ソーシャルメディアにおける俗語の検出

青木 竜哉 † 笹野 遼平 ‡ 高村 大也 ‡ 奥村 学 ‡
東京工業大学 工学院 † 東京工業大学 科学技術創成研究院 ‡
†aoki@lr.pi.titech.ac.jp ‡{sasano, takamura, oku}@pi.titech.ac.jp

1 はじめに

Twitter をはじめとしたソーシャルメディアでは，「サーバ」の意味で使用される場合がある「鯖」などのように，一般的な辞書に含まれる語が，辞書に掲載されていない意味で使用されるケースが存在する．文中のある語が辞書に掲載されていない意味で使用されていた場合，多くの人は一般的な用法ではないと判断できるが，その意味を特定するためには，なんらかの事前情報が必要であることが多い．本研究では，このような性質を持つ語の解析の手始めとして，辞書に掲載されていない意味で使用される俗語を対象に，SNS 中に出現した文における語が俗語として使われているかどうかを分類する手法を提案し，提案した分類手法を応用して俗語の自動検出を行う．

本研究では，まず，対象とする俗語がアノテートされたデータセットを作成する．作成するデータセットでは，コンピュータ，企業・サービス名，ネットスラングのカテゴリに出現した語から，俗語として使用されることのある 38 語を選定した上で，Twitter 上で各単語が出現する文をそれぞれ 100 文抽出し，各語が俗語として使われているかのアノテーションを行った．

続いて，各語が俗語的用法で使用されているかどうかの自動分類に取り組む．分類のアプローチとしては，俗語的用法で使用されている語は，その語の一般的用法で使用されている場合と周辺文脈が異なるとの仮定に基づき，テストデータとは別のコーパスを用いて事前に単語の使われ方を学習し，学習した単語の使われ方と，テストデータにおける使われ方が異なる場合に俗語と分類する手法を提案する．さらに，提案手法を応用して，俗語の自動検出を行った．

2 ソーシャルメディアと俗語

Twitter に代表されるソーシャルメディアにおいては，辞書に掲載されていない意味で使用されている語がしばしば出現する．例として，表 1 に Twitter における鯖の使われ方を示す．表 1 の上段と下段の例においては，同じ鯖という単語が出現しているが，その意味は異なり，上段における鯖 $_1$ は，青魚に分類される魚の鯖を示しているが，下段における鯖 $_2$ は，コンピュータサーバのことを意味している．本研究では，鯖 $_2$ のような俗語を検出することを目的とする．

表 1: Twitter における**鯖**の使われ方

今日，久々に**鯖** $_1$ の塩焼き食べたよ とても美味しかった 肉，魚，野菜など何でも好きだよ
なんで，急に**鯖** $_2$ 落ちしてるのかと思ったら スマップだったのか (ˆ ｑ ˆ)

本研究と関連している自然言語処理の研究分野として，新語義検出 [7] や新語義の用例のクラスタリング [2] が挙げられる．しかし，本研究で扱う俗語は，特定の状況において語義が変化するという性質を持つため，これらの一般的な語義に着目した研究とは枠組みが異なる．また，Bamman ら [1] は，話者の地域によって，語の持つ意味が異なるという点に着目し，状況に応じた語の意味表現ベクトルを獲得する手法を提案している．本研究も同様に，同一語の用法の違いに着目しているが，Bamman らが地域ごとの語の使われ方の違いに着目しているのに対し，本研究では語の使われ方が辞書的であるかそうでないかに着目する．

Web 上で使用される俗語に関する研究もいくつか存在する．Sboev[6] は，インターネットにおいてのみ使われる中国語の俗語表現の分析を行った．山田ら [8] は，有害情報を表す隠語に焦点を当てて，隠語を概念化するフレームワークを提案し，隠語表現の分類を行った．山田らは，隠語の知識を含んだ辞書を作成し，分類タスクを解いたが，作成した辞書のみでは隠語表現の多様性の対応に不十分であったと報告している．本研究は，俗語の検出を行うことに主眼をおいているため，表現の分析に重きをおいている Sboev の研究とは目的が異なる．山田らが有害情報を表す隠語に着目しているのに対して，本研究では，俗語という，より広いクラスを対象としている．また，山田らがドメインに特化した知識を用いているのに対し，本研究で提案する手法ではそのような知識を必要としない．

3 データセット

本研究で対象とする俗語に関するデータセットは存在しないため，まず，データセットの構築を行った．データのソースとして Twitter を採用し，2016 年 1 月 1 日から 2016 年 1 月 31 日に投稿されたツイートを対象としてデータセットを作成した．Twitter をデータのソースとして選択した理由は，Twitter においては，ある語が俗語として使われる場合と辞書的な意味で使われる場合が混在していると考えたためである．

表 2: 作成したデータセットの概要

カテゴリ	辞書的用法	俗語的用法	合計
企業・サービス名	324	227	551
コンピュータ	416	234	650
ネットスラング	817	814	1,631
合計	1,557	1,275	2,832

データセットの作成にあたって，まず，本研究で分類の対象とする 38 語を選定した．この 38 語は，一般的な辞書の見出しに含まれるものの，辞書に掲載されていない意味で使用される場合がある語であり，企業・サービス名，コンピュータ，ネットスラングの 3 つのカテゴリから人手で選定した．

次に，ツイートに対して形態素解析を行い，あらかじめ選定した 38 語が一般名詞であると解析されたツイートを無作為に 100 ツイート選択した．形態素解析には，MeCab[1] を使用し，IPA 辞書を用いて解析した．

最後に，選択したツイートにおいて，選定した単語が辞書的な意味で用いられているか，固有表現の一部となっているか，俗語として用いられているかという判断を 2 人のアノテータによって人手で行った．固有表現の一部となっている事例というのは，例えば「利尻島」の中の「尻」のような事例を示す．また，いずれかのアノテータが，与えられた情報だけからはツイート中で使用されている対象語の意味を決定できないと判断したツイートは，データセットから除外した．アノテーションが一致したツイートのうち，2 人のアノテータが辞書的な意味で用いられていると判断した事例と，俗語として用いられていると判断した事例の集合を最終的なデータセットとした[2]．アノテーションの一致率は 88.6% ($\kappa = 0.816$) であった．表 2 にデータセットの概要を示す．

作成したデータセットでは，単語ごとにラベルの偏りが見られた．アノテーションの結果に基づいて，38 語を，辞書的な用法の事例が多い語，俗語として用いられている事例が多い語，それ以外の語の 3 つに分類した．具体的には，7 割以上が辞書的用法としてアノテーションされた単語を辞書的ラベル優勢，7 割以上が俗語的用法としてアノテーションされた単語を俗語ラベル優勢，それ以外をラベル偏りなしとしてクラス分けを行った．表 3 に，アノテーション対象とした語の一覧をクラスごとに示す．

表 3: アノテーション対象とした 3 8 語の内訳

クラス	俗語				
俗語ラベル優勢	囲い	円盤	乙	垢	鯖
	支部	沼	草	裏山	密林
	地雷	凸	尼		
辞書的ラベル優勢	ピザ	安価	芋	駅弁	庭
	空気	串	狐	惨事	板
	尻	窓	蔵	林檎	洒落
	茸	鶴	渋		
ラベル偏りなし	藁	泥	鉄板	虹	養分
	升	ゆとり			

4 俗語の自動分類

4.1 提案手法

本研究では，作成したデータセットとは別のコーパスを用いた教師なし学習による分類手法を提案する．具体的には，まず，作成したデータセットとは別のコーパスで，単語の分散表現を学習し，そのコーパスにおける単語の使われ方を学習する．続いて，学習した単語の分散表現を用いて，学習コーパスにおける語の使われ方とテストデータにおける語の使われ方との類似度を計算し，計算した類似度が閾値を下回った場合，俗語と分類する．この類似度には，着目単語の分散表現と，周辺単語から予測される単語の分散表現の類似度を使用する．周辺単語は，着目単語に対して，前後の窓幅分の単語とする．窓幅は，分散表現の学習時に用いた値と同じ値を使用する．ただし，窓幅内に出現した未知語は類似度計算時に考慮しないこととする．図 1 に提案手法の概要を示す．

提案手法では，Skip-gram with Negative Sampling (SGNS)[4] によって単語の分散表現[3] を学習し，学習された着目単語の入力側のベクトル表現 v と，周辺単語の出力側のベクトル表現 v' の類似度の加重平均に基づいて，俗語を分類する．類似度として，加重平均を採用している理由は，本研究で使用した SGNS の学習過程では，着目している単語と距離の近い単語に重みを付けた学習が行われているためである．窓幅を m，着目単語との距離を d と定義すると，加重平均の重み α は，$\alpha = m + 1 - d$ によって計算される整数値とする．ただし，d は，着目単語を基準として，何単語離れているかを表す整数値である．文中の着目単語を w_t とし，着目単語の入力側のベクトル表現を v_{w_t}，w_t を基準として前後 m 単語の単語集合を $\mathbf{w_c}$，各周辺単語 $w_j \in \mathbf{w_c}$ の出力側のベクトル表現を v'_{w_j}，その重みを α_{w_j} と表すと，提案手法では，次の式によって，着目単語 w_t が俗語として使われているか，そうではないかを分類する．

$$score = \frac{\sum_{w_j \in \mathbf{w_c}} sim(v_{w_t},\ v'_{w_j}) \times \alpha_{w_j}}{\sum_{w_j \in \mathbf{w_c}} \alpha_{w_j}}, \tag{1}$$

$$usage = \begin{cases} \text{俗語的用法} & (score < \text{th}) \\ \text{辞書的用法} & (\text{otherwise}) \end{cases}. \tag{2}$$

式 (1) における $score$ は，事前に学習した単語 w_t の使われ方と，文中に出現する単語 w_t の使われ方の類似度を表し，計算された $score$ の値に対して閾値 th との大小関係によって単語 w_t の用法を分類する．$sim(v_{w_t},\ v'_{w_k})$ は，着目単語のベクトル表現 v_{w_t} とその周辺単語のベクトル表現 v'_{w_k} の類似度を表し，提案手法では，$sim(v_{w_t},\ v'_{w_k}) = sigmoid({v_{w_t}}^{\top} v'_{w_k})$ を使用する．

[1] http://taku910.github.io/mecab/

[2] 固有表現の一部となっている事例を除いたのは，将来的に固有表現認識を行うことにより機械的に除外できると考えたためである．

[3] Mikolov ら [4] における，単語の入力側のベクトル表現と出力側のベクトル表現をそれぞれ v，v' と表す．

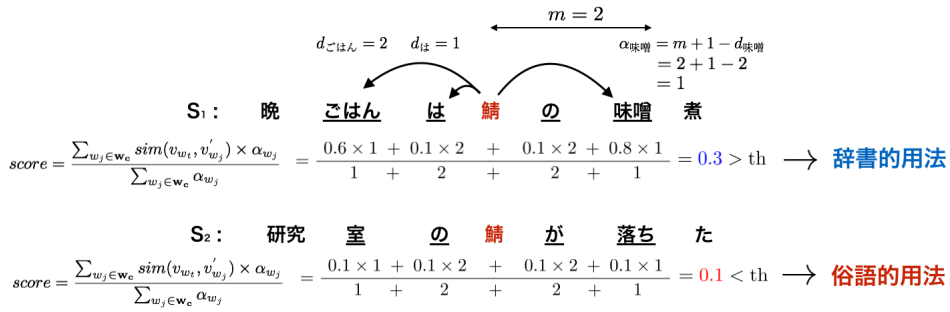

図 1: 提案手法の概要

表 4: 学習に使用したコーパスの内容

コーパス	内容	異なり語数	単語数
BCCWJ	現代日本語書き言葉均衡コーパス	131,913	1.1 億
Web	Web から無作為に抽出した文の集合	336,048	6.0 億
Wikipedia	2016 年 7 月時点の日本語 Wikipedia	1,081,154	8.9 億
新聞	1994-2004 年の新聞 [4]	1,204,914	15.0 億

4.2 実験

本研究では，4 種類のコーパスで学習した単語の分散表現を用いて俗語の自動分類を行った．特に，提案手法では，学習コーパスの用法との差異に基づいて俗語分類を行うため，均衡コーパスである BCCWJ における用法を学習することで，より俗語を分類することができると考える．表 4 に，各コーパスの内容を示す．

単語の分散表現の学習にあたっては，窓幅を 5，次元数を 300 とした．また，各コーパスにおいて，5 回未満の出現頻度の単語は，<unk> に置き換えて学習を行った．SGNS による単語の分散表現の学習には，Python ライブラリの一つである `gensim` [5] による実装を使用した．SGNS の学習時には，ネガティブサンプリングの数を 10 とした．閾値は，テストデータを単語レベルで無作為に二つのグループに分け，分割したグループを用いた二分割交差検定によって決定した．なお，閾値の範囲は，0.00 から 1.00 で 0.01 刻みで探索し，それぞれのグループで分類正解率が最も高くなる値をもう一方のグループの閾値として利用した．

分類の評価には，正解率を利用した．また，比較手法として，相互情報量を用いて学習した単語の分散表現に対して，特異値分解 (SVD) によって次元削減を行った単語の分散表現を使用した手法による実験も行った．相互情報量には，Positive Pointwise Mutual Information (PPMI) を使用する．この手法を用いる場合には，式 (1) において，$sim(v_{w_t}, v'_{w_k}) = \frac{{v_{w_t}}^\top v_{w_k}}{||v_{w_t}|| \times ||v_{w_k}||}$ を使用する [5]．PPMI と SVD による単語の分散表現の学習には，Levy ら [3] の実装を使用した．さらに，それぞれの手法において，式 (1) における α を $\alpha = 1$ に固定し，距離による重みを考慮しないモデルによる実験も行った．表 5 に，これらの実験の結果を示す．

[4]毎日新聞，日経新聞，読売新聞を対象とした．

[5]SVD による次元削減後の行列を $M = W \cdot \sum \cdot C^\top$ とすると，$v = W \cdot \sum$ ($\cdot$ は内積を表す)．単語の共起情報の類似度を図るため，$v' = v$ とする．

表 5: 各コーパスと手法における正解率

コーパス	提案手法 重みあり	提案手法 重みなし	SVD 重みあり	SVD 重みなし
BCCWJ	**.811**	**.804**	**.722**	**.704**
Web	.784	.775	.691	.687
Wikipedia	.730	.727	.653	.658
新聞	.734	.720	.689	.689

表 5 より，SVD を用いた手法と比較して，提案手法の方が正解率が高い傾向にあることが確認できる．SGNS によって学習した分散表現を用いることで，共起情報だけではなく，距離も考慮した単語の使われ方を学習することができたため，単語の辞書的用法と俗語的用法の違いを捉えることができたと考えられる．提案手法では，重みを考慮しないモデルと比較して，重みを考慮したモデルの方が正解率が高かった．このことから，俗語の分類においては，ある単語が俗語として扱われた場合は，その単語と特に距離の近い周辺単語の使われ方が，一般的な用法として扱われた場合の周辺単語の使われ方と異なることが多いという俗語の性質を示唆する結果となった．

また，それぞれの手法において学習コーパスとして BCCWJ を用いたモデルが最も正解率が高い傾向にあることが確認できる．BCCWJ は，コーパスの規模は小さいものの，コーパスの内容が均衡しているという特徴がある．この特徴によって，単語の一般的な使われ方を学習することができたため，俗語の分類タスクにおける正解率が高くなったと考えられる．

5 俗語の自動検出実験

式 (1) で計算される $score$ が十分に小さい場合，その用例は俗語として使用されている可能性が高いと考えられる．そこで前節で説明した手法を俗語の自動検出に応用することを考える．具体的には，俗語的用法が存在するかどうか未知の語を対象に，式 (1) で計算される $score$ が十分に小さい閾値 th_{small} 以下となった用例を俗語の候補として出力することを考える．

この際，閾値 th_{small} は，分散表現の学習に使用したコーパスにおける対象語の式 (1) によって計算される $score$ の分布から決定する．具体的には，学習コーパスにおける対象語の $score$ を昇順に並べ，先頭から r%を超えた最初の $score$ を閾値として使用する．本稿における実験では，r の値として 0%から 0.1%刻みで順に大きくしていった値を使用した．

まず，予備調査として，3 節で説明したデータセットを用いた実験を行った．このデータセットの作成時に選定した 38 語のうち，データセット中で俗語的用法で使用された事例が存在しなかった「庭, ピザ, 板, 洒落」の 4 語を除く 34 語を対象に，r を変化させた場合に 1 つ以上の俗語的用法が検出された語数（俗語検出数）を調査した．図 2 に結果を示す．縦軸は全 34 語を対象に俗語的用法と判定された用例の適合率を示している．学習コーパスごとに，もっとも左側の点は r=0.0%とした場合[6] の結果を表しており，以降の点は r を順に 0.1%，0.2%と大きくしていった際に俗語検出数に変化があった r に対する検出数と適合率を表す．

図 2 に示す結果から，俗語的用法が存在する語に限定した場合は，高い適合率のもとで多くの語に対し 1 つ以上の俗語的用法を検出可能であることが分かる．特に BCCWJ コーパスを使用し，r=2.3%とした場合，91.7%の適合率のもとで 30 語を俗語的用法を持つ語として検出できた．また，いずれのコーパスを用いた場合も，85%以上の適合率を維持しつつ，31 語を俗語的用法を持つ語として検出することができた．ここで，俗語的用法が 1 つも検出されなかった 3 語は，BCCWJ，Web，Wikipedia，新聞の各コーパスに対しそれぞれ「林檎, 鶴, 惨事」,「惨事, 串, 狐」,「鶴, 狐, 尻」,「尻, 惨事, 狐」の 3 語であり，いずれも表 3 に示すとおり辞書的ラベルが優勢な語であった．

続いて，より現実的な設定として，俗語的用法が存在するかどうか未知の語を対象に，俗語的用法を持つ語の候補を出力することを考える．俗語的用法と判定された用例の割合が多い語は，俗語として使用される可能性が高い語であるとの仮説に基づき，Twitter から抽出した 15,583,427 ツイート中に 1,000 回以上出現した一般名詞を対象として，俗語的用法と判定された用例の割合が多い語 50 語のうち，実際に俗語的用法として使われている事例が存在するかどうかを人手で判定した．この際，予備調査の結果から，コーパスとしては BCCWJ，r の値には 2.3%を使用した．人手判定の結果，50 語のうち俗語的用法が正しく抽出されたと判定された語は 10 語であった．出力された語の一部をその判定結果とともに表 6 に示す．「家宝」や「商業」などのように俗語的用法を持つ語を正しく出力できている一方で，「信長協奏曲」のような形で出現することが多い「協奏曲」などのように，固有名詞中の語が俗語的用法と判定されたために出力された語が散見された．この問題は，固有表現認識技術を活用することで軽減できると考えられる．

6 おわりに

本研究では，俗語の解析の手始めとして，事前に選定した俗語的用法を持つ 38 語を対象に，与えられた文における用法が俗語的用法であるかのアノテーションを行った．続いて，別のコーパスにおける対象語の使われ方との差異に基づき，各語が俗語的用法で使用されているかどうか自動分類する手法を提案し，構築したデータセットにおける実験の結果，81.1%の精度を達成した．さらに，提案した自動分類手法が俗語の自動検出に応用できることを示した．今後の課題としては，固有表現認識技術等を活用した俗語の自動検出精度の向上が挙げられる．

[6] 先頭の事例の *score* を閾値として使用することに相当．

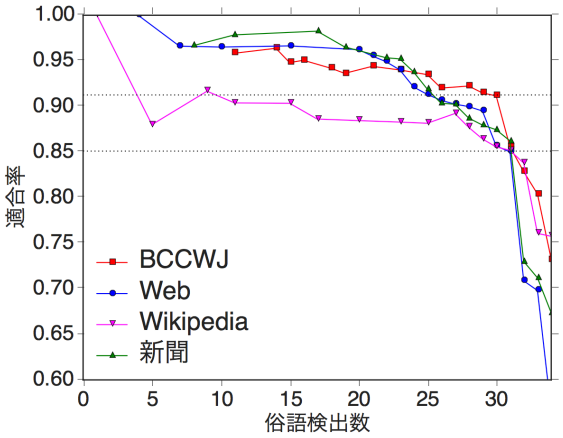

図 2: 適合率 – 検出できた俗語の数

表 6: 俗語的用法ありと判定された語と用法の例

人手判定	語	検出された用例
俗語的用法あり	家宝	日曜**家宝**行けそうで嬉しい！
	商業	衝動的に**商業**買った
	レート	友達と一緒に**レート**潜ってたら
	ペダル	いいねぇ…**ペダル**は夏が似合うし
俗語的用法なし	協奏曲	楽しみにしてた信長**協奏曲**速攻みてきた
	単位	勉強しないと**単位**が死ぬ気がする
	耐性	煽り**耐性**ねーなーって思う

参考文献

[1] David Bamman, Chris Dyer, and Noah A. Smith. Distributed representations of geographically situated language. In *Proceedings of ACL '14*, pp. 828–834, 2014.

[2] Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. Word sense induction for novel sense detection. In *Proceedings of EACL '12*, pp. 591–601, 2012.

[3] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 211–225, 2015.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS '13*, pp. 3111–3119, 2013.

[5] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of Workshop on NewChallenges for NLP Frameworks*, pp. 45–50, 2010.

[6] Aleksandr Sboev. The sources of new words and expressions in the Chinese internet language and the ways by which they enter the internet language. In *Proceedings of PACLIC '16*, pp. 355–361, 2016.

[7] 新納浩幸, 佐々木稔. 外れ値検出手法を利用した新語義の検出. 自然言語処理, Vol. 19, pp. 303–3027, 2012.

[8] 山田大, 安彦智史, 長谷川大, Michal Ptaszynski, 中村健二, 佐久田博司. ID 交換掲示板における書き込み有害性評価に向けた隠語概念化手法の提案. 言語処理学会 第 22 回年次大会 発表論文集, pp. 49–50, 2016.