BERT ベクトルを用いたオノマトペ由来の新動詞の検出

古宮嘉那子 [1] 宇野良子 [1] 浅原 正幸 [2]
[1] 東京農工大学　[2] 国立国語研究所/総合研究大学院大学
{kkomiya@go, ryokouno@cc}.tuat.ac.jp　masayu-a@ninjal.ac.jp

概要

本稿では、BERT ベクトルを用いてオノマトペ候補の分析を行う。国語研日本語ウェブコーパスから 2 モーラが繰り返される ABAB 型のオノマトペ候補を収集し、クラウドソーシングによる調査で分析対象のオノマトペ候補を 64 種類にしぼった。これらのオノマトペ候補の zero-shot の BERT ベクトルを、「ABAB」、「ABAB する」、「AB る」という三種類について主成分分析を用いて二次元上にプロットした。「ABAB」、「ABAB する」、「AB る」のクラスタ間距離を目視で遠い/近いに分けて分析し、これらのクラスタ間距離によるオノマトペ候補の分類を用いて、オノマトペ由来の新動詞の検出を試みた。

1　はじめに

「プンプン」のような 2 モーラ繰り返し型（ABAB 型）のオノマトペは日本語において最も一般的な形のひとつである。このタイプのオノマトペは「する」を伴った動詞用法を持つことが多い。また、「AB る」型の新語の動詞（新動詞）となることもある。新語とは言語集団がまだ慣習化していないとみなす語である [1]。例えば、「モフる」という新動詞は「モフモフ」「モフモフする」に由来する [2, 3]。しかし一方で、「ナデナデ」は ABAB 型だがオノマトペではなく、「撫でる」という動詞がもとになっていると思われ、オノマトペ由来の新語動詞の自動検出は容易ではない。本研究では、BERT[4] の出力ベクトルを用いて ABAB 型のオノマトペ候補を分析する。国語研日本語ウェブコーパスから ABAB 型のオノマトペ候補を機械的に収集し（4.1 節）、クラウドソーシングによる調査で分析対象のオノマトペ候補を 64 種類にしぼった（4.2 節）。これらの zero-shot の BERT ベクトルを、「ABAB」、「AB る」、「ABAB する」という三種類について主成分分析を用いて二次元上にプロットした（4.3 節）。「ABAB」、「ABAB する」、「AB る」のクラスタ間距離を目視で遠い/近いに分けてオノマトペ候補を分類し（4.4 節）、これらのクラスタ間距離によるオノマトペ候補の分類を用いて、オノマトペ由来の新動詞の検出を試みた。また、言語学的知見からこれらの分類の分析を行った（5 節）。

2　参考文献

Komiya ら [5] と Komiya ら [6] はそれぞれオノマトペの現れる文脈上の出現語を素性にしたベクトルまたはオノマトペの Word2Vec を用いてオノマトペのクラスタリングを行っている。また、乙武ら [7] は BERT ベクトルを用いてオノマトペの語義を分類する手法を提案している。BERT ベクトルを用いて日本語の意味について分析した論文にはほかに小林ら [8] などがある。また本研究の派生的な研究として、日本語の母語話者が、未知の単語の表層から「オノマトペらしさ」を評価できるかという実験を行った Uno ら [9] やクラウドソーシングによるオノマトペの類像性を調査することで、オノマトペ由来の新語抽出を試みた宇野ら [10] がある。

3　BERT ベクトルを用いたオノマトペ由来の新動詞の検出

本研究はオノマトペ由来の新動詞を検出することを目的としている。言語学では、少数の単語について人手で分析を行うことが一般的であるが、本研究では、自然言語処理の技術を用いて比較的多数のオノマトペの候補についての分析を行う。新語を研究対象とするため、対象オノマトペは、既存の辞書を利用せず、コーパスとクラウドソーシングを用いて選定した。また、検出に際し、アノテーションデータが存在しないため、教師なしの枠組みを用いる。そのため、BERT ベクトルを文脈ベクトルとして用いてオノマトペ候補を分類し、どのような特徴がオノマトペ由来の新語に相当するかを、分類をもとに読み解くことで分析するという手順を用いた。この際、恣意的な分類にならないよう、分類は第一著者が行い、分類後の分析を第二著者と第三著者が

行った。

4 実験

4.1 ウェブコーパスからのオノマトペ候補の抽出

2014 年 10 月から 12 月までに収集された『国語研日本語ウェブコーパス』[11] 25,836,947,421 語（国語研短単位）から、正規表現を用いて異なり 5,882 語の「ABAB」型のオノマトペ候補を抽出した。さらに可能な活用形を考慮したうえで「ABAB する」の頻度が、「AB る」の頻度よりも高い表現「ABAB」異なり語数 844 語（のべ 250 億語に相当）をクラウドソーシングの調査対象とした。具体的には ABAB にサ行変格活用の活用形展開したものを後置させた文字列の頻度の合計と、AB にラ行五段活用の活用形展開したものを後置した文字列の頻度の合計を評価した。また、オノマトペ候補は片仮名表記のものに限った。

4.2 クラウドソーシングよる調査

上記 844 語について Yahoo! クラウドソーシングを用いてアンケート調査を実施した。本稿では 4 種類実施したうちの本研究に主にかかわる 2 種類について示す。残りは宇野ら [10] を参照されたい。

（調査 1）「ABAB」型のうち、オノマトペであるものを抽出するために行った。「ABAB」型の 844 語を対象とし、「オノマトペか」について、0 から 5 の評定情報を各表現 20 人分収集した。調査は 2022 年 10 月 1 日 23:00-23:55 に実施し、異なり 674 人が参加した。

（調査 2）「ABAB する」由来の「AB る」を特定するために行った。844 語の「ABAB」に対応する、「ABAB する」と「AB る」の 2 つの表現対を対象とし、「【AB る】は【ABAB する】をもとにした語であるか」について、0 から 5 の評定情報を各表現 20 人分収集した。調査は 2022 年 10 月 1 日 23:00 より 10 月 2 日 00:10 まで実施し、異なり 667 人が参加した。

（調査 1）において、20%以上の人が「オノマトペである」と認識している 64 語を本研究の対象のオノマトペ候補と定義する。「ABAB」の「AB」部分を列挙すると、ドナ，ゴモ，ホエ，シマ，ホレ，クサ，シバ，ボケ，ツメ，テレ，グス，ネム，ボツ，ネジ，マゼ，モエ，ムレ，ウツ，シブ，ノビ，カス，カジ，ブレ，パシ，ウダ，アゲ，シコ，ヌメ，ヨワ，ウネ，チビ，ニギ，グズ，コネ，ドヤ，アセ，バテ，テカ，サワ，ジワ，ナデ，ヨタ，ネバ，ガチ，ヒエ，シナ，モジ，ガミ，ビク，ヨレ，イキ，ヒタ，チク，グネ，シク，モヤ，スベ，クネ，ゴワ，ヘナ，モフ，ホジ，ボコ，モグとなる（オノマトペらしさの評点の低い順（少数点以下 3 桁目を四捨五入））。

4.3 オノマトペ候補の BERT ベクトルの描画

オノマトペ候補 64 語について『国語研日本語ウェブコーパス』から収集した例文スニペットについて、BERT base Japanese[1] の zero-shot ベクトルを描画した。BERT へは BERT のトークナイザで分割して入力した。オノマトペ候補が 2 つ以上のトークンに分割された場合には、平均ベクトルを描画した。また、ベクトルは主成分分析[2] により二次元に圧縮して描画した。ABAB 型のオノマトペ候補の例文は「ABAB」「ABAB する」「AB る」の三種類について収集されており、それらを別の色とマークのプロットとして描画した。描画の際は可視性を担保するため、用例の数にかかわらず、最大表示数を「ABAB」「ABAB する」「AB る」の三種類についてそれぞれ 200 とした。またオノマトペ辞典 [12] に多義語として掲載されている場合に限り、辞書の例文の BERT ベクトルもプロットした。実験は GPU を利用し、Google Colab PRO+によって行った。

付録の図 1 に「ボコボコ」「ボコボコする」「ボコる」の BERT ベクトルのプロット図を示す。緑の星が「ABAB」（この例では「ボコボコ」）、水色の丸が「ABAB する」（この例では「ボコボコする」）、茶色の三角が「AB る」（この例では「ボコる」）を示す。「ABAB する」は必ず「ABAB」部分を含むので、「ABAB する」を表す水色の丸のプロットエリアは必ず「ABAB」を表す緑の星のプロットエリアに含まれることに注意されたい。また、図中の数字は、オノマトペ辞典中の例文である。数字の違いは語義の違いである。

4.4 目視による分類

本実験の対象の 64 種類のオノマトペ候補について、図 1 のようなプロット図を見て

(1) 「ABAB」（緑の星）と「AB る」（茶色の三角）の BERT ベクトルのクラスタ間距離

(2) 「ABAB する」（水色の丸）と「AB る」（茶色の

1) https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking

2) https://scikit-learn.org/1.5/modules/generated/sklearn.decomposition.PCA.html

三角）の BERT ベクトルのクラスタ間距離

を「遠い/近い」に目視で分け、それぞれのクラスタ間距離の「遠い/近い」に応じてオノマトペ候補を以下の 4 種類に分類した。

A 型　(1) と (2) が両方遠い
B 型　(1) が遠く、(2) が近い
C 型　(1) が近く、(2) が遠い
D 型　(1) と (2) が両方近い

この際、二つのクラスタが接しているかよりも、二つのクラスタの広がりが重なっているかどうかに注意して分類を行った。分類の結果、A型が 14、B型が 1、C型が 8、D型が 32 となった。残りの 9 つについてはプロットされた用例が少なすぎたため、分類対象外とした。D型は多かったので (1) のクラスタ間距離について、クラスタが重なっているもの（D1 型）と近いが重なっていないもの（D2 型）に大別した。その結果、D1 型が 18、D2 型が 14 となった。図 1 は D1 型の例である。付録の図 2 に A 型「カスカス」の例を示す。なお、この分類は、仮説ありきの分析にならないよう、言語学的な仮説を持たない研究者（第一著者）がクラウドソーシングの調査の結果を参照せずに行った。

付録の表 1 にＡＢる型動詞の分類結果を示す。オノマトペ候補を A、B、C、D1、D2 型の順に並べ替えたところ、第一著者は以下の二つの傾向に気づいた。

- クラウドソーシングの「【AB る】は【ABAB する】をもとにした語であるか」という調査結果の平均値を参照した際、A 型においては【AB る】は【ABAB する】をもとにした語ではないという回答に近くなり、D 型については【ABAB する】をもとにした語であるという回答に近くなる傾向
- D2 型のオノマトペ候補において「ABAB と AB る」（例：ウネウネとウネる）という言い方が自然である傾向

以降、これらの傾向についてクラウドソーシングの追加調査と分散分析にて確認する。

5　分析

5.1　クラウドソーシングの追加調査

（調査 3）「ABAB と AB る」という表現が自然であるかについての調査を、Yahoo! クラウドソーシングを用いて 2023 年 2 月 1 日に実施した。1209 名が参加し、半角・全角の文字を利用した 748 通りの表現について調査した。それぞれの表現につき「ABAB と AB る」の自然さの 0 から 5 の評定情報を 100 人により評価した。

5.2　調査結果の分散分析

付録の表 2 にクラスタ間距離による型分類（A, B, C, D1, D2）とクラウドソーシングの調査の平均の結果を示す。分散分析の結果が統計的に有意であるものにはアスタリスクを記した。（調査 1）の「オノマトペだと思うか」については、分散分析により B 型を除いて型ごとの差が統計的に有意であった（$F(3, 50)=3.459$, $p < 0.05$）。比較すると、A 型の平均値は C 型、D1 型、D2 型の平均値に比べて有意に低い（t 検定：$p<0.05$）。B 型は用例が少なすぎて検定を行うことができなかったが、A 型よりも高い値であることが表から見て取れる。この結果から、「ABAB」と「AB る」のクラスタ間距離が遠く、また「ABAB する」と「AB る」のクラスタ間距離が遠い場合には、オノマトペになりにくいことが分かる。

（調査 2）「【AB る】は【ABAB する】をもとにした語であるか」については A 型よりも D 型のほうが数値が高くはあるものの、分散分析によれば有意な差は認められなかった（$F(3, 50)=1.235$, $p = 0.306$）。有意差については用例数が少ないことが原因である可能性がある。

（調査 3）「「ABAB と AB る」の自然さ」については、分散分析による型ごとの差が統計的に有意であった（$F(3, 50)=2.697$, $p = 0.0557$）。D 型（D1 型と D2 型の合計）は少なくとも A 型と C 型よりも自然であると評価された（t 検定：$p<0.1$）。この結果から、「ABAB」と「AB る」のクラスタ間距離が近く、また「ABAB する」と「AB る」のクラスタ間距離が近い場合には、「ABAB と AB る」と言えることが分かる。「ABAB と AB る」が言えるか、というテストは D 型の判定に使える可能性がある。しかし、D2 型は、D1 型よりも数値が高くはあるものの、有意な差は認められなかった。有意差については用例数が少ないことが原因である可能性がある。

「ABAB と AB る」が自然となるのは、副詞「ABAB と」が表す様態と動詞が表す動作が不揃いでちぐはぐではないが、一方で重複し過ぎてもいない場合である、と考えられる。例えば、「ガミガミとガミる」は「ガミガミ」は文句を言う様を表し、「ガミる」は

損をすることを表し、意味が通らずに不自然な表現となる。一方で、「モフモフとモフモフする」や「ナデナデとナデナデする」は、意味の過剰な重複があり、不自然である。ところが、「モフモフとモフる」や「ナデナデとナデる」は自然だと判断される。これは、「ぴょんぴょんと跳ねる」が自然であるのと同様だと考えられる。ここでの観察は [3] の研究で「モフる」が「モフモフする」より類像性が低く、表す意味範囲が一部重複しつつも異なっており、一般動詞に近い特性をもつ、とした分析と一貫性を持つ。

5.3 分類の言語学的分析

4.4 節での分類について、言語学的に、特に「AB る」の分布を観察・分析した。表 1 の「AB る」のうち、日本最大の国語辞典である、日本国語大辞典 [13] に同音同義のエントリがなかったものに下線を付した。つまり、下線のある動詞は新語の可能が高い。

A 型では 14 語中、「テカる」のみ下線となる。「テカる」は [10] でも特殊な性質（出現時期からは新語と推定されるが、類像性の相対的高さは新語らしくない）を持つと示され、今回の特殊な分類結果もこの性質と関連する可能性が高い。残り 13 語はどれも、一般的な動詞である。また、5.2 節での分析によれば、A 型の ABAB の多くはオノマトペではないとされる。実例を見ると、「上げる」と「まずは新年会でアゲアゲスタート」の「アゲアゲ」のように、一般的な動詞の語幹を繰り返し、特定の文脈で用いられるものが多い。A 型は主に一般的な動詞を元とする ABAB の分類である。A 型は「ABAB」と「AB る」のクラスタ間距離と「ABAB する」と「AB る」のクラスタ間距離が両方遠い分類である。そのため、オノマトペでない「ABAB」と一般動詞の「AB る」が意味的に関係しつつも用いられる文脈が異なることと矛盾しない。

B 型は下線のない「アセる」1 語のみで、「焦る」あるいは「汗」から「アセアセ」となったと考えられる。1 語で十分な分析はできないが、A 型と共通点が多い。

C 型は、8 語のうち 6 語に下線がある。例を見ると、動詞は新語だが、オノマトペには由来ではないものが多い。例えば「ボツる」は「ボツボツ」のような凹凸の形状ではなく、没にするという意味で関係していない。C 型は「ABAB」と「AB る」のクラスタ間距離が近いが「ABAB する」と「AB る」のクラスタ間距離が遠い分類である。「ABAB」と「AB る」のクラスタ間距離が近い理由は現在のところ不明である。一方で、「ABAB する」と「AB る」のクラスタ間距離の遠さは、新動詞がオノマトペ由来でないことと一貫する。

D1 型では 18 語のうちの 3 語、D2 型では 14 語のうちの 3 語が新語の可能性がある。「ボコる」「モフる」のようなオノマトペ由来の新語が観察される。ただし、A 型同様「ナデる」のような、一般動詞も多く含まれる。D 型は「ABAB」「ABAB する」「AB る」のクラスタがどれも近くに存在する。この性質はオノマトペ由来の新動詞がこの型に含まれやすいことと矛盾しない。ただし、「撫でる」から「ナデナデ」が派生したケースのように、オノマトペと動詞の由来が逆であるものも多く含む。由来が逆であってもクラスタが近いこととは矛盾しないため、オノマトペ由来の新動詞の検出には、このグループに絞り込んだ後、由来がどちらなのかを見極める必要がありそうである。

大まかな傾向としては、A 型と B 型には一般動詞が、C 型にはオノマトペには由来しない新動詞が、そして、D 型にはオノマトペ由来の新動詞と一般動詞が、分類されていると言える。

6 まとめと展望

本研究では、BERT ベクトルを用いてオノマトペ由来の新動詞の検出を試みた。新語を分析対象とするため、オノマトペ候補はコーパスとクラウドソーシングによる調査で選出した。「ABAB」「ABAB する」「AB る」についてコーパス上の用例を zero-shot の BERT ベクトルにして主成分分析で圧縮し、二次元に描写した。このベクトルの「ABAB」「ABAB する」「AB る」のクラスタ間距離を目視で「近い/遠い」に分け、これをもとに 4 つの型に分類した。この分類を用いてオノマトペ由来の新動詞の検出を試みたところ、オノマトペ由来の新動詞は「ABAB」「ABAB する」「AB る」のクラスタがどれも近くに存在する D 型に多く見られることが分かった。今後の研究として、目視ではなく機械的にクラスタ間距離を求めること、文脈ベクトルの改良などが考えられる。

謝辞

本研究は JSPS 科研費 JP21K12603、JP22K12145、及び国立国語研究所共同研究プロジェクト「アノテーションデータを用いた実証的計算心理言語学」の助成を受けたものです.

参考文献

[1] Hans-Jörg Schmid. New words in the mind: concept-formation and entrenchment of neologisms. **Anglia**, Vol. 126, pp. 1–36, 2008.

[2] 宇野良子, 鍛治伸裕, 喜連川優. 新動詞の成立にみる意味と形の変化の相関ー「ファブる」と「モフる」の分析からー. 日本認知言語学会論文集, 第 10 巻, pp. 377–386, 2010.

[3] 宇野良子, 鍛治伸裕, 喜連川優. ウェブコーパスの広がりから現れるオノマトペの二つの境界. ひつじ書房, 『篠原和子・宇野良子（編）『オノマトペ研究の射程ー近づく音と意味ー』, pp. 245–260, 2013.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the NAACL-HLT 2019, Volume 1**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[5] Kanako Komiya and Yoshiyuki Kotani. Classification of japanese onomatopoeias using hierarchical clustering depending on contexts. In **The 2011 English International Joint Conference on Computer Science and Software Engineering(JCSSE)**, pp. 108–113, 2011.

[6] Kanako Komiya, Minoru Sasaki, and Hiroyuki Shinnou. Comparison of distributed representations and context vectors for japanese onomatopoeia classification. In **CICLING 2018**, p. no 57, 2018.

[7] 乙武北斗, 内田ゆず, 高丸圭一, 木村泰知. Bert による周辺文脈を考慮したオノマトペの語義分類手法の提案. 知能と情報, Vol. 32, No. 1, pp. 518–522, 2020.

[8] 小林千真, 相田太一, 岡照晃, 小町守. Bert を用いた日本語の意味変化の分析. 自然言語処理, Vol. 30, No. 2, pp. 713–747, 2023.

[9] Ryoko Uno, Kanako Komiya, and Masayuki Asahara. How do we categorize known and unknown ideophones? a case study of japanese reduplicated ideophones. In **International Conference on Cognitive Linguistics 16 (ICLC16)**, 2023.

[10] 宇野良子, 古宮嘉那子, 浅原正幸. オノマトペ由来の新動詞分析のための大規模アンケート調査. 認知科学会第 40 回大会論文集, pp. 339–342, 2023.

[11] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. Archiving and analyzing techniques of the ultra-largescale web-based corpus project of ninjal, japan. **The Journal of National and International Library and Information Issues**, Vol. 25, No. 12, pp. 129–148, 2014.

[12] 小野正弘. 擬音語・擬態語 4500 日本語オノマトペ辞典. 小学館, 2007.

[13] 北原保雄, 久保田淳, 谷脇理史, 徳川宗賢, 林大, 前田富祺, 松井栄一, 渡辺実. 日本国語大辞典〔第 2 版〕. 小学館, 2000-02.

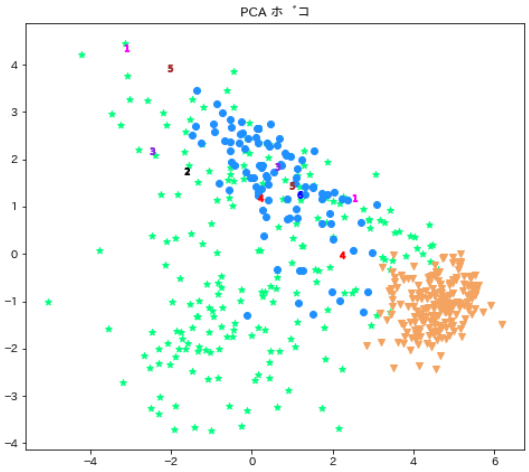

図 1　ボコボコの BERT ベクトルのプロット図

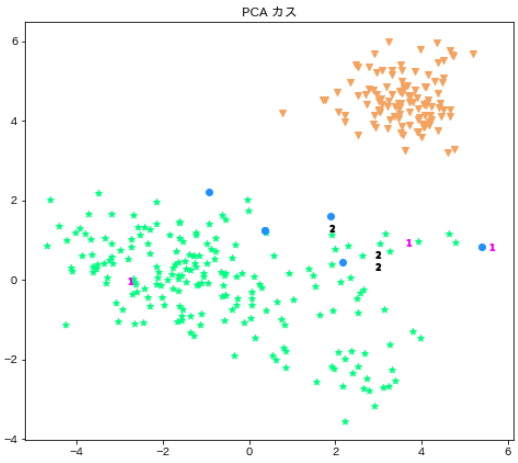

図 2　カスカスの BERT ベクトルのプロット図

	(1)ABAB\AB る間	(2)ABAB する \AB る間	AB る
A 型	遠い	遠い	アゲる、ボケる、ブレる、グズる、ホレる カスる、クサる、モエる、ムレる、スベる テカる、テレる、ツメる、ウツる
B 型	遠い	近い	アセる
C 型	近い	遠い	ビクる、ボツる、ガチる、ガミる、ヒタる モジる、パシる、ヨレる
D1 型	近い（重ならない）	近い	ボコる、チビる、ドナる、ドヤる、グネる ヘナる、イキる、カジる、マゼる、モグる モヤる、ネバる、ネムる、ニギる、ノビる サワる、シバる、ヨタる
D2 型	重なる	近い	チクる、グスる、ヒエる、ホジる、コネる クネる、モフる、ナデる、ニジる、ヌメる シコる、シクる、ウダる、ウネる

表 1　「ABAB」\「AB る」、「ABAB する」\「AB る」のクラスタ間距離による A B る型動詞の分類

型	数	調査 1	調査 2	調査 3	
A	14	2.87	2.36	1.87	
B	1	3.45	2.60	2.16	
C	8	3.63*	2.21	1.82	
D1	18	3.39*	2.44	2.25*	2.09
D2	14	3.55*	2.97		2.40
全体	55	3.33	2.52	2.07	

表 2　クラスタ間距離による型分類（A, B, C, D1, D2）とクラウドソーシングの調査結果