

駄洒落を含む対話における可読性と面白さの相関分析

花房竜馬¹ 荒木健治²¹北海道大学大学院情報科学院 ²北海道大学大学院情報科学研究院

(hanabusa, araki)@ist.hokudai.ac.jp

概要

スマートスピーカーなどのデバイスに搭載されている対話エージェントとの対話は人間同士の対話と同等の面白さを感じることができない。この問題を解決するためには、対話中の文脈情報を考慮した高度なユーモア処理を実現する必要がある。これまで対話の文脈情報を含むユーモアの面白さについて、評価及び分析はなされていない。本稿では対話文脈を考慮した駄洒落の面白さを評価する標準データベースを構築するために、駄洒落を含む対話を収集し、12,000件の駄洒落が含まれる対話に対して対話の面白さを複数人で評価し、さらに、対話の可読性が面白さに与える影響について、相関係数を用いて分析した結果について述べる。

1 はじめに

新型コロナウイルス感染症の蔓延による生活様式の大きな変化に伴い、精神的ストレスを感じる人が急増している[1]。このような社会的問題に対処するために、新たな生活様式に沿った孤独や不安を緩和するためのメンタルケアの確立が必要となる。

孤独や不安を緩和するためのメンタルケアとして、友人や家族との対話が挙げられる。しかし、対話相手の心理状態によって笑いを生むことが困難なことも少なくない。これらのことから、対話相手を担う存在として、対話によって面白さや楽しさを与える対話システムの実現が緊急の課題となっている。

現在、スマートフォンやスマートスピーカーの普及によって、人間と対話エージェントとの対話が日常的に行われている。これらのデバイスに搭載される対話エージェントは対話によって何らかのタスクを実行することを目的としたシステムである。そのため、対話によって面白さや楽しさを与えることができない。この点を踏まえ、対話による面白さや楽

しさを与えることを目的とした非タスク指向型システム(雑談システム)の研究が行われている[2,3]。しかし、これらの研究では人間同士の対話と同等以上の面白さや楽しさを与えるというレベルに達していない。この問題を解決するためには、人間と同等の応答精度に加えて、対話中の文脈情報からユーモア表現を理解して笑い、適切なタイミングで対話相手を飽きさせないようにユーモアを発話するといった高度なユーモア処理が必要となる。

これまでのユーモアに関する研究として、我々は67,000件の駄洒落を収録した大規模なデータベースの構築[4,5]を行った。駄洒落の面白さを評価するための手法を確立するために、複数人による駄洒落の面白さの評価及び分析[6]も行った。

しかし、これらの研究において、文脈情報は1文内の文内文脈に限定されている。そのため、対話の文脈情報を含むユーモアの面白さについて、評価及び分析はなされていない。

本稿では、対話文脈を考慮した駄洒落の面白さを評価する標準データベースの構築のために、駄洒落を含む対話の収集を行い、複数人によって対話の面白さを評価し、さらに、対話の可読性が面白さに与える影響について、対話の面白さと表層情報との相関を用いて分析した結果について述べる。

2 駄洒落を含む対話の収集

人間同士の雑談対話を収録したコーパスとして、日本語日常会話コーパス[7]や名大会話コーパス[8]が挙げられる。しかし、これらのコーパスは一般的な会話文章を収録しているため、コーパス中に含まれる駄洒落は非常に少ないと考えられる。このことから、既存のコーパスから駄洒落を含む対話を大量に収集することは困難である。

ユーザ同士の対話が Web 上で公開されたソーシャル・ネットワーキング・サービスとして、Twitterⁱ

ⁱ <https://twitter.com/>

表 1 収集結果の件数及び割合

	件数	割合[%]
駄洒落を含む対話	26,739	5.9
駄洒落を含む会話	81,454	18.0
1者による連続発話	1,758	0.4
駄洒落を含むツイート単体	341,904	75.7
総計	451,855	100.0

が挙げられる。日本において、Twitter は 2021 年時点で約 5,310 万人が利用している。そのため、既存のコーパスと比べて、Twitter には駄洒落を含む対話が比較的多いと考えられる。このことから、本稿では Twitter を用いて、駄洒落を含む対話を収集する。

一般的に、Twitter 上に存在する駄洒落を含む対話の構成は以下の 3 種類が挙げられる。

- ① 最初に駄洒落を含むツイートが出現する対話
- ② 中間に駄洒落を含むツイートが出現する対話
- ③ 最後に駄洒落を含むツイートが出現する対話

駄洒落を含む対話の面白さは駄洒落を含むツイートの前後のツイートによって変化すると考えられる。また、駄洒落を含むツイートから前後 5 件以上のツイートは駄洒落とは関連性のない話題が進行すると考えられる。これらのことから、駄洒落を含む対話の範囲は駄洒落を含むツイートを基準とした前後それぞれ 0 から 5 つのツイートとする。

本稿において、駄洒落を含むツイートの前後のツイートの数が 0 となるものは駄洒落を含むツイート単体とする。

2.1 収集方法

駄洒落を含む対話を収集するために、10,000 件の駄洒落を用いて、Twitter 上で完全一致検索を行い、検索結果をクロウリングする。

駄洒落を含む対話の範囲の基準となるツイートは完全一致検索に用いた駄洒落を含むツイートとする。

完全一致検索に用いる駄洒落は駄洒落データベースに収録されている番号 1~10,000 の駄洒落を使用する。駄洒落データベースは 67,000 件の駄洒落が収録された大規模なデータベースであり、Web サイト上ⁱⁱで一般公開されている。収録された駄洒落には面

ⁱⁱ <http://arakilab.media.eng.hokudai.ac.jp/~araki/dajare.htm>

白さの度合いを 5 段階で示す面白さのスコアが付与されている。駄洒落データベースにおける面白さのスコアの評価はクラウドソーシングによって雇用した 3 名の評価者によって行われている。

クロウリングには Twitter から提供されている学術研究向け API である Twitter API for Academic Researchⁱⁱⁱを使用する。

2.2 収集結果

表 1 に収集結果の件数及び割合を示す。ここで、本稿における対話は 2 者による言葉の掛け合い、会話は 3 者以上による言葉の掛け合いとする。表 1 に示すように、駄洒落を含む対話は 26,739 件となった。また、駄洒落を含むツイート単体の件数は 341,902 件となった。このことから、Twitter 上で行われる駄洒落を含む対話は駄洒落を含むツイート単体よりも比較的少ないことが確認された。

また、収集されたものには駄洒落を含む会話が 81,454 件(18.0%)、1 者による連続発話が 1,758 件(0.4%)、が含まれている。

完全一致検索に用いた駄洒落 10,000 件のうち、収集された駄洒落は 1,147 件(11.5%)となった。

3 駄洒落を含む対話の面白さの評価

これまでの研究において、駄洒落を含む対話の面白さの指標は存在しない。このことから、駄洒落を含む対話に対して、クラウドソーシングで雇用した 16 名の評価者に面白さを評価していただいた。

3.1 評価方法

評価対象は駄洒落を含む対話 26,739 件から無作為に抽出した 12,000 件とする。

駄洒落を含む対話の面白さの評価には 200 件あたり 1 時間前後が必要である。そのため、12,000 件の作業には膨大な時間的コストがかかる。それに加えて、駄洒落単体の面白さの評価[6]において、膨大な量の評価を行うと駄洒落に対する飽きが生じ、面白さが低く評価されることも確認されている。このことから、12,000 件の駄洒落を含む対話を 1,000 件単位で分割した。その後、3 名による評価が付与されるように評価者を割り当てた。

図 1、表 2 に駄洒落を含む対話の面白さの評価の

ⁱⁱⁱ <https://developer.twitter.com/ja/products/twitter-api/academic-research>

◆対話番号:	8030
◆ID:	378933
◆対話に含まれる駄洒落: 敗者は歯医者	
~~~~~	~~~~~
U1:	虫歯稀なら滅多に蝕まれないよ
U2:	今虫歯に蝕まれてるっぽくて厳しい
U1:	<b>敗者は歯医者</b> にかかるしかねえな
U2:	敗者なのでオイオイ泣いてる
~~~~~	~~~~~
☆対話の面白さの評価:	4
----	----

図 1 駄洒落を含む対話の面白さの評価の例

表 2 面白さのスコア

スコア	説明
5	とても面白い
4	面白い
3	普通
2	面白くない
1	非常に面白くない (駄洒落ではないと思うものを含む)

例と面白さのスコアを示す。図 1 における太字は収集時の完全一致検索に用いた駄洒落の種表現と変形表現を示す。また、対話番号は 1~12,000 件の駄洒落を含む対話の番号を示す。U1, U2 はそれぞれ対話中の話者を示す記号である。対話の面白さの評価は評価者によって表 2 に示す面白さのスコアが付与される。面白さのスコアは面白さの度合いを示す 1~5 の整数値である。

3.2 評価結果

図 2 に評価者全員のスコアの分布を示す。図 2 に示すように 1 点から 3 点が全体の 91.3% を占めている。一方、4 点は 7.2%、5 点は 1.5% と極端に少ない結果となった。そのため、評価者全員のスコアの平均は 1.89 ポイントとなった。

1 点から 3 点の割合が 91.3% と比較的多い要因として、Twitter 上の対話には第三者から面白さを評価される前提がないことが挙げられる。Twitter における対話はユーザ同士のコミュニケーションとして行われる。そのため、一般的なユーザは第三者から面白いと感じられるような対話をするという意図がない。このことから、駄洒落を含む対話に対して第三者が対話の面白さを評価した場合、面白くないまたは普通と感じるものが比較的多いと考えられる。

表 3 平均スコアと表層情報の相関

	ピアソンの相関係数
対話の長さ	-0.18
発話の平均文字数	-0.31
駄洒落の出現位置	-0.23

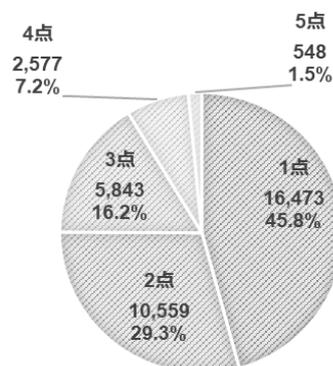


図 2 評価者全員のスコアの分布

4 可読性の面白さへの影響

駄洒落を含む対話の面白さを変化させる要因として、対話内容の可読性が挙げられる。駄洒落を含む対話の可読性を表す表層情報として(A)対話の長さ、(B)発話の平均文字数、(C)駄洒落の出現位置の 3 種類が考えられる。この点を踏まえ、本稿では駄洒落を含む対話の面白さの指標と(A)、(B)、(C)の表層情報との相関係数を算出する。これにより、対話の可読性が面白さに与える影響を分析する。

4.1 可読性と面白さの相関分析

駄洒落を含む対話の面白さの指標として、3 名の評価者のスコアの平均を用いる。本稿において、3 名の評価者のスコアの平均を平均スコアと呼ぶ。

2 章で述べたように、本稿における対話の範囲は駄洒落を含むツイートを基準とした前後それぞれ 0 から 5 つのツイートとした。このことから、駄洒落を含む対話の長さは 2 から 11 の整数値である。

駄洒落の出現位置は 1 から 3 の整数値で表す。そのため、対話の 1 ツイート目に駄洒落が出現したものは 1、対話の最後のツイートに駄洒落が出現したものは 3、それら以外は 2 となる。

12,000 件の駄洒落を含む対話 1 件ごとの平均スコアと 3 種類の表層情報の相関をピアソンの相関係数を用いて算出する。

・対話に含まれる駄洒落: 筑波山に嘯みつくばあさん
 ・3名の駄洒落を含む対話の面白さスコア: 5, 5, 4
 ・平均スコア: 4.67
 U1: 明日は筑波山行こう。
 U2: 筑波山に嘯みつくばあさん
 U1: ちょっとだけうまい、ちょっとだけ
 U2: 最近思い付いて頭から離れない

図 3 評価者全員のスコアが高い対話の例

・対話に含まれる駄洒落: 北海道はでっかいどう
 ・3名の駄洒落を含む対話の面白さスコア: 2, 1, 2
 ・平均スコア: 1.67
 U1: ひとつ鍋美味しくて大好きです ♡ 六花亭さんバ
 ターサンド以外も全部美味しいです ☺
 U2: 美味しいですね!!! 六花亭で一番好きかも
 しれないです お土産のイメージ強いですけど、
 時々自分用に買いたくなります(笑)
 U1: 北海道は六花亭のおやつがデパートでバラ売り
 されてて羨ましい!!! 行って行くたびに思っ
 てます笑 いつも物産展でも買っちゃいますね ☺
 U2: ばら売りされてますねえ…好きなのをぼいぼい
 っとかごに入れて買って、紙袋に入れて持って
 帰ります♪ なんかお話ししたら食べたくなっ
 てきました ☺ しかし最寄りの六花亭が札幌なの
 で私もなかなか遠いです…笑
 U1: そうか、北海道はでっかいどう! でした…北海
 道といえども簡単には手に入らないんですね ☺
 物産展やらないかなー!

図 4 評価者全員のスコアが低い対話の例

表 3 に平均スコアと表層情報の相関係数を示す。
 表 3 に示すように、平均スコアと発話の平均文字数、
 駄洒落の出現位置の相関係数は-0.31, -0.23 となり、
 弱い負の相関があることが確認された。また、平均
 スコアと対話の長さの相関係数は-0.18 となった。こ
 のことから、対話の長さが平均スコアに及ぼす影響
 は他 2 つの表層情報と比べて小さいと考えられる。

4.2 考察

図 3 に評価者全員のスコアが高い対話の例を示す。
 図 3 における U1, U2 はそれぞれ対話中の話者を示
 す記号である。3 名の面白さの評価は 5, 5, 4 とな
 っている。そのため、平均スコアは 4.67 となる。

図 3 の例では U1 の「明日は筑波山に行こう」と
 いう発話に対して、U2 が「筑波山にかみつくばあさ

ん」という駄洒落を返している。

図 3 のような対話は文章量が少なく、内容を理解
 することが容易である。また、この対話は駄洒落に
 関する話題によって進行している。そのため、第三
 者から駄洒落をユーモアとして認識することも容易
 であると考えられる。

図 4 に評価者全員のスコアが低い対話の例を示す。
 図 4 においても U1, U2 はそれぞれ対話中の話者を
 示す記号である。3 名の面白さの評価は 2, 1, 2 と
 なっている。そのため、平均スコアは 1.67 となる。

図 4 の例では北海道の洋菓子屋に関する話題から
 U1 が「北海道はでっかいどう」という駄洒落を含む
 発話を行うことで、対話が終了している。

図 4 のような対話は発話の平均文字数が 62 文字
 であるため、図 3 のような対話と比べて文章量が多
 い。さらに、対話の最後に駄洒落を含む発話が出現
 しており、対話者のリアクションがない、そのため、
 第三者から駄洒落をユーモアとして認識することが
 困難であると考えられる。

これらのことから、可読性が高い対話は駄洒落を
 ユーモアとして認識することが容易であり、面白さ
 が高い傾向があると考えられる。

5 おわりに

本稿では、対話文脈を考慮した駄洒落の面白さを
 評価する標準データベースを構築するために、はじ
 めに、Twitter 上に存在する 26,739 件の駄洒落を含む
 対話を収集し、12,000 件の対話に対して面白さの評
 価を行った。評価の結果として、評価者全員のスコ
 アの 91.3% は 1 点から 3 点であることが確認された。

次に、対話の面白さがどのような要因によって変
 化するのかについて、対話の可読性に着目した分析
 を行った。ここで、12,000 件の駄洒落を含む対話 1
 件ごとの面白さのスコアの平均と対話の表層情報の
 相関をピアソンの相関係数を用いて算出した。その
 結果、対話の面白さと発話の平均文字数及び駄洒落
 の出現位置には弱い負の相関があることが示された。

今後の課題として、まず駄洒落データベース中の
 10,001 番以降の駄洒落を用いた駄洒落を含む対話の
 追加収集を行うことが挙げられる。次に、第三者か
 ら面白いと感じられる対話をより多く収集するため
 に、人手で駄洒落を含む対話を作成することが挙げ
 られる。最後に、収集及び作成した対話を用いて対
 話文脈を含む駄洒落の面白さを評価する標準デー
 タベースを構築する予定である。

謝辞

本研究は JSPS 科研費 21K12007 の助成を受けたものである。

参考文献

- [1] 厚生労働省, “新型コロナウイルス感染症に係るメンタルヘルスに関する調査結果の概要について”, https://www.mhlw.go.jp/stf/newpage_15766.html, (参照 2020-08-01).
- [2] Rafal Rzepka, Shinsuke Higuchi, Michal Ptaszynski, Pawel Dybala and Kenji Araki: When Your Users Are Not Serious, 人工知能学会論文誌, Vol.25, No. 1, pp.114-121, 2010.
- [3] Arnaud Jordan and Kenji Araki, Real-time Language-Independent Algorithm for Dialogue Agents, 知能と情報 (日本知能情報ファジィ学会誌), Vol.28, No.1, pp.535-555, 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, 佐山公一, 内田ゆず, 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] 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を用いた面白さの評価及び分析, 人工知能学会第2種研究会 ことば工学研究会資料, SIG-LSE-B902-4, pp.63-76, 2019.
- [7] Fujimura,Itsuko, Shoji Chiba and Mieko Ohso, Lexical and Grammatical Features of Spoken and written Japanese in Contrast: Exploring a lexical profiling approach to comparing spoken and Written corpora, Proceedings of the VIIth GSCP International Conference. Speech and Corpora, pp.393-398, 2012.
- [8] Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yuka Watanabe and Yasuyuki Usuda, Design and Evaluation of the Corpus of Everyday Japanese Conversation, Proceedings of LREC2022, pp.5587-5594, 2022.

A 面白さの評価の詳細

A.1 評価者の属性及び作業件数

本稿では、評価にかかる時間的コストの高さと評価者の駄洒落を含む対話に対する飽きを考慮し、クラウドソーシングを用いて 16 名の評価者を雇用した。また、評価者 16 名には 1 件の対話に対して、3 名の評価が付与されるように 1,000~3,000 件の作業を割り当てた。ここで、1,000 件単位の対話に対する評価者の割り当ては雇用順に沿って行った。

表 4 に評価者の属性及び評価を実施した作業番号を示す。クラウドソーシングの Web サイト上には業務の依頼者から作業者に対して返信速度・納品実績などを 5 段階で評価した数値が与えられている。今回の評価者は全員評価値が 5.0 であるため、一定レベルの能力を満たしているものと考えられる。

A.2 駄洒落を含む対話に対する飽き

駄洒落単体の面白さの評価[6]において、膨大な量の評価を行うと駄洒落に対する飽きが生じ、面白さが低く評価されることが確認されている。このことから、作業番号 1 から 3,000 の各評価者のスコアと平均スコアの 500 件ごとの平均値を算出し、駄洒落を含む対話に対する飽きを分析する。

図 5 にスコアの平均値の推移を示す。図 5 に示すようにいずれの評価者も 500 件ごとにスコアが低下している。しかし、500 件時点と 3,000 件時点の平均スコアの差は 0.2 ポイントとなった。このことから、駄洒落を含む対話に対する飽きの影響は小さいと考えられる。

A.3 評価者間の一致度

面白さの評価における一致度の算出にはコーヘンの κ 係数を用いる。ここで、分析する一致度は駄洒落を含む対話に対して 5 段階の面白さの度合いを付与するという順位尺度に基づいた評価に対する一致度である。

12,000 件の面白さの評価に対する一致度を算出するために、はじめに 1,000 件ごとの 2 者間の一致度を算出する。ここで、1,000 件ごとの評価者の組み合わせは 3 通りであるため、3 つの一致度が算出される。その後、36 個の一致度の平均を全体の一致度として算出する。

表 4 評価者の属性及び評価を実施した作業番号

作業者名	年代	性別	経歴	作業番号
評価者1	30代前半	女性	自営業	1~3,000
評価者2	30代前半	女性	その他職種	1~3,000
評価者3	30代前半	男性	営業職	1~3,000
評価者4	50代前半	女性	請負	3,001~4,000
評価者5	30代後半	女性	専業主婦	4,001~6,000
評価者6	30代前半	女性	薬剤師	3,001~6,000
評価者7	30代後半	男性	事務職	3,001~6,000
評価者8	50代後半	女性	プログラマ	6,001~9,000
評価者9	20代後半	女性	事務職	6,001~9,000
評価者10	30代前半	男性	その他職種	6,001~9,000
評価者11	40代前半	女性	専業主婦	9,001~11,000
評価者12	40代後半	女性	旅行業	11,001~12,000
評価者13	20代後半	男性	その他職種	9,001~12,000
評価者14	30代前半	女性	事務職	9,001~10,000
評価者15	20代後半	女性	学生	10,001~11,000
評価者16	40代前半	女性	専業主婦	11,001~12,000

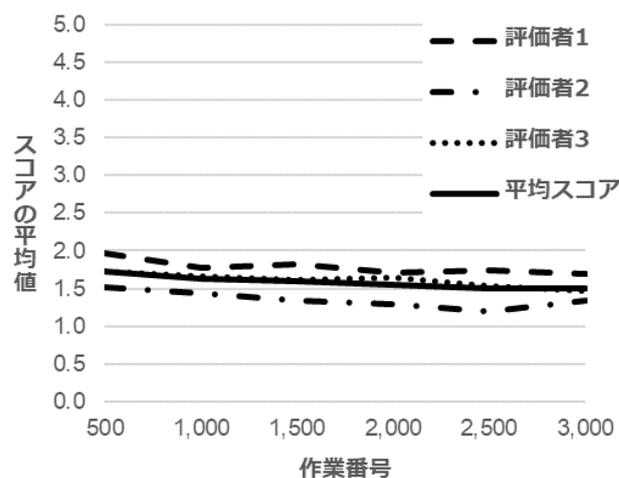


図 5 平均値の推移

12,000 件の面白さの評価に対する一致度は 0.29 となり、おおむね一致という結果となった。駄洒落単体の面白さの評価[6]における評価者間の一致度は 0.10 未満となっている。そのため、駄洒落を含む対話の面白さの評価は駄洒落単体の評価に比べて評価者間で一致する割合が高い結果となった。