2N1-1

Twitterデータを用いたユーモア語句自動生成手法に関する一検討

A Study on Humor Generation Method Using Twitter Data

岩倉亮介 *1 Ryosuke Iwakura　吉川大弘 *1 Tomohiro Yoshikawa　ジメネスフェリックス *1 Felix Jimenez　古橋武 *1 Takeshi Furuhashi

*1名古屋大学工学研究科
Graduate School of Engineering Nagoya University

In recent years, researches on automatic dialogue systems (Chat dialogue systems) attract attention. Dialogue systems are classified into task-oriented and non-task-oriented one. In non-task-oriented dialogue systems such as a chat, it is important that a user wants to continue the dialogue over a long period of time. In order to improve dialogue continuity, utterances including humor are effective. However, there is a problem that the conventional methods often generate utterances that are not accepted as humor. In this paper, using Twitter data, we investigate the improvement of humor acceptability in dialogue systems.

1. はじめに

人間とロボット/コンピュータが自然言語を用いて会話を行う，自動対話システムの研究が近年注目されている．自動対話システムの代表例としては，Apple 社の「Siri」*1 や NTT Docomo の「しゃべってコンシェル」*2 などが挙げられる．これら対話システムは大別して，タスク指向型と非タスク指向型に分類される．

タスク指向型対話システムは，対話を通じてユーザの質問や要求に対して適切な情報を提供するなど，特定のタスクの達成を目的とする．一方，非タスク指向型対話システムは，特定のタスクの達成を目的とせず，自由な対話である雑談によってユーザを楽しませることを目的とする．このような非タスク指向型対話システムにおいては，対話の継続性が重要とされるが，対話が単調だと飽きやすいという問題点がある．

これまで，ユーザが継続的に対話を続けたいと感じるためには,「ユーモア表現」を含む発話 [1] と「ユーザの趣味・嗜好」を考慮した発話 [2] が有効なことが示されている．

ユーモア表現を含む発話の生成としては，ユーザからの入力に含まれる名詞に着目し，Twitter から収集したデータに対して word2vec[3] を用いて単語間類似度を考慮する手法が提案されている [4]．ここで用いられている word2vec とは，大規模コーパスから教師なしで単語の分散表現を学習する手法であり，学習された分散表現から cos 類似度を用いて単語同士の意味的な関係を捉えることができる．また，ユーザの発話履歴から，ユーザの発話極性がポジティブ/ニュートラル/ネガティブのいずれかを判断し，ユーモアの生成に考慮する手法も提案されている [5]．また，ユーザの趣味・嗜好を考慮した発話としては，小林らの手法が提案されている [2]．[2] では，ユーザの嗜好情報や知人関係を記憶し，システムの発話に反映することで，対話の継続性を高めている．

しかしこれまで,「ユーモア表現」と「ユーザの趣味・嗜好」の両者を同時に考慮した研究は報告されていない．そこで本稿では，ユーモア表現の生成に，趣味・嗜好を考慮する手法を提案する．具体的には藤倉ら [4] の手法をベースとし，ユーザの趣味や嗜好を取り入れることを目指す．また，提案手法で生成されたユーモアを用いて被験者実験を行い，その性能を評価する．加えて，共起する形容詞を用いて実験結果の解析を行う．

連絡先: 岩倉亮介，名古屋大学大学院工学研究科，名古屋市千種区不老町，052-789-2793，052-789-3166，iwakura@cmplx.cse.nagoya-u.ac.jp

∗1 http://www.apple.com/jp/ios/siri/

∗2 https://www.nttdocomo.co.jp/service/shabette_concier/

2. 従来手法

提案手法のベースとして用いる従来手法 [4] のユーモア生成の手順を以下に示す．なお，対話システムにおいてユーザから単一の文が入力された状況を想定し，この入力文に含まれる名詞を入力名詞とする．

1. Twitter からツイートデータを取得する．
2. 1. で得られたデータに対し，形態素解析器 MeCab[6] を用いて「連体修飾要素 + 名詞」の組み合わせを抽出し，データベースを構築する．なお，修飾要素は修飾語または修飾部を意味している．
3. word2vec により学習した分散表現を用いて，入力名詞とデータベース中にある名詞との cos 類似度を計算し，類似度が高い/低い/ランダムな名詞を選択する．
4. 3. で選択した名詞と組み合わせになっている修飾要素をデータベース中から全て取り出し，その中で登場回数が少ない修飾要素の組み合わせを選択する．
5. 名詞を入力名詞と置換し,「ですか？」を付与して疑問文の形で出力とする．

3. 提案手法

3.1 概要

システムはユーザの趣味や嗜好を考慮した上で，ユーザからの入力文に含まれる名詞を使用したユーモアを出力として生成する．従来手法と同様，ユーモアを生成する際に，word2vec で学習された分散表現により入力名詞との cos 類似度を求める．入力名詞に対して似た意味の単語や異なる意味の単語を用いることで，ユーモア受容性の向上を期待する．ユーモア受容性とは，対話システムの応答に対するユーモアの感じやすさを意味する [4].

3.2 ユーモア生成方法

提案手法として，2 章で示した従来手法の手順 1 と 2 の間に趣味・嗜好を考慮する手順を加えた流れを以下に示す．なお，単純化のため，従来手法の手順 2 で「修飾要素 + 名詞」であった部分を「名詞 $_1$ + の + 名詞 $_2$」に変更する．

1. Twitter からツイートデータを取得する．
2. 取得したツイートデータからユーザの趣味に関する単語を含む文を抽出する．
3. 2. で得られた文に対し，形態素解析器 MeCab[6] を用いて「名詞 $_1$ + の + 名詞 $_2$」の組み合わせを抽出し，データベースを構築する．
4. word2vec により学習した分散表現を用いて，入力名詞とデータベース中にある名詞 $_2$ との cos 類似度を計算し，類似度が高い/低い/ランダムな名詞 $_2$ を選択する．
5. 手順 4 で選択した名詞 $_2$ と組み合わせになっている名詞 $_1$ をデータベース中から全て取り出し，その中で登場回数が少ない名詞 $_1$ の組み合わせを選択する．
6. 名詞 $_2$ を入力名詞と置換し，「ですか？」を付与して疑問文の形で出力とする．

ユーザの趣味を「サッカー」とし，入力文に「ぬいぐるみ」が含まれる場合のユーモア生成例を図 1 に示す．

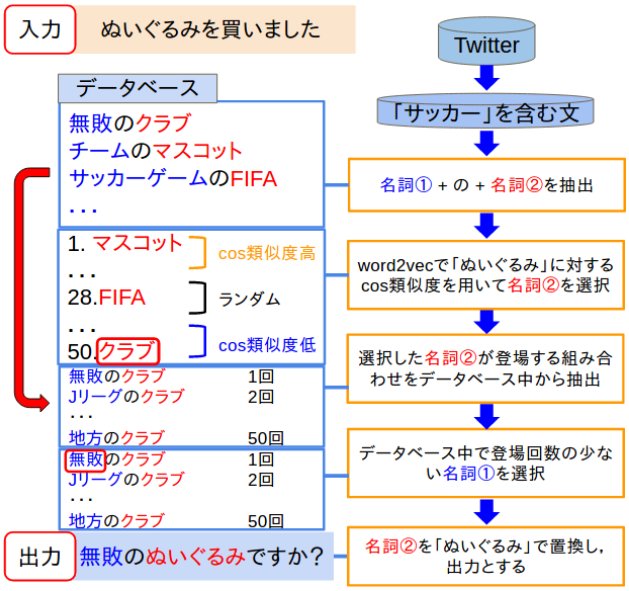


図 1: ユーモア生成フロー

4. 実験方法

ユーザの趣味・嗜好とユーモア受容性との関係を調査するため，趣味としてスポーツを想定し，3.2 節で説明したユーモア生成方法を用いて評価実験を行った．

データには，Twitter から 2016/10/03〜2016/11/17 の間に取得した約 2000 万ツイートを利用した．スポーツとしては，[7] で紹介されている人気スポーツ上位 5 種類の「野球」,「サッカー」,「テニス」,「ウォーキング」,「バレーボール」を選択した．

各スポーツについて，入力名詞に対する単語間類似度の上位 15 個，下位 15 個，その他ランダムに取得した 20 個の計 50 個ずつ，スポーツ 5 種類で計 250 個のユーモア語句を生成した．また比較として，「名詞 + の + 名詞」の形式で従来手法においても同様に 50 個を生成し，計 300 個のユーモア語句に対して面白いか否かの評価を行った．ここでは，入力名詞には，「ぬいぐるみ」,「結婚」,「メガネ」の 3 つを用い，大学生 3 人がそれぞれユーモアの評価を行った．なお今回の実験では，ユーモア発話生成に対する基礎的な検討のため，出力には「ですか？」を付与せずに行った．また，各評価者の趣味・嗜好とユーモア受容性との関係を調査するため，各スポーツに対する興味の度合いを 1〜5 の 5 段階で評価してもらった．

5. 結果と考察

表 1 に面白いと判断されたユーモア生成例，図 2 に評価実験の結果，表 2 に各スポーツに対する各評価者の興味度をそれぞれ示す．また，図 3 に面白いとされたユーモアの数と興味度の関係を表す散布図，表 3 にその相関係数をそれぞれ示す．

図 2 と表 2 から，例えば評価者 1 のテニスへの興味度は 4 で，興味度がそれぞれ 3 と 1 のサッカーや野球よりも面白いと評価した数は上回っているが，興味度が 2 のウォーキングの場合よりは下回っているなど，必ずしも興味度と面白さとの相関は高くないことがわかる．また全評価者において，従来手法で生成したユーモアを面白いとする数が最大または 2 番目となっている．さらに表 3 に示した相関係数で無相関検定（有意水準：5 %）を行った結果，有意な相関であるとはいえなかった（p=0.57）．

以上の結果から，スポーツへの興味度と面白いと判断する数には意味のある相関がないという結果となった．さらに，趣味・嗜好を考慮することによるユーモア受容性の向上は確認できなかった．しかし，図 2 において，従来手法で生成されたユーモアでさえ面白いと判断された数は非常に少ないため，ユーモア生成方法として改善の余地があると考えられる．

趣味・嗜好としてスポーツを考慮してもユーモア受容性が向上しなかった原因として，スポーツが評価者の趣味として十分ではなかったためと考えられる．このことは，表 2 において興味度の最大値である 5 を選択している評価者がいないことからわかる．また，各スポーツ名で文を限定したことにより，ユーモアの候補となる組み合わせの数が減少したためであること，さらにスポーツの考慮の方法として，直接スポーツ名を含む文を対象としたためなどが考えられる．さらに表 1 において，各スポーツを考慮しながらも，出力されたユーモア文にはそのスポーツと関係ない単語が結果的に選ばれてしまっていることは，今後検討・改善していく必要があると考えられる．

6. ユーモアの要因の解析

5 章ではスポーツへの興味度とユーモア受容性との関連について論じたが，本節ではユーモアと感じる要因そのものについて考察する．評価実験で得られた結果について解析を行うため，面白いと判断されたものについて，ユーモアの要因となるような特徴が見られるかについて調査した．

今回の実験で用いた「名詞 + の + 名詞」のような単純な形においては，二つの名詞の関連性が面白いと判断される要因になると想定される．例として，表 1 で示したユーモア生成例のうち「ヤクザのぬいぐるみ」と「久しぶりの結婚」を考える．これらはそれぞれ，“ぬいぐるみは可愛いもの”，“結婚は

表 1: 面白いと判断されたユーモア生成例

スポーツ名	面白いと判断されたユーモア生成例
（従来手法）	ヤクザのぬいぐるみですか？ 離婚前提の結婚ですか？
野球	遂行中のぬいぐるみですか？ 久しぶりの結婚ですか？
サッカー	債務超過のぬいぐるみですか？ デブ合宿の結婚ですか？
テニス	毎日仕事のぬいぐるみですか？ さだまさし好きの結婚ですか？
ウォーキング	悲劇的運命のぬいぐるみですか？ 20 分程度の結婚ですか？
バレーボール	墓だらけのぬいぐるみですか？ 一人の結婚ですか？

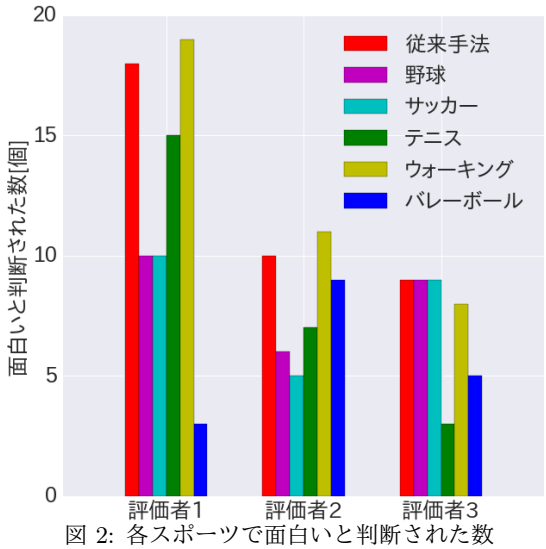


図 2: 各スポーツで面白いと判断された数

表 2: 各スポーツに対する評価者の興味度

	バレーボール	テニス	サッカー	野球	ウォーキング
評価者 1	2	4	3	1	2
評価者 2	2	1	1	2	2
評価者 3	4	2	3	4	4

表 3: 面白いとされた数と興味度との相関係数

	相関係数
評価者 1	0.26
評価者 2	0.61
評価者 3	0.56
全体	0.16

頻繁にはしないもの”というような一般常識からズレていたことが，面白いとされた要因としてあると考えられる．

しかし，一般常識を読み取るためには，名詞のニュアンスを理解する必要がある．そこで，名詞のニュアンスを表すものとして共起する形容詞を考える．形容詞は名詞を修飾する役割をもつため，文章として名詞のニュアンスを表している可能性があると考えられる．

そこでここでは，名詞に対してその名詞と共起する形容詞

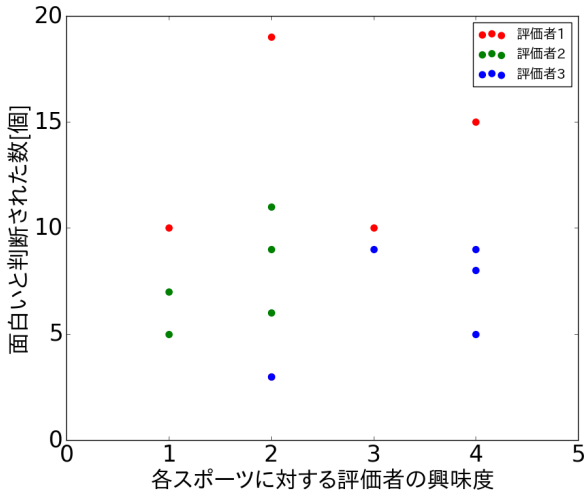


図 3: 面白いとされたユーモアの数と興味度の関係

との関係を調査した．Twitter から取得したデータにおいて，出現する形容詞をカウントし，文章内にある名詞と形容詞が共起した回数を算出した．なお今回は，各名詞に対して比較するため，共起としてカウントするのは Twitter データ中で出現回数が上位 100 個の形容詞とした．

以下に結果を示す．表 4 に，Twitter データ中で出現回数が上位 10 個の形容詞を示す．また図 4 に，「ぬいぐるみ」，「ヤクザ」に対する形容詞の共起回数をそれぞれ示す．図 4 において，横軸は左から Twitter での出現回数が多い順に 100 個の形容詞に対応している．出現回数順であるため，グラフは基本的に右肩下がりとなるが，それぞれの図において共起回数が突出した部分があることが確認できる．また，表 5 に，それぞれの名詞で共起回数の多い形容詞上位 3 個を示す．表 5 から，「ヤクザ」と「ぬいぐるみ」では「可愛い」と「怖い」といった対義語に近い形容詞がそれぞれ共起しやすいことがわかる．そのため，単語のニュアンスとして反対の意味が現れていると考えられる．このように，ユーモアと感じる要因は，共起する形容詞の頻度分布の違いで解析できる可能性がある．

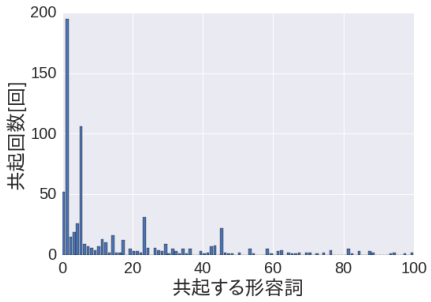
表 4: Twitter における形容詞の出現回数上位 10

1 位	2 位	3 位	4 位	5 位
いい	可愛い	楽しい	すごい	嬉しい
6 位	7 位	8 位	9 位	10 位
欲しい	やばい	早い	多い	強い

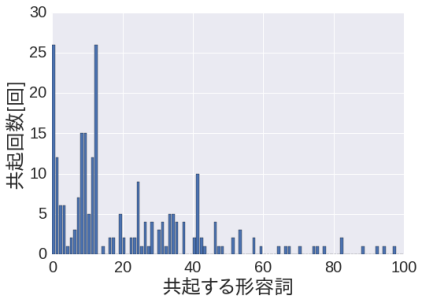
表 5: 名詞と共起する形容詞の上位 3

	1 位	2 位	3 位
ぬいぐるみ	可愛い	欲しい	いい
ヤクザ	いい	多い	怖い

また，形容詞との共起で類似した意味の名詞が捉えられるかを確認するため，100 種類の形容詞の共起回数を要素としたベクトルを用いて名詞間の cos 類似度を求めた．表 6 に，「ぬいぐるみ」，「ヤクザ」，「結婚」，「名古屋」のそれぞれに対する cos 類似度が上位 5 個の名詞を示す．表 6 から，「ぬいぐるみ」，「名古屋」では比較的類似した意味の名詞が取得できてい



(a) ぬいぐるみ



(b) ヤクザ

図 4: 形容詞 100 種類の共起回数

ることが確認できる．一方,「ヤクザ」では特徴を表した名詞が取得できている．この結果から，共起する形容詞を用いることで，類似した意味の名詞や特徴を表した名詞が取得できると考えられる．cos 類似度上位のものだけでなく，cos 類似度下位の名詞や cos 類似度が特定の値となる名詞を用いることで，ユーモア受容性の高いユーモア発話の生成が行える可能性があると期待できる．

表 6: 各名詞に対する cos 類似度上位 5

	ぬいぐるみ	ヤクザ	名古屋
1 位	ポーチ	詐欺	大阪
2 位	イヤリング	職員	地元
3 位	ワンピ	キレる	会場
4 位	チャーム	切断	福岡
5 位	キーホルダー	恐怖	京都

7. まとめ

対話の継続性向上を目的として，ユーモア発話の生成に趣味・嗜好を考慮する手法を提案した．提案手法で生成したユーモアについて評価実験を行った結果，従来手法と比較して，趣味・嗜好を考慮したことによるユーモア受容性の向上は見られなかった．また，実験に用いたスポーツへの興味度とユーモア受容性との間にはほぼ相関がないという結果となった．一方で，実験結果について形容詞を用いた解析を行った結果，ユーモアと感じる要因の一つとして，共起する形容詞の頻度分布の違いが関連している可能性があること，また，形容詞を用いた関連語の取得が可能であることを示した．

今後の課題としては，ユーモア受容性の向上を目指して，趣味・嗜好の考慮方法に対する検討と，形容詞を用いたユーモアの分析およびユーモア発話の生成方法に対する検討が挙げられる．

謝辞

本研究は，文部科学省科学研究費 (基盤研究 (B), No.16H02889) の助成を受けたものです．

This work was supported by MEXT KAKENHI (Grant-in-Aid for Scientific Research (B), No.16H02889).

参考文献

[1] 宮澤幸希, et al. ”音声対話システムにおける継続欲求の高いインタラクションの要因.” 電子情報通信学会論文誌 A 95.1 (2012): 27-36.

[2] 小林峻也, and 萩原将文. ”ユーザの嗜好や人間関係を考慮する非タスク指向型対話システム.” 人工知能学会論文誌 31.1 (2016): DSF-A_1.

[3] Mikolov, Tomas, et al. ”Efficient estimation of word representations in vector space.” arXiv preprint arXiv:1301.3781 (2013).

[4] 藤倉将平, 小川義人, and 菊池英明. ”非タスク指向対話システムにおけるユーモア応答生成手法.” 人工知能学会全国大会論文集 29 (2015): 1-4.

[5] 松井辰哉, and 萩原将文. ”発話極性を考慮したユーモアを有する非タスク指向型対話システム.” 日本感性工学会論文誌 14.1 (2015): 9-16.

[6] MeCab. ”Yet Another Part-of-Speech and Morphological Analizer.” http://taku910.github.io/mecab/ports.pdf

[7] MACROMILL. ”日本人が最も好きなスポーツ” https://www.macromill.com/r_data/20151009sports/20151009sports.pdf