単語の分散表現によるユーモアの解釈

三嶋 祐輝　　（指導教員　西村 俊二)

令和 2 年 1 月 24 日

An Interpretation of Humor Using Distributed Representation of Words

Yuki Mishima　(Academic Advisor　Shunji Nishimura)

概要：近年，人工知能開発技術の発展に伴い，対話型エージェントの普及が進んでいる．しかし，現行の対話型エージェントのほとんどが，無感情かつ人間味のないコミュニケーションを行っている．したがって，対話型エージェントの普及に伴い，ユーザがエージェントに対して親しみを感じることができず，エージェントがコミュニケーションの相手として成立しないという問題が存在する [3]．井上 [5] は，人間同士が互いに親しみを感じ，円滑な人間関係を築きあげるにはユーモアが必要であると述べている．つまり，エージェントが人の代わりに，ユーモアを持ってユーザと対話を行うことができれば，ユーザはエージェントに対して親しみを感じることができると考える．そこで本研究では，対話型エージェントにユーモアの定量的解釈を組み込むことで，ユーザが親しみを感じる対話型エージェントの設計を目的とする．本稿では，膨大なテキストコーパスを fastText に学習させて，単語を実数値ベクトルで表したモデルを作成する．その後，学習済みモデルを用いて漫才のボケ単語の分散表現を得る．この分散表現から，コサイン類似度を用いてツッコミの単語を導出する．最後に，アンケートを行うことで，提案手法によるツッコミの単語の有効性を評価する．

キーワード：fastText, 分散表現,MeCab, コサイン類似度

1.　緒言

1.1　背景と目的

近年，人工知能技術の開発の進展に伴い，対話型エージェントの開発が進んでいる [1][2]．対話型エージェントは，ユーザからの問いかけや要求に対して，適切な回答，動作をする機能およびサービスのことである．対話型エージェントは，スマートフォンのデジタルアシスタントやショッピングモール，駅案内など我々のさまざまな生活シーンへと浸透している．しかし，現行の対話型エージェントのほとんどが，無感情かつ人間味のないコミュニケーションを行っている．したがって，現在運営されている対話型エージェントは，ユーザがエージェントに対して親しみを感じることができず，エージェントがコミュニケーションの相手として成立しないという問題が存在する [3]．例えば，老人ホームに介護およびコミュニケーションのための対話型エージェントを要したロボットが導入されたとする．ロボットが堅苦しい話ばかりすると，老人との親和的関係を築きあげることは難しい．2016 年に行われた，介護分野におけるコミュニケーションロボットの活用に関する大規模実証試験で，要介護者の約 3 分の 1 に生活の改善効果が認められたと報告された [4]．改善項目の中でも，「コミュニケーション」は全体の約 5 分の 1 を占めている．老人にとってコミュニケーションロボットとの対話は心理的観点から重要であると考えられる．

井上は，人間同士が互いに親しみを感じ，円滑な人間関係を築きあげるにはユーモアが必要であり，ユーモアは場の緊張を解くことや，人間関係の修復を図ることに役立つと述べている [5]．つまり，エージェントが人の代わりに，ユーモアを持ってユーザと対話を行うことで，ユーザはエージェントに対して親しみを感じることができると考えられる．しかし，ユーモアをエージェントに組み込むためには，ユーモアを定量的に評価する必要がある．

本研究では，対話型エージェントにユーモアの定量的解釈を組み込むことで，ユーザが親しみを感じる対話型エージェントの設計を目的とする．本研究で扱うユーモアは「他者から面白いと判断され，楽しいという感情や笑い，微笑みを生み出すようなすべてのコミュニケーション」と定義する [6]．

人間が意図的にユーモアを生成するコンテンツとして，ギャグや落語，漫才などがある．その中でも漫才は，ボケ側の常識から少しずれた発言や，冗談に対してツッコミ側がボケ側の発言の間違いや，冗談を素早く指摘するという対話形式で展開される．漫才は対話形式でユーモアを扱えることから，ユーモアのある対話型エージェントを設計するために，漫才を用いることは有効な手法であると考える．そこで本研究では，対話型エージェントにユーモアを定量的に理解させるために，漫才に焦点を

当て，ボケを表す単語とツッコミを表す単語の関係性を調べる．関係性を調べることにより，ツッコミの単語を導出する方法を提案し，実験により提案手法の有効性を検証する．

1.2 提案手法の概要

自然言語処理ライブラリである fastText を用いることにより，単語の分散表現を知ることができる．本研究では，fastText を用いて漫才におけるボケ単語の分散表現を得る．得られた分散表現に対してコサイン類似度を用いることにより，ツッコミ単語を導出する手法を提案する．図 1 に提案手法の流れを示す．具体的に，まず漫才は一般的に映像を媒体として提示されるため，本手法で用いるために文字起こしをする．その後，膨大なテキストコーパスを得るために，Wikipedia の日本語記事を形態素解析エンジンである MeCab[7] を用いて分かち書きする．その分かち書き済みデータを fastText に学習させ，単語を実数値ベクトルで表したモデルを生成する．学習済み fastText のモデルを用いて，文字起こしした漫才のボケ単語に該当する部分の単語の分散表現を得る．分散表現に対してコサイン類似度を用いることで，ボケ単語に最も近しい単語一覧を得る．得られた単語一覧の中から，ボケ単語に対してのツッコミ単語を導出する．導出されたツッコミ単語の有用性を評価するためにアンケートを行う．

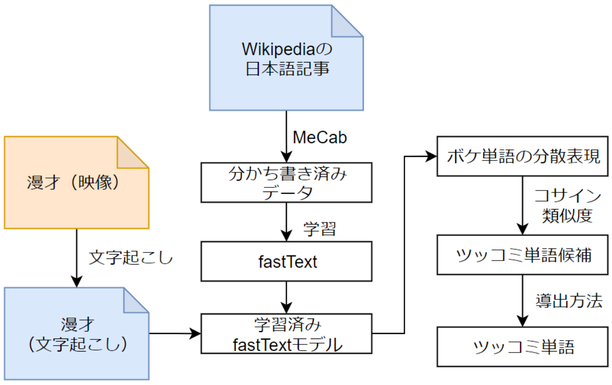

図 1　提案手法の流れ

1.3 関連研究

分散表現とユーモアに関する研究として呉らは，エージェントがユーモアを交えて対話を行うために，word2vec によって得られる分散表現からボケの生成を行った [3]．word2vec によって得られる単語の分散表現により，単語の意味を定量的に表すことができる．しかし，word2vec によって得られる分散表現には，単語の活用形を扱うことができない．そのため，word2vec では学習データに存在しない未知語に対して分散表現を得ることができないという問題がある．本研究では，呉らの提案手法をもとに，分散表現を得るために word2vec ではなく，fastText を用いることで分散表現の精度の向上を目指す．上記以外のユーモアに関する研究として伊藤は，不適合モデルと不適合—解決モデルの概念的な食い違いを整理し，両者を発展的に統合するモデルの提案と妥当性を示した [8]．本研究では，ユーモアを概念的に捉えるのではなく，分散表現を用いて，定量的に捉えるという点で異なる．ユーモアを定量的に捉えることにより，対話型エージェントにユーモアを定量的に解釈させることができ，ユーザからのボケに対して，的確なツッコミができると考える．

2. 理論

本章では分散表現，fastText および MeCab の概要を述べる．2.1 では分散表現の概要について述べる．2.2 では，fastText の概要および分散表現を得る方法について述べる．2.3 では，MeCab の概要について述べる．

2.1 分散表現

単語の分散表現とは，単語の意味を多次元の実数値ベクトルで表す方法であり，単語を定量的に示すことを目的としている．「単語の意味は，周囲の単語によって形成される」という分布仮説に基づいており，単語単体には意味がなく，その単語が文章中の存在位置によって単語に意味が与えられるという考えである．例えば，'I drink water' という文章と'We drink beer' という文章のように，'drink' という単語の近くには飲み物の意味を持つ単語が存在しやすいと考えられる．また，'I guzzle water' という文章と'We guzzle beer' という文章があるとき，'guzzle' という単語が'drink' という単語と同じような文脈で用いられることが分かる．

もし単語の意味に従って正しく単語をベクトル化できていた場合，その単語でなくても似た意味を持つ単語であれば，近い実数値ベクトルを持つこととなる．実数値ベクトル同士にコサイン類似度等の類似度計算を用いることで，ベクトル空間上における単語間の近似度を測定することが可能となる．式（1）にコサイン類似度の計算式を示す．単語の分散表現を応用することで，単語の分散表現をもとに文章のベクトル化が可能となり，文章間の類似度を測定することも可能となる．

$$sim(x, y) = \frac{\sum_{i=1}^{n} v_{x,i} \cdot v_{y,i}}{\sqrt{\sum_{i=1}^{n} v_{x,i}^2}\sqrt{\sum_{i=1}^{n} v_{y,i}^2}} \qquad (1)$$

- x，y：単語
- v_x，v_y：分散表現
- n：次元数

2.2 fastText

fastText[9] とは，Facebook AI Research が開発した自

然言語処理を高速かつ高精度に行うことができるライブラリである．fastText に膨大なテキストコーパスを学習させることにより，単語の集合が出力される．この出力された単語にはベクトルが付与されており，これを単語の分散表現と呼ぶ．fastText では，二つのモデルが使用されている．一つは，**図 2** にある周辺の単語から対象の単語を推測する Continuous Bag-of-Words モデル（以降，CBOW モデルと略記）．?に該当する単語を，周辺単語である’you’ と’goodbye’ から推測する．もう一つは，**図 3** にある対象の単語から周囲の複数ある単語を推測する skip-gram モデルである．?に該当する単語を，中心の単語である’say’ から推測する．fastText の根源となったものとして，Mikolov らによって提案された word2vec[10] がある．fastText では word2vec で考慮されていなかった部分語を扱うことができる．これにより，原形と活用形のような共通の部分を持つ単語に対して，その関係性を学習することができる．また，word2vec と比べて学習が高速であり，より高精度なモデルを生成することができる．fastText の学習パラメータの詳細を表 1 に示す．

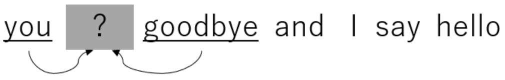

図 2 CBOW モデル

図 3 skip-gram モデル

表 1 fastText のパラメータ

-minCount	これより少ない出現回数の単語は無視する
-t	この値より出現率の高い単語は，その単語の出現率から計算された確率値に従って無視される
-lr	値が高いほど学習の速度が上がる
-dim	単語の分散表現の次元数
-epoch	学習回数，多すぎると過学習になる
-thread	学習時に使うスレッド数，CPU に応じて設定

2.3 MeCab

MeCab とは，オープンソース形態素解析エンジンであり，言語や辞書，コーパスに依存しない汎用的なものである．リスト 1 のように日本語のテキストコーパスを，語彙ごとに分かち書き，品詞の分類を行うことができる．リスト 1 に”私はペンを持っています”という文章に対して，分かち書きを行ったものを示す．
またリスト 2 に，リスト 1 と同様の文章に対して，品詞の分類を行ったものを示す．

リスト 1. MeCab を用いた分かち書き

```
私 は ペン を 持っ て い ます
```

リスト 2. MeCab を用いた品詞の分類

```
私 名詞, 代名詞, 一般,*,*,*, 私, ワタシ, ワタシ
は 助詞, 係助詞,*,*,*,*, は, ハ, ワ
ペン 名詞, 一般,*,*,*,*, ペン, ペン, ペン
を 助詞, 格助詞, 一般,*,*,*, を, ヲ, ヲ
持っ 動詞, 自立,*,*, 五段・タ行, 連用タ接続, 持つ, モッ, モッ
て 助詞, 接続助詞,*,*,*,*, て, テ, テ
い 動詞, 非自立,*,*, 一段, 連用形, いる, イ, イ
ます 助動詞,*,*,*, 特殊・マス, 基本形, ます, マス, マス
```

3. 提案手法

本章では提案手法の概要およびその手順について述べる．3.1 ではテキストコーパスの詳細，データの整形方法および fastText のモデルの作成方法を述べる．3.2 では漫才におけるツッコミとボケの役割およびボケ箇所の考え方について述べる．3.3 では漫才に応じて，ツッコミ単語を導出する方法を述べる．

3.1 fastText のモデル作成

膨大なテキストコーパスを得るために Wikipedia の日本語記事を用いる．ネット上の Wikipedia の日本語記事をそのまま fastText の学習に用いることができないため，分かち書きを行う必要がある．MeCab を用いて記事を分かち書きし，分かち書き済みテキストコーパスを fastText に学習させる．本研究では，新語に対応させるために MeCab に IPA-Neologd 辞書を用いる．2.2 節におけるパラメータは，dim を 300，epoch を 10，minCount を 20 とし，他は Default の設定を用いる．fastText の学習済みモデルには，学習元となったテキストコーパス上に存在する単語の分散表現が出力され，このモデルを用いることによって任意の単語に対する分散表現を求めることができる．

3.2 ボケ箇所

漫才は，一名または複数名の演者で構成され，それぞれをボケとツッコミと呼ばれる役割に分けることができる．ボケは漫才の中でツッコミのきっかけとなるような面白いことを言う役割である．主に常識から少しずれたことや，冗談などを言う．ツッコミは，ボケの間違いや冗談を素早く指摘し，笑いどころを観客側に提供する役割である．ときにはボケに対して軽い平手や，胸のあたりを手の甲でたたいて指摘するボディーランゲージを行うこともある．提案手法では，実際の漫才における単語ツッコミのみに注目し，ツッコミまでの一連の流れを題材として扱う．本研究における単語ツッコミとは，ボケに対して名詞かつ単語で返すものとする．リスト 3 に単

リスト 3. 単語ツッコミの例

ボケ：「4 時 64 分 4 時 64 分」
ツッコミ：「パラレルワールド！」

リスト 4. C1 の例

ボケ：「彼女と別れて四十九日なんですよ」
ツッコミ：「法事！」

リスト 5. C2 の例

ボケ：「大阪府警の署長が到着しました」
ツッコミ：「署長や」
ボケ：「署長の前田敦子です」
ツッコミ：「一日署長！」

語ツッコミの例を示す．リスト 3 ではツッコミの下線部にあたる’ パラレルワールド’ が単語ツッコミに該当する．なお題材を選ぶ際，ボケ箇所を単語で表すのが難しいもの，文脈が複雑なもの，適切な単語が見当たらないものに関しては対象外とする．

ボケ箇所とはボケ側における，ツッコミを生成するもととなる概念のことである．ボケ箇所を単語で表したものをボケ単語と呼ぶ．単語でのツッコミは，このボケ箇所をもとに生成されると考える．提案手法において，ボケ単語は二種類存在すると仮定する．一つ目は，ボケの文章中の常識とは違うもの（以降，C1 と記述）である．具体例としてリスト 4 があげられる．リスト 4 では下線部にあたる’ 四十九日’ という単語が，ボケ箇所に該当している．常識的に，この部分には’ 三カ月’ や’ 一年’ といった単語が当てはまると考えられる．そのため，この’ 四十九日’ という単語が非常識を感じさせ，ツッコミを誘発させるものとなっている．

二つ目は，漫才の話題に依存し，かつ常識とは違うもの（以降，C2 と記述）である．具体例としてリスト 5 があげられる．リスト 4 との違いは，ボケの単語のみではツッコミを考えることができないため，それまでの文脈を考慮しなければならないという点である．リスト 5 では，ニュースという一連の話題の中で，ボケ箇所である’ 署長’ と’ 前田敦子（アイドル）’ という単語の組み合わせが聞き手側に非常識性を感じさせ，ツッコミを誘発させるものとなっている．

実際に用いた漫才一覧を表 2 に示す．漫才 1 から 3 までは C1 に該当し，漫才 4 から 6 までは C2 に該当すると考える．表 3 に各漫才に対するボケ単語一覧を示す．漫才 2 では，’ カスタネット’ と’ リコーダー’ という単語が，ツッコミを誘発させるものとなっている．ゆえに，ボケ単語は，’ カスタネット’ と’ リコーダー’ となる．漫才 3 では，ツッコミ側が歌っている際に，ボケ側が合いの手を入れている．この合いの手が国会のヤジのようであるため，本来の合いの手とのギャップが生じ，そのギャップが，ツッコミを誘発させるものとなっている．ゆえに，

表 2 実験に用いた漫才

漫才 1	ボケ：「彼女と別れて四十九日なんですよ] ツッコミ：「法事！」
漫才 2	ツッコミ：「なんか楽器できるん？」 ボケ：「一応カスタネットとリコーダー」 ツッコミ：「必修！」
漫才 3	ツッコミ：「駐車場のネコはアクビをしながら～♪」 ボケ：「何言ってんだ！」 ツッコミ：「今日も一日を過ごしてゆく～♪」 ボケ：「どういうことだ！」 ツッコミ：「何も変わらない穏やかな街並～♪」 ボケ：「もっと端的に話をしろ！」 ツッコミ：「野党！」
漫才 4	ボケ：「大阪府警の署長が到着しました」 ツッコミ：「署長や」 ボケ：「署長の前田敦子です」 ツッコミ：「一日署長！」
漫才 5	ボケ：「おばあさんが川で洗濯をしていると，さらさらとそうめんが流れてきました」 ツッコミ：「流しそうめん！」
漫才 6	ボケ：「花火が上がるぞ」 ツッコミ：「花火ええやん」 ボケ：小規模な花火を体で表現 ツッコミ：「線香花火！」

表 3 各漫才に対するボケ単語一覧

漫才 1	四十九日
漫才 2	カスタネット，リコーダー
漫才 3	国会，ヤジ
漫才 4	アイドル，署長
漫才 5	そうめん，水流
漫才 6	花火，小さい

ボケ単語は，ボケの一連の台詞を要約して’ 国会’ と’ ヤジ’ となる．漫才 5 では，昔話の桃太郎という話題の中で，本来ならば川から桃が流れてくるという場面を，そうめんが流れてくると表現している．これにより，本来の場面との不一致が生じ，この不一致がツッコミを誘発させるものとなっている．ゆえに，ボケ単語は，’ そうめん’ と川の流れを要約して’ 水流’ となる．漫才 6 では，遊園地という話題の中で，常識ならば打ち上げ花火を期待するところを，ボケ側がボディランゲージで小さな花火を表現する．このボディランゲージが，期待とのギャップが生じさせ，ツッコミを誘発させるものとなっている．ゆえに，ボケ単語は，’ 花火’ とボケ側のボディーランゲージを考慮した’ 小さい’ となる．

3.3 ツッコミ単語の導出

ツッコミ単語とは，表 2 において下線が引かれている箇所になる．ツッコミ単語を導出する提案手順は二種類ある．まず初めに，導出方法 1 の詳細を以下に示す．

(1) 3.1 節で作成した fastText のモデルを用いて，一つ

または二つのボケ単語の分散表現を得る．

(2) ボケ単語が一つの場合はそのまま，ボケ単語が二つの場合は，二つのボケ単語の分散表現を加算したものに対してコサイン類似度を用いることで，ボケ単語の類似単語を得る．このとき，本手法では名詞かつ単語でのツッコミを考えているため，類似単語は名詞かつ単語のみに限定する．

(3) 類似単語の中から，ボケ単語の同位概念に当たるものを人手で除く．

(4) 残った類似単語の中から，ポピュラーな単語のみを考慮するため，Google 検索件数の多いもの上位三つを残す．残った単語をツッコミ単語とする．

次に, 導出方法 2 の詳細を以下に示す．

(1) 3.1 節で作成した fastText のモデルを用いて，二つのボケ単語の分散表現を得る．

(2) 二つのボケ単語の分散表現同士を加算し一つにする（以降，加算ボケ単語と記述）．その後，加算ボケ単語の分散表現に対してコサイン類似度を用いることで，加算ボケ単語の類似単語を得る．このとき，本手法では名詞かつ単語でのツッコミを考えているため，類似単語は名詞かつ単語のみに限定する．

(3) 類似単語と，二つのボケ単語のそれぞれに対して，コサイン類似度を測り，得られた値の差の絶対値を，類似度単語に対する評価値とする．評価値（以降，eval と記述）の計算式を式（2）に示す．この eval の値が 0 に近いほど，類似単語が分散表現の空間上において，二つのボケ単語に近い存在ということになる．

(4) eval の小さい類似単語の上位三つをツッコミ単語とする．

$$eval = |sim(x, y) - sim(x, z)| \quad (2)$$

- x : 類似単語
- y，z : ボケ単語

導出方法 1 で同位概念を除く理由としては，ボケと同じ概念の単語が，ツッコミになるとは考えにくいからである．例えば，表 2 の漫才 2 において，ツッコミとして「オカリナ！」と言っても，ボケに対するツッコミとして成立しないと考える．導出方法 1 でポピュラーな単語を考慮する理由としては，漫才におけるツッコミは，誰にでも分かる単語でないとツッコミとして成り立たないからである．例えば，ツッコミとして専門用語を用いても，その手の専門家には通じるかもしれないが，一般の人には通じないと考える．

導出方法 2 でコサイン類似度の差を測る理由としては，二つのボケ単語に対する類似単語が，ボケ単語のどちらに近しいかを知るためである．仮に eval の値が大きいと，その類似単語がボケ単語のどちらかに片寄っているということになる．そうすると，二つのボケ単語に対するツッコミではなくなってしまうと考える．例として，表 2 における漫才 4 のツッコミ単語について考える．ボケ単語にあたる’アイドル’＋’署長’ の類似単語の一つとして，’バラエティーアイドル’ がある．’アイドル’ と’バラエティーアイドル’ のコサイン類似度は 0.767 に対して，’署長’ と’バラエティーアイドル’ のコサイン類似度は 0.155 である．ゆえに，二つのボケ単語の類似単語である’バラエティーアイドル’ は’アイドル’ に近い単語であるといえる．そのため，’バラエティーアイドル’ という単語は，二つのボケ単語に対するツッコミ単語として不適格であると考える．次に，先ほどの二つのボケ単語に対する類似単語のもう一つとして’婦警’ がある．’アイドル’ と’婦警’ のコサイン類似度は 0.384 に対して，’署長’ と’婦警’ のコサイン類似度は 0.407 である．これらのコサイン類似度が近い値を取ることから eval の値が小さくなり，その結果，’婦警’ という単語が二つのボケ単語に対して似た向きを取り，ツッコミ単語として的確であると考える．

4. 実験と考察

4.1 実験の目的

本研究は，ユーザがエージェントと対話をする際に，エージェント側がユーモアのあるツッコミをすることで，ユーザが親しみを持つことができるという考えのもと，ユーモアのある対話型エージェントの設計のための提案である．提案手法の有用性を評価するには，実際にユーザにツッコミの評価をしてもらうことが必要である．そこでアンケートを行い，分散表現によるユーモアとユーザの感性によるユーモアの違いを検証する．

4.2 実験の手順

表 2 における漫才 1，2 は，C1 に属しているため，導出方法 1 によりツッコミ単語を導出した．漫才 3 は C1 に属しているが，ボケ箇所が’国会のヤジ’ という二つの単語で表されており，ボケ単語が二つになる．そのため，二つに近い単語を知りたいということから導出方法 2 によりツッコミ単語を導出した．漫才 4，5，6 は，C2 に属しているため，導出方法 2 によりツッコミ単語を導出した．表 4 に導出されたツッコミ単語を示す．

漫才 1，4 に関しては実際の漫才のツッコミと同等のツッコミ単語を導出することができたが，それ以外の漫才に関しては，導出することができなかった．理由として，漫才 2 では，’カスタネット’ と’リコーダー’ の分散表現を加算したものに対する類似単語の上位 50 個が楽器の名称になり，導出方法 1 における同位概念を除く処理の際，類似単語がすべて除かれ，ツッコミ単語の候補が

無くなったためである．漫才 3 では，’国会’ と’ヤジ’ のどちらにも近い単語として，’大島理森’ と’バカヤロー解散’ と’政治とカネ’ が得られた．しかし，どの単語も実際の漫才におけるツッコミと同等なものとは考えられなかったため，ツッコミ単語として採用していない．漫才 5 では，’そうめん’ と’水流’ のどちらにも近い単語として’鮎’ と’ニゴロブナ’ と’味噌川ダム’ が得られた．しかし，これも漫才 3 と同じ理由でツッコミ単語として採用していない．漫才 6 では，’小さい’ と’花火’ のどちらにも近い単語として’かきくらべ’ と’紙火薬’ と’四尺’ が得られた．しかし，これも漫才 3 と同じ理由でツッコミ単語として採用していない．

表 4　導出されたツッコミ単語

	Top1	Top2	Top3
漫才 1	葬式	忌	法事
漫才 2	なし	なし	なし
漫才 3	大島理森	バカヤロー解散	政治とカネ
漫才 4	スケバン刑事	一日署長	婦警
漫才 5	鮎	ニゴロブナ	味噌川ダム
漫才 6	かきくらべ	紙火薬	四尺

アンケートでは，ツッコミ単語の導出に成功したもの（漫才 1，4）に関しては，実際の漫才の一連の流れ（本物のツッコミは隠してあるもの）と，提案手法によって導出されたツッコミ単語の候補を三つ選択肢として与え，どれが一番ツッコミ単語として成り立っているのか集計を取った．このとき，提案手法による導出過程から生じる先入観をなくすために，各ツッコミ単語の生成方法はアンケート回答者には伝えていない．ツッコミ単語の導出に失敗したと思われるもの（漫才 2，3，5，6）に関しては，導出方法 1，2 によって導出されたツッコミ単語を採用しないという判断の妥当性を検証するために，二つのボケ単語を提示し，その二つから連想される単語（以降，連想単語と記述）を複数個回答してもらった．

4.3　提案手法と実験結果の比較

本実験では，10 代 11 名，20 代 12 名，30 代 5 名，40 代 1 名の合計 29 名に対してアンケートを行った．

図 **4** は漫才 1 に関するアンケート結果である．’ボケに対して適切だと思われるツッコミを選んでください’ という質問に対して，法事が 76.7 %，忌が 16.7 %，葬式が 6.6 %という結果が得られた．

図 **5** は漫才 4 に関するアンケートの結果である．’ボケに対して適切だと思われるツッコミを選んでください’ という質問に対して，一日署長が 66.7 %，婦警が 26.7 %，スケバン刑事が 6.6 %であった．

表 5 に，漫才 2 に関するものを示す．分散表現によるツッコミ単語では，楽器およびその名称ばかりになり，同位概念を考慮した場合，ツッコミ単語となるものはなかった．

アンケートによる連想単語では，’小学校’ や’小学校の音楽’，’小学生’ など，ボケ箇所と異なる概念のもの上位となった．

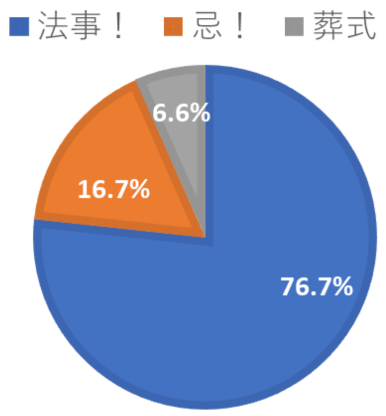

図 4　漫才 1 に関するアンケート結果

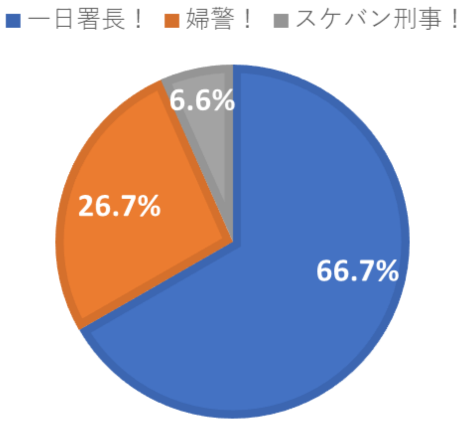

図 5　漫才 4 に関するアンケート結果

表 5　導出された単語とアンケートによる連想単語

	導出されたツッコミ単語	連想単語
1	なし	楽器
2	なし	小学校
3	なし	小学生

表 6　導出された単語とアンケートによる連想単語

	導出されたツッコミ単語	連想単語
1	大島理森	野党
2	バカヤロー解散	安倍晋三
3	政治とカネ	政治

表 6 に，漫才 3 に関するものを示す．分散表現によるツッコミ単語の上位が，固有名詞である’大島理森’，歴史上の出来事かつポピュラーでない’バカヤロー解散’，名詞句である’政治とカネ’ となったため，どれもツッコミ単語として成り立たないと考えた．

アンケートによる連想単語では，' 野党'，' 安倍晋三'，' 政治' が上位となった．一番回答数の多かった' 野党' は，実際の漫才で用いられている単語である．

表 7 導出された単語とアンケートによる連想単語

	導出されたツッコミ単語	連想単語
1	鮎	流しそうめん
2	ニゴロブナ	夏
3	味噌川ダム	水

表 7 に，漫才 5 に関するものを示す．分散表現によるツッコミ単語の上位が，魚の名称である' 鮎' と' ニゴロブナ'，ダムの名称である' 味噌川ダム' となったため，どれもツッコミとして成り立たないと考えた．

アンケートによる連想単語では，' 流しそうめん'，' 夏'，' 水' が上位となった．一番回答数の多かった' 流しそうめん' は，実際の漫才で用いられている単語である．

表 8 導出された単語とアンケートによる連想単語

	導出されたツッコミ単語	連想単語
1	かきくらべ	線香花火
2	紙火薬	手持ち花火
3	四尺	失敗した花火

表 8 に，漫才 6 に関するものを示す．分散表現によるツッコミ単語の上位が，祭りの名称である' 紙火薬'，花火の一種である' 紙火薬'，打ち上げ花火の玉の一種である' 四尺' となったため，どれもツッコミとして成り立たないと考えた．

アンケートによる連想単語では，' 線香花火'，' 手持ち花火'，' 失敗した花火' が上位となった．一番回答数の多かった' 線香花火' は，実際の漫才で用いられている単語である．

4.4 提案手法と実験結果の考察

4.4.1 導出できたケース

図 4 について，提案手法によって導出されたツッコミ単語である' 法事' が，アンケート結果において 76.7 %であったことから，導出方法 1 はこのボケ箇所に対して有効であるといえる．

図 5 について，提案手法によって導出されたツッコミ単語である' 一日署長' が，アンケート結果において 66.7 %であったことから，導出方法 2 はこのボケ箇所に対して有効であるといえる．

4.4.2 人の連想からは導出できたケース

表 6 について，提案手法による導出方法では，ツッコミとして有効な単語を導出することができなかった．連想単語には，実際の漫才で用いられている' 野党' が一番多く回答された．' 野党' という単語は，ボケ単語に対する類似単語に表れていたが，導出方法 1 の過程で除かれてしまった．

表 7 について，提案手法による導出方法 2 で，' 鮎' という単語が現れるのは，鮎そうめんという食べ物が存在するからであると考える．その結果，提案手法でツッコミとして有効な単語を導出することができなかった．連想単語には，実際の漫才で用いられている' 流しそうめん' が一番多く回答された．' 流しそうめん' という単語は，ボケ単語に対する類似単語に表れていたが，漫才 3 と同様，導出過程において除かれてしまった．

表 8 について，提案手法による導出方法では，ツッコミとして有効な単語を導出することができなかった．また，ボケ単語に対する類似単語として，' 大きい'，' 小さく'，' 小さな' 等の形容詞が多くみられた．これらは，ボケ単語の一つである' 小さい' に分散表現が近い値をとるものだと考えられる．このことから，視覚的要素であるボディーランゲージをボケ箇所として扱う場合，形容詞ではなく名詞で表す必要があると考える．連想単語には，実際の漫才で用いられている' 線香花火' が一番多く回答された．' 線香花火' という単語は，ボケ単語に対する類似単語に表れていたが，漫才 3 と同様，導出過程において除かれてしまった．

以上の結果を踏まえて，人の連想からツッコミを導出する方法は正しいといえるが，提案方法には改善が必要であると考えられる．

4.4.3 人の連想からも導出できなかったケース

表 5 について，連想単語である' 楽器' は，ボケ箇所に対する類似単語でも現れていた．その点，' 小学校' や' 小学生' は類似単語に現れていなかった．' カスタネット' と' リコーダー' は小学校教育において，必修なものである．このことを踏まえると，連想単語の項目に' 小学校' 等の単語が現れるのは理解できる．実際の漫才では，ツッコミ側が「なんか楽器できるん？」と質問をする．これは，経験したことのある楽器を教えて，というニュアンスの質問である．しかし，今回のアンケートでは回答者に対して，そのようなニュアンスは伝えていない．もし，アンケートの質問に"やったことのある"という文脈込みで単語を連想してもらうことができていれば，連想単語に' 必修' という単語が現れていたと考える．また，表 5 の結果から，人間が連想する単語と分散表現で導出される単語には違いがあることがわかる．そのため，人間は上位概念と同位概念の区別ができるのに対して，分散表現では区別ができない．これは，「単語の意味は，周囲の単語によって形成される」という分散表現のアイデアによる影響であると考える．

5. 今後の課題

本研究の今後の課題について述べる．

(1) Facebook社がfastTextの後継として公開した自然言語処理ライブラリであるStarSpace[11]を用いて，単語の分散表現を獲得する．その後，Sparse Composite Document Vectors[12]を用いて，単語を複数のクラスタに分類し，分散表現に単語のidf特徴を掛け合わせ，単語の分散表現の精度向上をさせる．これにより，ボケ単語に対する類似単語の精度改善を図る．
(2) 時系列データを扱うことのできるRecurrent Neural Network（RNN）を用いて，漫才の一連の文脈を考慮することによって，扱える漫才の文脈の数を増やし，漫才単位でのボケとツッコミの関係性を調べる方法を提案する．
(3) 日本語の概念辞書である，単語の同義語，上位・下位語などがまとめられているWordNet[13]を用いて，分散表現から同位概念の除去を行い，ツッコミ単語の精度向上を図る．
(4) 文章データの分散表現を獲得することができるDoc2Vec[14]を用いて，文章単位でツッコミの導出をする手法を検討する．

6. 結言

本研究では，ユーモアの定量的解釈を目標として，単語の分散表現を用いて実際の漫才におけるボケ箇所からツッコミ単語を導出する方法を提案した．その結果，漫才によってはツッコミを導出できるケース（4.4.1節）もあったが，導出できないケース（4.4.2節と4.4.3節）もあった．漫才のボケとツッコミにおいて，概念レベルを区別することは必須かつ重要であると考えるが，現状，単語の分散表現から概念レベルを区別する方法がない．ツッコミ単語が導出できたケース（4.4.1節）に関しては，ツッコミ候補の単語を三種類の中から選択させるアンケートを行い，提案した手法の有用性を評価した．図4と図5の結果より，単語の分散表現をツッコミ単語生成に用いることは，一部の漫才に対して有効であった．ツッコミ単語の導出ができないケース（4.4.2節と4.4.3節）に関しては，単語の自由記述式アンケートを実施することにより，単語の分散表現と人間の連想力の違いを評価した．表5，6，7，8の結果より，提案した手法によって定量的に導出された単語と人間の連想によって導出される単語には大きな違いがあることを確認した．

7. 謝辞

本研究に際して，さまざまなご指導を頂きました西村俊二講師に深謝いたします．また，この研究の機会をくださった情報工学科の先生方，そして多くの知識やご指導をくださいました同研究室の先輩・同期の皆様，アンケートにご回答くださいました皆様に厚く御礼申し上げます．

参考文献

[1] 田中剛, 角森唯子, 門畑祥子, 藤本拓:”自然言語プラットフォームによるマルチエージェント対話システムの実現”,SIG-SLUD,5(02),2016
[2] 鈴木天詩, 宮田章裕:”認知症の理解を支援する対話型エージェントの基礎検討”,IPSJ SIG Technical Report,2017-GN-102 No.9,2017
[3] 呉建朗, 中原涼太, 長岡大二, 中辻真, 宮田章裕:”ボケて返す対話型エージェント”,TVRSJ,Vol.23 No4 pp.231-238,2018
[4] 大川弥生:”介護分野におけるコミュニケーションロボットの活用に関する大規模実証試験報告書”, 国立研究開発法人,2016
[5] 井上:”笑い学”研究について, 笑い学研究,No.9,pp.3-5,2002
[6] 塚脇涼太, 深田博己, 樋口匡貴:”ユーモア表出が表出者自身の不安および抑うるに及ぶ影響過程”, 実験社会心理学研究 51(1), 43-51, 2011-09
[7] T.Kudo, K.Yamamoto, Y.Matsumoto: “Applying Conditional Random Fields to Japanese Morphological Analysis” Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004),p.230-237,2004
[8] 伊藤大幸:”ユーモアの生成過程における論理的不適合および構造的不適合の役割”, 認知科学 17(2),297-312,2010
[9] Piotr Bojanowski,Edouard Grave,Armand Joulin,Tomas Mikolov:”Enriching Word Vectors with Subword Information” Transactions of the Association for Computationinal Linguistics,vol.5,pp.135-146,2017
[10] Mikolov,Tomas, et al :”Efficient estimation of word representations in vector space.” arXiv preprint arXiv:1301.3781,2013
[11] Ledell Wu,Adam Fisch,Sumit Chopra,Keith Adams,Antoine Bordes and Jason Weston,:”StarSpace:Embed All The Things!” The Thirty-Second AAAI Conference on Artificial Intelligence.AAAI18 - NLP and Machine Learning,2018
[12] Dheeraj Mekala,Vivek Gupta,Bhargavi Paranjape,Harish Karnick,:”SCDV:Sparse Composite Document Vectors using soft clustering over distributional representations”Proceedings of the 2017 Conderence on Empirical Methods in Natural Language Processing.pages 659-669,Copenhagen,Denmark,September 7-11,2017
[13] Hitoshi Ishihara,Francis Bond,Kiyotaka Uchimoto,Masao Uchiyama and Kyoko Kanzaki:”Development of Japanese WordNet”,in LREC-2008
[14] T.Mikolov,I.Sutukever,K.Chen,G.S.Corrado,J.Dean, ”Deistributed Representations of Words and Phrases and their Compositionality”,Advances in neural information processing systems,2013