ISSN 2186-5647

5-9

Word2Vec の類似度平均化による“あるある文章”の共感度判定の検討

日大生産工（院）　○野口　啓太
日大生産工　豊谷　純，大前　佑斗

1. はじめに

Apple 社の Siri[1)]や Amazon 社の Alexa[2)]のような対話システムは，現代社会において広く浸透し始めている．それらの例として，海外では不動産売買や，体重体組成計に対話システムを活用する試みがなされている[3),4)]．今後このような対話システムは，長期的かつ継続的に利用され続けることが求められる．その際，人間が親近感を抱くようなデザインを，対話システムにもたせることは重要であるといえる．

親近感を抱かせる要素の一つとして，ユーモアがあげられ，人同士の場合，親和的な関係構築には“笑い”の要素が不可欠であるとされる[5)]．さらに，人と対話システムの場面においてもユーモアの要素は対話を継続させたいと思わせる効果があると報告されている[6)]．

したがって，ユーモアを定量的に表現することは，ユーモア要素のある対話システムを実現する上で，非常に重要であるといえる．一方，ユーモアはいくつかの種類に分類される[7)]．そのひとつに“あるある”がある．あるあるとは，「日常生活における事象・見聞などで，多くの人の共感を得ることのできる話題．また，その話題で笑いをとる演芸のこと」（実用日本語表現辞典）とされている．つまり，あるあるにはユーモアの側面と多くの人の共感を促す側面があり，ユーモアの要素を持つ対話システムの実現において有効であるといえる．よって本研究では，上記のような特徴を持つあるあるを活用するため，定量的に判定することを目的とする．

そこで本研究では，Word2Vec を用いて単語同士の類似度からあるある文章の共感度を測定する手法を検討する．

2. “あるある”に関する考察

“あるある”の具体例は以下のようなものが一般的である．

・スプーンを洗おうとすると水がかかる
・爪切りや耳掻きが行方不明になる
・母が電話するとき時声が高くなる
・借りたボールペンの書き心地に感動する

上記のような“あるある”の特徴の一つに，「過去の日常生活において，何回か経験したことがある事柄」というものがあると考えられる．そのため，「食器を洗った後は手を拭く」「電車に乗って席に座る」といった毎回経験し当たり前であることや，「電車の中でハードル走をする」「電話をしながらラジオ体操をする」といった日常生活で起こりえないことは，“あるある”にはならないと言える．つまり，“あるある”になるような適切な範囲が存在すると言える．

そこで本研究では，その範囲を定量的に判定するために“あるある”の文章（以下，あるある文章とする）が示す場面の発生頻度を，0～4 の 5 段階で表現した．また，そのような値を本研究では共感度とした．

3. Word2Vec

本研究で用いる Word2Vec とは，Tomas

Verification using the average of similarity in Word2Vec to determine the empathy of "Common topics"

Keita NOGUCHI, Jun TOYOTANI and Yuto OOMAE

Mikolov らによって提案された，自然言語処理に利用されるニューラルネットワークモデルの一つである[8]. Word2Vec では，単語の分散表現を作成し，単語同士のベクトル演算を可能にする．それにより，単語と単語の類似度の算出や，単語から単語の足し算や引き算が可能になる．さらに，Word2Vec は「単語の意味はその周辺語の単語によって形成される」という分布仮設というアイデアに基づいている．つまり，類似度が高いという結果が出た単語同士は，単語同士を入れ替えたとしても文章が成り立つ関係と捉えることができ，意味的に近いとわかる．

3．1　類似度を関連度の指標

分布仮説のイメージとして，図 1 のような例が挙げられる．

図 1 の場合，電車と新幹線という単語のそれぞれの周辺語には，新大阪や東京，駅といった単語がある．つまり，電車と新幹線という単語は，この場合似た意味になるのではないかと予測される．

一方，図 2 に示すように“電車”と“線路“というように意味的には異なるが，関連性が高い関係の単語同士も存在する．

このような場合も，ある程度周辺語は似てくると考えられる．つまり，関連性が低い単語同士と比べて，関連性が高い単語同士は類似度が高くなると考えられる．よって，Word2Vec における単語同士の類似度は関連度としても用いることが可能であると考えられる．

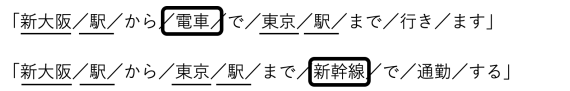

図 1　分布仮説による類似度のイメージ

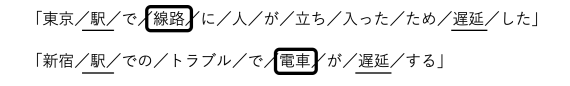

図 2　類似度を関連度として捉える場合

上記より，あるある文章内の名詞や動詞といった単語を抽出し，類似度を求めることで文章の内容に対する共感度を予測する指標になると言える．つまり，単語同士の類似度が全体的に低い文章は，単語同士の関連性が低い「電車の中でハードル走をする」「電話をしながらラジオ体操をする」といった日常では起こりえない内容の文章になると予想される．また，単語間の類似度が全体的に高い文章は，単語同士の関連性が高い「食器を洗った後は手を拭く」「電車に乗って席に座る」といったあたり前の内容の文章になると考えられる．

よって本研究では，Word2Vec を用いて単語同士の類似度の平均値を，関連度と捉え，共感度を予測する特徴量として用いた．

4.　提案手法

本研究で用いたあるある文章は，SNS の Twitter にて#あるあると検索をかけて取得した．また，あるあるといっても種類は多岐にわたるため，身近な話題である交通機関に関するあるあるに限定し，59 個のあるあるを対象にした．さらに用いた Word2Vec における，ニューラルネットワーク学習済みモデルは，単語数が 211673 であり，次元数が 300 次元であったものを使用した[9]．

Word2Vec では，先の説明通り単語を用いて類似度などの値を求める．そのため，あるある文章中から単語を抽出しなければならない．この時，抽出した単語が文章内容を反映したものではなかった場合，出力結果も文章内容を反映したものではなくなってしまう．そのため単語の並びのみで，文章内容を表現できる単語を抽出する必要がある．一方で，抽出する単語数が多いと処理の負担が大きくなってしまう．そのため，抽出する単語数は必要最低限の数に抑える必要がある．そこで本研究では図 3 に示す，文章中における「どこで」「なにを」「どうした」にあたる 3 単語を抽出した．その後，図 4 のように各単語間の類似度の平均値を，Word2Vec を用いて求めた．

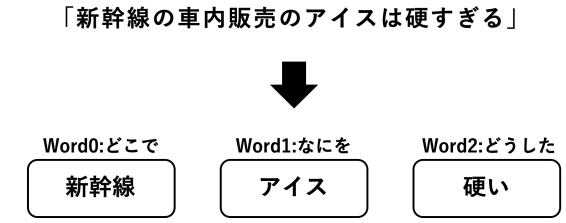

図 3　単語の抽出の例

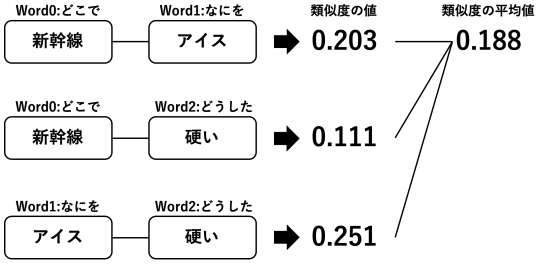

図 4　類似度の平均値算出の流れ

さらに，出力で予測する共感度として，各あるある文章に対して『0:まったくない』『1:ごくまれにある』『2:たまにある』『3:よくある』『4:いつもある』の 5 段階で評価した．この際，『0:まったくない』や『4:いつもある』に該当する文章が上記の取得方法では得られない．そのため『0:まったくない』に該当する「タクシーで木星まで行く」や，『4:いつもある』に該当する「タクシーはドライバーが運転する」といったオリジナルの文章を作成した．さらに評価は筆者が行い，研究室の学生らの意見を取り入れて一般的なものになるように調整した．

最後に上記の類似度の平均値から，共感感度の値を予測する．そのため，類似度の平均値を説明変数 x とし，共感度を目的変数 y として単回帰分析をした．

5.　結果

単回帰分析の結果，重相関係数 R は 0.567 となり相関が確認された．また，説明変数にかかる係数は，有意水準 1%で有意であった．回帰直線の切片は-0.293 であり，説明変数の係数は 13.26 であった．以上より，回帰式は以下のようになる．

$$\mathbf{y = 13.26x - 0.293} \qquad (1)$$

さらに，人が評価した共感度と類似度の平均値の相関を示すグラフは図 5 ようになった．

また，入力したあるある文章から共感度の判定をし，あるある度の振り分けまでの流れを図 6 にフローチャートで表した．あるある度の振分けは，『まったくない』『いつもない』といった，あるあるではないと出力された文章以外を対象にした．つまり，出力された共感度が 1 以上かつ 3 以下の場合あるあると判定し，1 より小さい場合はあたりまえと判定するものとした．

実際にあるある文章を判定した例を表 1．に示す．あるあるもしくは，あたりまえと判定される文章はあったが，ありえないと判定される文章は本研究に用いたデータ内にはなかった．これは表 1 の例の場合，“タクシー”と“ナス”の類似度は低くなるが，“ナス”と“育て”の類似度が高くなり平均値に影響してしまうためだと考えられる．そのため，ありえないと判定する，より適切な範囲を求める必要がある．

表 1　共感度判定とあるある判定の結果

あるある文章	類似度の平均値	共感度の評価値	共感度の予測値	あるある判定
新幹線のリクライニングを倒すときに緊張する	0.192	3	2.25	あるある
タクシーの匂いがきつい	0.209	3	2.47	あるある
タクシーでナスを育てる	0.104	0	1.08	あるある
駅でハイエナにぶつかる	0.123	0	1.34	あるある
電車で席に座る	0.317	4	3.91	あたりまえ
タクシーはドライバーが運転する	0.385	4	4.81	あたりまえ

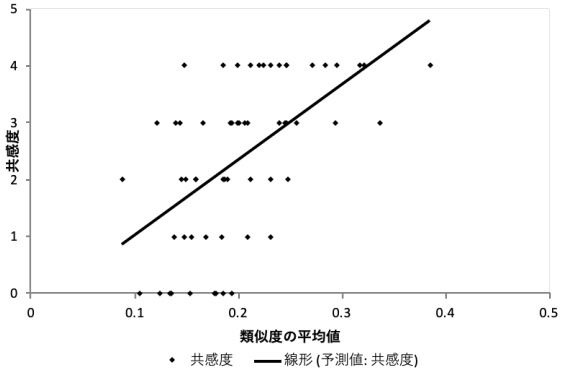

図 5　共感度と類似度の平均値の相関グラフ

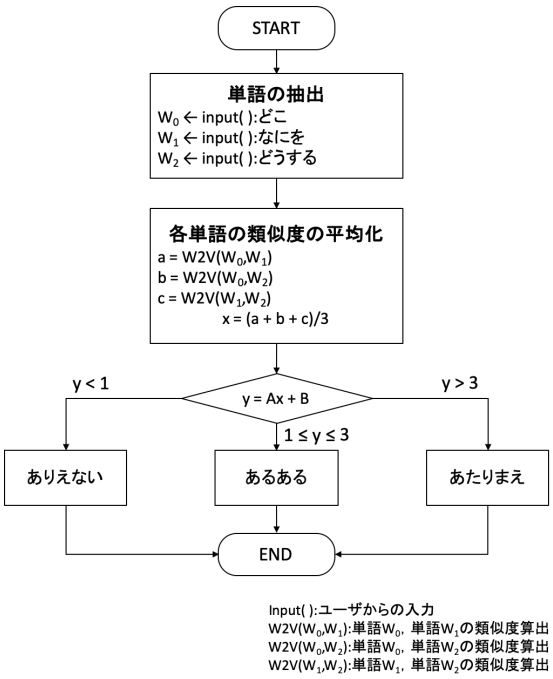

図 6　あるある度の振り分けまでの流れ

6. まとめ

本研究では，計算機によるユーモアの定量的な表現の実現のため，ユーモアの一つのである“あるある”を対象にその共感度とあるある度を予測し判定する指標を提案，検証した．

その結果，類似度の平均値と人が評価するあるある文章に対する共感度は相関関係であると明らかになった．これは，人々が日常で経験する事象の頻度が高いということは，それらの事象を示す文章が学習用のデータ内に多く存在することになり，その文章に登場する単語同士の結びつきが強くなるためだと考えられる．

今後の課題として，本研究では対象の文章が“あるある”かそうでないかの判定をすることは可能だが，ユーモアがあるかないかを判定することはできない．そのため，“あるある”と判定された文章に対してユーモア度を判定する手法を研究していく．

7. 参考文献

1) Siri，https://www.apple.com/jp/siri/，（参照:2020/10/07）

2) Alexa とできること，https://www.amazon.co.jp/meet-alexa/b?ie=UTF8&node=5485773051，（参考:2020/10/07）

3) T. W. Bickmore．J. Cassell.，“Relational agents: a model and implementation of building user trust”，In Proc. CHI, (2001) pp. 396-403.

4) T. W. Bickmore．R. W. Picard.，“Establishing and maintaining long-term human-computer relationships.”, ACM Transactions on Computer-Human Interaction (TOCHI), Vol.12 No.2 (2005) pp.293–327

5) 井上宏，「笑い学」研究について，笑い学研究，No. 9 (2002) pp. 3-15

6) 宮澤幸希，常世徹，桝井祐介，松尾智信，菊池英明，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，Vol.J-95-A No.1 (2012) pp.27-36

7) 上野行良，ユーモア現象に関する諸研究とユーモアの分類化について，社会心理学研究，Vol.7 No.2 (1992) pp.112-120

8) T. Mikolov. I. Sutskever. K. Chen. G. Corrado. J. Dean., “Distributed representations of words and phrases and their compositionality”. In NIPS, (2014).

9) Qiita,(2017),https://qiita.com/Hironsan/items/513b9f93752ecee9e670 ,(参照：2020/10/07)