BERT を用いて状況文コーパスを学習させることにより作成した分類器に対して，学習データとして使用しなかった，残りの 1,000 件のデータを用いて評価実験を行う．評価指数には，式 1 に示す正解率（Accuracy）を用いた．この式において，TP は真陽性，FP は偽陽性，TN は真陰性，FN は偽陰性を各々示す．

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

皮肉状況と非皮肉状況を分類する分類器の正解率は 0.867 となった．

*2 https://colab.research.google.com/

*3 https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/

6.2 皮肉文に含まれる皮肉状況の検出検証

5. 章で作成した皮肉状況と非皮肉状況を分類する分類器は，文が状況を表す文であるとき，皮肉状況か非皮肉状況かを分類している．しかし，この分類器が分類している文は,「と（い | ゆ | 言）う皮肉」,「と（い | ゆ | 言）う奇跡」であり，皮肉文に含まれる皮肉状況を検出できるか検証する必要がある．評価実験 2 では，皮肉状況文コーパスを用いて皮肉状況と非皮肉状況を分類する分類器が，皮肉文に含まれている皮肉状況を検出可能かの検証を行う．これにより,「と（い | ゆ | 言）う皮肉」が皮肉状況であることに加えて，分類器が皮肉状況の検出に有効であるかを検証する．皮肉状況文コーパスの文を分類器により分類することで，皮肉文に含まれる皮肉状況の検出を行う．評価指数は評価実験 1 と同様に，正解率を用いた．評価実験 2 の結果は，皮肉状況文コーパスの 3,094 件のうち，2,440 件を皮肉状況として検出した．正解率を算出すると，0.788 であった．

7. 考察

7.1 評価実験 1

皮肉状況と非皮肉状況を分類する分類器の精度を評価することを目的とした，評価実験 1 の結果について考察する．この分類器での正解率は 0.867 であり，適切に状況文コーパスを分類できていることが確認できた．これにより，状況文コーパスを用いて作成した分類器は，状況を表す文がある場合，皮肉状況と非皮肉状況を分類することができ，皮肉状況を検出できることが示された．

評価実験 1 で，誤判定されたデータを表 1 に示す．分類器が皮肉状況と誤判定した文を確認すると，表 1 中の例 8 のように，文に書かれている状況だけでは皮肉かどうか判断することができない文が存在した．このような文では，前後に連なる状況を表す文や評価表現により皮肉状況かどうかを判断することができる．例えば，表 1 中の例 7 の状況文に「直らないとされていたのに」という状況や「人間はすごいな」のような評価表現があれば，非皮肉状況となる．反対に,「高額の治療を予約したのに」と言う状況や「医者はやはり必要だ」のような評価表現があれば，皮肉状況となる．このように状況文コーパスだけでは，皮肉状況と非皮肉状況が判断できない文が存在した．このような文をコーパスから除き学習することや，その前後に連なる文を考慮することにより判断できると考える．

また，非皮肉状況として採用した「と（い | ゆ | 言）う奇跡」では，奇跡的な状況を表す文を皮肉状況ではない状況としている．この奇跡的な状況のうち，表 1 中の例 8 のように奇跡的な状況が起きることが，皮肉状況になり,「と（い | ゆ | 言）う奇跡」が皮肉として用いられている文が存在した．分類器としては，誤判定となっているが，この文は状況文コーパスの作成の際にノイズとして含まれた文であり，皮肉状況を表す文である．そのため，学習における改善ではなく，状況文コーパスの作成段階でのノイズの除去の方法を考える必要がある．

分類器が非皮肉状況と誤判定した文を確認すると，表 1 中の例 9 のように，ポジティブな内容が含まれている文が存在した．この文の，前処理により削除した部分には，笑顔の方が綺麗であると思っていたが，真顔の方が綺麗だと言われた事実が皮肉であるという内容の文であった．このように，ポジティブな状況が皮肉状況になる文も存在するため,「真顔」という単語に対する前提知識の考慮や,「の方が」などの文のパターンによる検出を組み込む必要がある．表 1 中の例 10 の文でも，前処理により削除した部分に，自由をやめて，良い子を演じてい

表 1: 評価実験 1 において誤判定された文

皮肉と誤判定	例 7 寝て起きたら症状が回復する
	例 8 合致している情報が何一つとしてない
非皮肉と誤判定	例 9 わたしは真顔の方が綺麗
	例 10 良い子はみんなご褒美がもらえる

表 2: 評価実験 2 において非皮肉状況と誤判定された例

非皮肉と誤判定	例 11 ご近所さんはすでにビニールプールでおおはしゃぎ
	例 12 居眠りしても年収 1000 万越えできる

ることに対して，皮肉な状況であるという内容の文であった．前処理で「と（い | ゆ | 言）う皮肉」以降の文を削除したことにより，必要な文脈を読み取ることができない文も存在することが確認できた.「皮肉」や「奇跡」という手がかり語を削除しつつ，文として成立する前処理を考える必要がある．

7.2 評価実験 2

皮肉状況と非皮肉状況を分類する分類器が，皮肉文に含まれている皮肉状況の検出を検証することを目的とした，評価実験 2 の結果について考察する．この分類器が，皮肉状況文コーパスから皮肉状況を検出した正解率は 0.788 であり，皮肉状況を検出できていることがわかった．

評価実験 2 で誤判定された文を表 2 に示す．分類器が非皮肉状況とご判定した文を確認すると，表 1 中の例 11 のように前後の文から皮肉状況と判断できない文が存在した．この例では「お元気そうでなにより」という皮肉表現が続く文であったが，「ご近所さん」という皮肉対象の「おおはしゃぎ」という状況を皮肉状況と判断することができなかった．人がこの皮肉文から皮肉状況を判断する際，その文から読み取れる背景や前提知識を含めて判断する．例えば，文から読み取れる背景として，「家同士が近い」ということや，前提知識として,「ご近所が大はしゃぎするとうるさい」というようなことが挙げられる．そのため，このような皮肉状況を検出するためには，皮肉状況の前後の文を考慮することや，単語ごとの前提知識を組み込んだ特徴量を算出することが必要であると考える．

表 2 中の例 12 のように状況文コーパスにおける，非皮肉状況と類似している文が存在した．この例では,「居眠り」という否定的な状況に「年収 1000 万超えできる」という肯定的な表現が，奇跡的な状況と類似している文のため，御されたと考える．これは状況文コーパスにおいて皮肉状況と非皮肉状況に共通した特徴量があるため，分類を明確にはできていないことが原因であると考えられる．状況文コーパスにおける非皮肉状況に「と（い | 言）う奇跡」以外の状況文を加えることで，精度の向上が可能であると考える．

8. おわりに

本研究では，皮肉状況に着目し皮肉状況を検出することで，皮肉を検出することを目指した．その端緒として，皮肉状況の検出が可能か検証を行った．まず，Twitter 上の投稿されているツイートを収集し，状況を表す文のうち，皮肉状況と非皮肉状況からなる，状況文コーパスを作成した．次に，BERT を用いてその状況文コーパスを学習し，皮肉状況と非皮肉状況を分類する分類器を作成した．分類器の精度検証として正解率を算出したところ，正解率は 0.867 で，皮肉状況と非皮肉状況の分類が可能であることが示唆された．最後に，分類器が皮肉文に含まれる皮肉状況を検出可能か，検証を行った．結果は，皮肉状況を検出した正解率が 0.788 であり，分類器による皮肉状況の検出，有効であることがわかった．本論文では，状況文から皮肉状況を検出した．今後は，皮肉表現となる肯定表現を考慮し皮肉状況を検出や，皮肉状況から皮肉文を検出する手法について検討する．

参考文献

[Hankyol 20] Hankyol, L., Youngjae, Y., and Gunhee, K.: Augmenting data for sarcasm detection with unlabeled conversation context, *arXiv preprint arXiv:2006.06259* (2020)

[Jacob 18] Jacob, D., Ming-Wei, C., Kenton, L., and Kristina, T.: Bert:Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018)

[伊藤 21] 伊藤 圭吾, 荒澤 孔明, 服部 峻：誹謗中傷による被害を減らすためのツイートにおけるトゲワード検出, 電子情報通信学会技術研究報告, Vol. 120, No. 311, pp. 7–12 (2021)

[諏訪 21] 諏訪 光輔, 張 建偉：BERT 及び絵文字を利用した日本語文における皮肉の検出, 第 13 回データ工学と情報マネジメントに関するフォーラム, H31-4 (2021)

[石坂 11] 石坂 達也, 山本 和英：Web 上の誹謗中傷を表す文の自動検出, 言語処理学会第 17 回年次大会, No. E1-6, pp. 131–134 (2011)

[肥合 18] 肥合 智史, 嶋田 和孝：皮肉検出のための皮肉の対象の推定, 電子情報通信学会技術研究報告, Vol. 118, No. 210, pp. 7–11 (2018)