

# クラウドソーシングによる大喜利の面白さの構成要素の分析

中川 裕貴 村脇 有吾 河原 大輔 黒橋 禎夫

京都大学 大学院情報学研究科

{nakagawa, murawaki, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

## 1 はじめに

人と計算機が高度で円滑なコミュニケーションを行うためには面白さの計算機処理は重要である。しかし、面白さの自動理解は依然として困難な課題であり、観測されるテキストとその背後にある面白さとの間の大きな隔たりはほとんど埋められていない。

自然言語処理における一般的な取り組みは、テキストから出発して自動理解に到達しようと試みる。典型的には、面白さの理解を教師あり機械学習問題として定式化し、定量化された面白さを特徴量の組み合わせにより予測しようとする。しかし、特徴量として用いられるのは、bag-of-words、n-gram、単語埋め込み、感情極性といった、いずれも表層の手がかりにすぎず、面白さとの落差はあまりにも大きい。

そこで、本研究では、面白さの側から歩み寄りという新たな取り組みを提案する。鍵となるアイデアは、面白さを相対的にわかりやすい構成要素に分解するというものである。たとえ構成要素を計算機に予測させることが現時点では難しいとしても、それが面白さよりも容易であれば、この困難な課題を漸進的に解決するための指針として機能することが期待される。

本研究では、面白さ理解のための題材として大喜利に着目する。大喜利は、お題と回答で完結するシンプルな形式であるとともに、データの入手が容易であることから分析に適している。まず、定量分析に必要な大喜利データセットをクローリングにより構築する。次に、先行研究を参考に、面白さの構成要素として6個の候補を挙げる。クラウドソーシングを用いて、面白さと同時に構成要素候補を尺度として、人手により大喜利の評点付けを行う。予備調査により6個から3個に絞り込んだ構成要素を用いて面白さを予測したところ、ある程度高い精度が得られた。この結果は、大喜利の面白さがそれらの要素によって説明できることを示唆する。最後に、人手により定量化された構成要素を参考にして、計算機による予測に向けた予備的検討を行う。

## 2 関連研究

自然言語処理における面白さの研究対象の多くはジョークである。ジョークの研究は主に認識と生成という2つの側面から行われている。ジョークの認識は入力された文または文章がジョークか否かを二値分類する問題として定式化されることが多い。Yangらは言語的な特徴量を手掛かりにジョークの自動認識を試みている[4]。彼らはジョークの面白さに繋がる言語的な構成要素として incongruity, ambiguity, phonetic style, personal affect の4つが存在することを示している。他にも極性表現を特徴量に組み入れた研究[3]やCNN (convolutional neural network) を用いた研究[2]がある。一方、ジョークの生成の例としては、Binstedらの駄洒落なぞなぞ生成システム JAPE[1]や、ニューラルネットを用いたシステム[5]が挙げられる。

日本における主なジョークは駄洒落である。駄洒落の従来研究には、駄洒落なぞなぞ生成システム BOKE [6]や駄洒落の識別を試みたもの[7]がある。駄洒落は音韻という表層の手がかりが効果的という点で比較的簡単な問題といえる。

## 3 大喜利データセット

### 3.1 大喜利

大喜利は与えられたお題に対してひねりのある回答をすることで笑いを誘う演芸である。テレビ番組として放送されることも多く、広く一般の人にも親しまれている。本研究で扱う大喜利の一例を示す。

(1) 【お題】友達からのお見舞い、花やフルーツが並ぶ中、貧乏な田中くんが持ってきたものとは…?

【回答】隣の病室にあったマスクメロン

(とんち道場<sup>4</sup>から引用)

本研究が大喜利を対象とするのは次の理由からである。まず、駄洒落と比較すると、大喜利の面白さは意味内容への依存が大きく、より挑戦的な課題である。一方で、お題と回答で完結するシンプルな形式であることから、漫才のような複雑な談話形式とくらべて自

サイト	お題数	回答数	お題平均文字数	回答平均文字数	回答者
IPPON グランプリ	67	543	26.7	16.3	芸人
大喜利グランプリ	614	30,798	27.9	20.0	一般人
大喜利どっとねっと	822	4,465	29.1	16.5	一般人
とんち道場	3,785	47,626	22.6	17.8	一般人
大喜利プラス	85,366	1,893,885	21.0	12.5	一般人

表 1: データセットに含まれる大喜利のお題と回答の数と平均文字数

動理解に向けた分析を行いやすい。加えて、同じお題に対して面白さに違いのある回答を複数得やすいという点も分析に好都合である。

### 3.2 大喜利データセットの構築

定量分析を可能とするために、まずは大喜利データセットを構築した。構築は次の5サイトのクローリングにより行った。(1) IPPON グランプリ (第1回から第8回までを書き起こしたサイト)<sup>1</sup>、(2) 大喜利グランプリ<sup>2</sup>、(3) 大喜利どっとねっと<sup>3</sup>、(4) とんち道場<sup>4</sup>、および(5) 大喜利プラス<sup>5</sup>。IPPON グランプリはテレビ番組で放送された芸人による大喜利を収集したものであり、残りはいずれも一般人による大喜利を集積したものである。

各サイトからはお題、回答、およびその他の情報(順位や票数など)を取得した。IPPON グランプリにはイラストのような非言語的手段を用いる回答も含まれるが、今回はテキストデータのみから成り立つお題と回答を収集した。構築したデータセットの統計を表1に示す。

## 4 面白さの構成要素

### 4.1 大喜利の面白さ

本節では大喜利の面白さに繋がる要素を検討する。大喜利は出題者に回答者が答えるという形式を取ることから、回答は出題されたお題を踏まえたものでなければならない。また、回答は聞き手にとってわかりやすいものである必要がある。反対にこれらの条件を満たしていなければ、回答として成り立たず、したがって面白さも生まれまいだろう。

上記の条件は一般に会話に必要なものであり、もちろんそれだけで面白さを生み出すわけではない。面白い回答には聞き手の予想を裏切る意外性が欠かせない。つまり、お題と回答の間に何らかの文脈的なずれがありつつ、それでも回答が聞き手にとって納得感が

<sup>1</sup><https://blog.goo.ne.jp/takawo2222>

<sup>2</sup><https://oogiri.jp>

<sup>3</sup><http://www.ohgiri.net/ohgiri/>

<sup>4</sup><http://tonchi.jp>

<sup>5</sup><http://oogiri.biz>

設定した設問ID: 10000

大喜利のお題と回答について以下の質問に答えて下さい。

お題	殴られても蹴られてもチェ・ホンマンが難さなかった物とは？
回答	西山茉希の飲みかけのペットボトル

[質問] 回答はお題の内容を踏まえていますか？  
次の選択肢からあなたの感覚に近いものを選んで下さい。

5: 非常にそう思う

4: ややそう思う

3: どちらとも言えない

2: あまりそうは思わない

1: 全くそうは思わない

図 1: クラウドソーシングのタスク回答画面

あるという絶妙なバランスが面白さを生み出すと考えられる。

### 4.2 特徴量の設計

前節での議論を踏まえて大喜利データセットに人手で付与する特徴量の設計する。特徴量の付与にはクラウドソーシングを用いるため、わかりやすさに配慮して、以下のように質問形式を採用する。

- F1: 回答はお題の内容を踏まえていますか？  
 F2: 回答はお題に対するボケとしてわかりやすいですか？  
 F3: 回答は簡潔に書かれていますか？  
 F4: 回答は現実世界でおこりえそうですか？  
 F5: 回答はあなたの想定したボケと異なりますか？  
 F6: 回答に新しさ、もしくは新鮮味を感じますか？

### 4.3 特徴量と面白さの評点付け

3節で得られた大喜利データセットに対して、クラウドソーシングを用いて、特徴量と面白さを人手で評点付けする。クラウドソーシングのタスク回答画面を図1に示す。ワーカは、特徴量と面白さのそれぞれについて、非常にそう思う場合(5)から全くそうは思わない場合(1)までの5段階の評点を付与する。

特徴量	平均二乗誤差
全て	0.0674
F1	0.0952
F2	0.0839
F3	0.1662
F4	0.2100
F5	0.2192
F6	0.1256

表 2: 各特徴量単独での回帰モデルの精度

特徴量	平均二乗誤差
F1 F2 F4 F5 F6	0.0671
F1 F2 F3 F4 F5 F6	0.0674
F1 F2 F4 F6	0.0690
F1 F2 F6	0.0703
F1 F2 F3 F4 F6	0.0712

表 3: 特徴量の組み合わせの精度上位 5 件

各お題・回答ペアに対して、15 人の評点を収集する。今回はその平均値を特徴量と面白さのスコアとして採用する。なお、お題・回答ペアに対して、特徴量と面白さを一通り単一のワークに評点付けさせる。これは、将来的に、ワークの個人差を反映したスコア集計を行えるようにするためである。

#### 4.4 構成要素を用いた面白さの予測

クラウドソーシングにより得られたスコアを用いて、4.2 節で設計した特徴量が有効かを検証する。この問題は、各お題・回答ペアについて、その面白さを特徴量を用いて予測する回帰問題として定式化できる。回帰モデルとしては、gradient boosting と random forests を組み合わせたアンサンブル学習器である XGBoost Regression<sup>6</sup>を用いる。10 分割交差確認を用い、平均二乗誤差をモデルの精度とする。

もし提案する特徴量によって面白さが高い精度で予測できるなら、面白さはそれらの構成要素によって説明できるといえる。また、特徴量の様々な組み合わせを比較することで、面白さの説明に有効な特徴量の組み合わせを特定できる。なお、大量のデータに対して効率的にクラウドソーシングを行うために、小規模な予備調査により有効な特徴量を絞り込む。その後、それらの特徴量に限定して大規模なクラウドソーシングを行い、改めて精度評価を行う。

## 5 実験

### 5.1 予備調査による特徴量の絞り込み

最初に小規模な予備調査を行い、面白さの予測に有効な特徴量を絞り込んだ。まず、大喜利データセット

<sup>6</sup><https://github.com/dmlc/xgboost>

特徴量	平均二乗誤差
全て	0.0483
F1	0.0884
F2	0.0768
F6	0.0660
F1 F2	0.0740
F2 F6	0.0487
F1 F6	0.0559
平均値ベースライン	0.2252

表 4: 大規模データにおける面白さの予測精度

から少数のお題・回答ペアを抽出した。特定の大喜利サイトに偏らないようにするために、各サイトからお題・回答ペアをそれぞれ 50 個ランダムに抽出し、計 250 ペアを得た。これらのデータに対し、Yahoo!クラウドソーシング<sup>7</sup>を用いて特徴量と面白さに評点付けを行った。最後に回帰モデルを訓練し、平均二乗誤差を得た。

結果を表 2、3 に示す。表 3 に頻出する特徴量は F1「回答はお題の内容を踏まえていますか?」、F2「回答はお題に対するボケとしてわかりやすいですか?」、F6「回答に新しさ、もしくは新鮮味を感じますか?」の 3 つであり、これらが面白さに最も寄与する特徴量といえる。以後これらの特徴量を便宜的にそれぞれ F1「関係性」、F2「わかりやすさ」、F6「新しさ」とよぶ。

### 5.2 面白さの予測結果と分析

予備調査の結果をもとに、3 つの特徴量を用いて大規模なクラウドソーシングを行った。各大喜利サイトからお題・回答ペアを各 543 個ランダムに抽出し、計 2,715 ペアを得た。これらに対して再びクラウドソーシングを用いた評点付けと回帰モデルによる予測精度の評価を行った。

予測の結果を表 4 に示す。特徴量単独での精度から、面白さとのつながりの深さは、上から順番に、F6「新しさ」、F2「わかりやすさ」、F1「関係性」と判断できる。比較のために、お題・回答ペアに対する面白さのスコアをデータセット全体における平均を常に出力するベースラインも示す。いずれの特徴量もこのベースラインよりも大幅に低い誤差を達成している。この結果から、大喜利の面白さがそれらの要素によって相当程度説明できていると結論付けられる。

## 6 面白さの自動推定に向けた検討

最後に、機械的に計量可能な特徴量を用いた面白さの自動推定を試みる。次の 3 つの機械的特徴量は、5 節で有効性が示された特徴量との関連性が期待され、

<sup>7</sup><https://crowdsourcing.yahoo.co.jp>

特徴量	平均二乗誤差
Embedding	0.2301
平均単語難易度	0.2302
LDA	0.2295
全て	0.2286
平均値ベースライン	0.2252

表 5: 機械的特徴量による面白さの予測精度

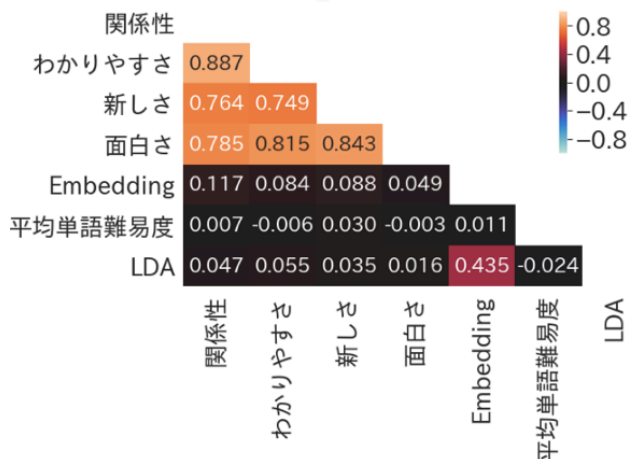


図 2: 特徴量間の相関関係

なおかつ現在の自然言語処理技術で比較的簡単に計量可能なものである。

**Embedding** お題と回答の文ベクトルのコサイン類似度。文ベクトルはテキスト中に出現する単語のベクトルの平均とする。単語ベクトルとしては、Web 文章 100 億ページから学習された 256 次元の word2vec モデルを用いる。F1「関係性」との関連性が期待される。

**平均単語難易度** 回答に含まれる単語の難易度の平均。単語難易度は、クラウドソーシングによって構築された約 26,000 語からなるデータベース [8] に基づく。F2「わかりやすさ」との関連性が期待される。

**LDA** お題と回答のトピックベクトルの類似度。トピックベクトルは、ブログコーパス 100 万文書で学習された 100 次元のオンライン LDA を用いて求める。F6「新しさ」に関連する、すなわちお題と回答間のトピックのずれを捉えることが期待される。

表 5 に機械的特徴量を用いた面白さの予測実験の結果を示す。Embedding、平均単語難易度、LDA、それら全ての組み合わせのいずれも平均値ベースラインよりも悪い予測精度を返した。特徴量間のピアソンの相関係数を図 2 に示す。いずれの機械的特徴量も、人手による特徴量あるいは面白さとは無相関に近いが、Embedding と F1「関係性」の相関係数は 0.117 と相対的に大きな値となった。平均単語難易度は他の特徴

量との相関係数が極めて小さく、LDA は期待された F6「新しさ」よりも F2「わかりやすさ」との間の相関係数の方が相対的に高くなった。これらの結果は、面白さの自動推定の手がかりとして Embedding が相対的に有効であるものの、全体として、現時点で得られる手がかりの有効性はいずれも限定的であり、大幅な改善の余地があることを示唆している。

## 7 おわりに

本研究では、大喜利を対象として、面白さを構成する要素をクラウドソーシングにより評価付けし、回帰モデルを用いてその有効性を示した。実験結果は、大喜利の面白さが主に 3 つ要素（お題と回答の関係性、回答のわかりやすさ、回答に新しさや新鮮味）の組み合わせで説明できる可能性を示唆する。さらに、これらの特徴量を参考に機械的特徴量を設計し自動推定を試みた。現時点ではいずれの機械的特徴も有効性は限定的であったものの、定量化された面白さの構成要素は今後の研究の指針として役立つと考えている。

## 謝辞

大喜利データセット作成にあたってクローラを許可していただいた大喜利プラス管理者の水口様に厚く御礼申し上げます。

## 参考文献

- [1] Kim Binsted and Graeme Ritchie. Computational rules for generating punning riddles. *HUMOR-International Journal of Humor Research*, Vol. 10, No. 1, pp. 25–76, 1997.
- [2] Peng-Yu Chen and Von-Wun Soo. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Vol. 2, pp. 113–117, 2018.
- [3] Lizhen Liu, Donghai Zhang, and Wei Song. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2, pp. 586–591, 2018.
- [4] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2367–2376, 2015.
- [5] Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1650–1660, 2018.
- [6] キムビンステッド, 修滝澤. 日本語駄洒落なぞなぞ生成システム “BOKE”. *人工知能学会誌*, Vol. 13, No. 6, pp. 920–927, nov 1998.
- [7] 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.
- [8] 水谷勇介, 河原大輔, 黒橋禎夫. クラウドソーシングを用いた習得時期の想起質問に基づく単語難易度データベースの構築. *言語処理学会第 25 回年次大会 発表論文集*, 2019. (to appear).