駄洒落検出用フレームワークの公開と動画コメントデータへの適用

A Pun Detection Framework and its Application to Video Comment Data

谷津 元樹[1*]　荒木 健治[2]
Motoki YATSU[1]　Kenji ARAKI[2]

[1] 青山学院大学　[2] 北海道大学
[1] Aoyama Gakuin University　[2] Hokkaido University

Abstract: 近年，ロボットや対話エイジェントによる言語的ユーモア理解の必要性が高まっている．この目的のために深層学習を含む多様な技術が利用可能となっているが，現状では実験の度に新たなシステムを構築する必要がある．そこで，駄洒落の認識・理解のためのシステムの開発研究を促進するため，『駄洒落データベース』に準拠した入力を利用する，教師あり機械学習を用いた駄洒落認識フレームワークを構築した．本発表では，フレームワークの紹介及び動画コメントデータにおける駄洒落検出タスクへの適用例について述べる．

1　はじめに

ヒトの社会的な生活において，ユーモアは欠かすことにできない心理的現象であるという認識から，近年，ロボットや対話エイジェントによる言語的ユーモア理解の必要性が高まっている．ユーモアには多種の表現方法があるが，言語を用いたユーモアにおいて重要な一角を占めているのが駄洒落である．

著者らは，駄洒落を人間とロボット並びに対話エイジェントとの相互的接触において，対人コミュニケーションと同様に，駄洒落を言語的ユーモアとして活用する方法を研究し，システムの構築によりその有効性を実証してきた．それは駄洒落の生成手法 [1] に始まり，駄洒落の検出 [2]，そしてこれらを利用した実際の対話エイジェントの構築 [3] へと至っている．

駄洒落に限らず，言語処理の研究においては良質なコーパスが必要となる．近年まで，駄洒落を充分な規模収録し，なおかつ研究用途に利用可能なデータセットは存在しなかった．荒木ら [4, 5] は初めて，その構築及び公開を行ない，また研究の進展につれて求められるようになった面白さの評価結果を取り入れる等，コーパスとしての質の改善を続けている．

2　駄洒落認識フレームワーク

駄洒落を検出し面白さを自動評価するためのモジュールまたはシステムの構築を支援するために開発したフレームワークについて述べる．

*連絡先： 青山学院大学理工学部情報テクノロジー学科
相模原市中央区淵野辺 2 丁目 5-10-1
E-mail: yatsu@it.aoyama.ac.jp

本フレームワークは，駄洒落データベース [4, 5] のフォーマットに準拠した入力データを訓練データとして教師あり学習を行い，プレインテキストでの入力文に対する駄洒落の検出または自動評価を行なうシステムの開発を簡便化することを目的とし，Python 言語を用いて構築したものである．

教師あり機械学習モデルのもつ素性関数を，Python モジュール `features.py` に記述することにより，自由に変更することが可能である．2019 年 11 月時点の公開版では，以下の素性関数を実装している：

i. 音韻類似度素性．
本素性では，入力文を形態素解析する際に得られる読み情報よりモーラ単位の音素ペア列を作成し，子音および母音について種表現候補と変形表現候補同士の類似度を求める．音素同士の音韻類似度の算出には，事前に開発データより Smith-Waterman 法 [6] でのアライメントにより構築された (子/母) 音の音韻類似度行列 [7] を用いる．

ii. Bag-of-words 素性．

本フレームワークは，GitLab リポジトリ[1]において公開中である．

3　動画コメントデータ

オープンデータであり豊富な自然発話文のコーパスである動画コメントからの駄洒落検出について述べる．国立情報学研究所 IDR より公開されているニコニコ

[1] https://gitlab.com/m-yatsu/djr_wpsm

データセット[2]は，動画投稿サイト・ニコニコ動画[3]にアップロードされた動画に対し 2017 年 3 月より 2018 年 11 月までの間に書き込まれた全コメントの，JSON 形式によるダンプデータである．

3.1 動画コメントデータからの駄洒落検出例

前述の動画コメントデータより，**2** 節にて述べたフレームワークを用いて検出に成功した駄洒落の例を下記に示す各行冒頭はニコニコ動画における動画 ID を示す．

sm11284310: その裁量は最良
sm17602682: こーしゅかの発音はこーっすか？
sm19000864: IKEA に行けや！
sm19009608: ←こんなマジニいたらマジに恐い
sm19027346: 「色についていろいろ教えました」
sm20191681: 遭難したんですかそうなんです・・・
sm24090089: 遊園地ならぬ終焉地ってか
sm25601593: シメっぽい曲でメシが食えるのかw
sm34056964: 内装は無いそうです (ﾎﾞｿｯ)

例示したものはすべて，音韻的に類似した部分である種表現及び変形表現が同一入力文中に共起する**併置型**の類型に属するものである．一方で，種表現が入力文中に生起せず，文脈上あるいは共有された知識の一部として存在する**重畳型**の駄洒落に関しては，現時点では検出の成果は得られていない．

3.2 動画コメントデータに見られる駄洒落の頻度傾向

ニコニココメントデータより無作為抽出した 300 件のコメントに対し，第 1 著者により駄洒落の類型分類を行なったところ，重畳型駄洒落の占める比率は 42.2% となった．これは 2018 年版の駄洒落データベース [5] に対し行われた頻度分析において，重畳型は総分類数の 1.5%に留まったことと対照的であり，オープンデータにおける駄洒落類型の存在比率の多様性を示しているといえる．

4 おわりに

本稿では，教師あり機械学習を用いた駄洒落認識フレームワークの開発成果について述べ，オープンデータの一例である動画コメントデータにおける駄洒落検出タスクへの適用について考察を行なった．

今後の展開として考えられる事項について述べる．

まず，オープンデータにおいて多数潜在していると考えられる重畳型駄洒落の検出を試み，データに基づく客観的基準を用いたその面白さの定量的評価を目指すことは有益と考えられる．

また，**3.2** 節で考察したように，重畳型駄洒落は，オープンデータにおいて潜在的に大きな存在であることが予想される．非テキスト的情報からの種表現の参照のされ方に関し，意味・形態・音韻にわたる多様な言語現象を考慮しつつ，今後はより詳細な分析を進めたい．

謝辞

本研究は科研費（基盤研究 (C)17K00294）の助成を受けたものである．

参考文献

[1] J. Sjobergh and K. Araki, ”Robots Make Things Funnier, New Frontiers in Artificial Intelligence”, Lecture Notes in Artificial Intelligence, vol. 5447, pp. 306–313, 2009.

[2] 谷津元樹, 荒木健治: ”子音の音韻類似性及び SVM を用いた駄洒落検出手法”, 知能と情報, vol. 28, no. 5, pp. 875–886, 2016.

[3] 谷津元樹, 荒木健治: ”話題遷移に適応した駄洒落ユーモア統合型対話システムの性能評価”, 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B601-3, pp.23–27, 2016.

[4] 荒木健治, 内田ゆず, 佐山公一, 谷津元樹: ”駄洒落データベースの構築及び分析”, 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B702-3, pp. 13–24, 2017.

[5] 荒木健治, 内田ゆず, 佐山公一, 谷津元樹: ”駄洒落データベースの拡張及び分析”, 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B803-1, pp. 1–15, 2018.

[6] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” Journal of molecular biology, vol. 147, no. 1, pp. 195–197, 1981.

[7] B. Hixon, E. Schneider, and S. L. Epstein, “Phonemic Similarity Metrics to Compare Pronunciation Methods,” in Proc of 12th Annual Conference of InterSpeech, pp. 825–828, 2011.

[2] https://www.nii.ac.jp/dsc/idr/nico/
[3] http://nicovideo.jp