BERT の文脈理解を理解する
Understanding what BERT understands
—Onomatopea understood with contexts—

浅川 伸一 †, 近藤 公久 ‡
Shin Asakawa, Tadahisa Kondo
† 東京女子大学, ‡ 工学院大学
Tokyo Women's Christian University, Kogakuin University
asakawa@ieee.org

概要

ニューラルネットワークによる言語モデルでは，トランスフォーマー [1, 2] に基づくモデルが支配的となっている．これらのモデルの表現能力を用いて認知過程の理解を模索することには意味があるだろう．本稿では，オノマトペを題材に，表現の簡潔さと，それらの意味の豊富さを記述し，分類することにより，トランスフォーマーベースのモデルの応用可能性を示した．

キーワード：BERT, SentenceBERT (SBERT), トランスフォーマー, オノマトペ，文脈，微調整 (fine-tuning)

1. はじめに

トランスフォーマーや，その応用である BERT に基づく諸モデルでは，マルチヘッド自己注意機構に構成要素とする大規模ネットワークを，大規模な訓練コーパスを用いて **事前学習 (pre-training)** を施す．加えて，**下流課題 (downstream tasks)** に対して，**微調整 (fine-tuning)** や **転移学習 (transfer learning)** が行われる．事前学習と微調整とにより，**少撃学習 (few-shot learning)** または **零撃学習 (zero shot learning)** が実現される．近年話題となっている多くのモデルでは，概ね上述の訓練プロトコルに従っていると考えられる．GPT-3 [3] や，Gopher, Chinchilla, PaLM なども同じ枠組みとみなしうる．Chinchilla はわずかに異なるスケーリング則の成功を証明したが，他のモデルと同様，大量のデータと計算を使用する大規模なトランスフォーマーに基づくのモデルであることに変わりない．DALL-E 2，Imagen，Parti はトランスフォーマーを超える技術を追加したテキストを画像へ変換するモデルという点では異なるものの，ほとんど同じ流れに沿っている．Flamingo と Gato も GPT-3 から少し離れて，より一般的でマルチモーダルなアプローチを取っている．だが同じアイデアを新しい課題に適用したリミックスに過ぎないと言えよう．

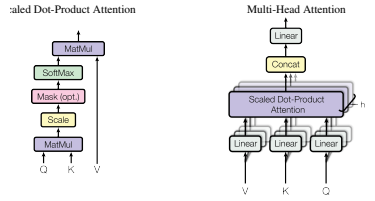

図 1　一つの入力情報から，クエリ，キー，バリューの 3 つのベクトルを作成し，マルチヘッド自己注意機構への入力とすることでトランスフォーマーの基本構成要素となる．文献 [1] Fig. 2 より

微調整や転移学習による精度向上は，モデルの持つ豊富な内部表現を反映していると考えられる．日本語でも事前訓練済みモデルが複数公開されており活況を呈している．我々はこのことを鑑み，オノマトペの群化，記述を通して文脈理解への応用を検討した．具体的には微調整を工夫して，課題に対して精度良く出力を調整するという枠組みは，機械学習の慣習としてだけではなく，認知過程における文脈情報の果たす役割ともみなすことが可能であろう．今回は，このようなグランド仮説に基づいている．

我々は，単語埋め込みモデルを用いた意味推論において，モデルのパラメータが張る部分空間を文脈として捉えることを提案している [4, 5]．この提案モデルは，従来モデルでは，操作や検索過程などの定義が曖昧であった心的操作を表現可能である．このことを，日本語 wikipedia を用いて訓練した単語のベクトル表現集合を意味空間とし，これらに対して射影変換によって文脈効果が得られるかを検討した．本稿では，この考え方をさらに進めて，BERT の微調整により文脈情報を扱うことに焦点を当て，オノマトペの理解過程を取り上げた．

微調整を敷衍して考えれば，我々が特定の課題を解

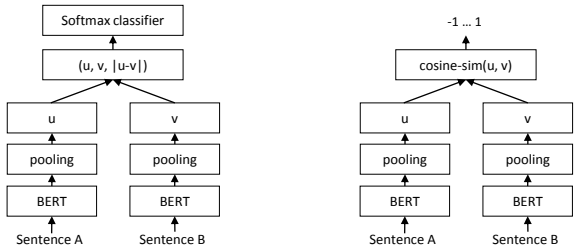

図 2　SBERT で用いられた微調整のためのシャムネット (左)，類似性判断ネットの構成 (右)。文献 [6] Fig. 3 より

く際にも，課題毎に微調整を行っているとも考えることができる．事前学習では，各課題に対して正解することが難しいが，逆に言えば，どの課題にも適応可能な準備がなされていると解釈することできる．一方で，微調整ではパラメータ空間を課題に適合するように変形させるため，内部表象が変化したと考えられる．このような微調整は，認知科学分野における熟達化の過程に喩えられるような変化ともみなしうる．熟達化の過程を内部パラメータの変化とみなすことにより，認知科学と人工ニューラルネットワークとの橋渡しをすることにもつながる．本稿では，上述の問題意識から，微調整における精度向上と，その意味についての解釈を試みた．

1.1　微調整における良好な文脈表現を得るための工夫

古典的な，TF-IDF による文のベクトル化を始めとして，任意の文章をベクトル化する手法には，多くの提案がなされてきた．word2vec[7] によって得られた単語ベクトルを平均化する手法や，seq2seq 翻訳モデル [8] も，それらの一つとして位置づけられよう．BERT に基づく手法としては，位置符号化器の出力を加え合わせるだけでなく，文符号化トークンとして [CLS] 利用する提案がなされた [1].

ところが [CLS] トークンでは，2 文の類似度判断に十分な精度が得られないことが指摘されている [6]. より良い文ベクトル表現を求めて，微調整の工夫として，目的関数と，出力調整が挙げられる．**対比学習 (contrastive learning)** [9] あるいは，対比学習に影響を与えた **雑音対比推定 (Negative Contrastive Estimation:NCE)** [10] は目的関数の工夫に相当する．一方，CLIP や DALL-E などは出力調整の一種であるプロンプトエンジニアリングとみなしうる．

目的関数に適切な制約を加えることで，正事例間のバイアスを小さくし，負事例との相互情報量を低く抑えるように作用すると考えられる．我々の持つ単語の意味や概念の形成過程を考える上でも，あるいは，推論過程をモデル化する場合においても，適切な正および負事例の選定と，それらを用いた表現学習が重要になると思われる．これらの技法は，ラグランジェ乗数を用いた制約付き最適化として定式化可能であるが，制約付き最適化という枠組みを超えて，適切な表現学習のために必須の概念であるとみなすことができよう．

Sentence BERT (以下 SBERT) [6] は，BERT [1] に対して良好な文ベクトルを得るための微調整と捉えることができる．SBERT は文を 1 つのベクトル表現（分散表現）し，文間の類似性を求めることを可能にする．オリジナル BERT では [CLS] トークンに文ベクトルが表現されていることとなっている．しかし，[6] によれば，BERT の [CLS] トークンによる文ベクトルでは，cos 類似度を用いた文間の類似度比較や，cos 類似度を用いたスピアマン順位相関での比較で，良好な精度が得られるとは言い難い．

加えて SBERT は BERT に比して計算量が少ないという利点がある．BART の検索時の計算複雑さは $O(NK)$ のオーダーであるのに対して，SBERT の複雑さはたかだか $O(N+K)$ である．ここで N は文章数，K はトピック数である．

類似した文ベクトルを得る手法に BART がある．BART を使ってトピックが単語に似ているかどうかを調べる場合，文と潜在的なトピック（[SEP] トークンで区切られる）を連結して BART トランスフォーマーを通す必要がある．これはすべての潜在的なトピックに対して行う必要がある．BART は 2 文が中立 (お互いに関係ない)，含意，矛盾である確率を出力する．このような計算効率から考えて，今回は SBERT を採用した．

2.　実験

2.1　マスク化言語モデルによるオノマトペ推定

2.1.1　手続き

本研究では事前訓練済の日本語 BERT モデル，Huggingface [1] のマスク化言語モデルを用いた．日本語オノマトペ辞典 [11] に掲載されているオノマトペ

[1] https://github.com/huggingface/transformers

のうち日本語 wikipedia に出現するオノマトペ 1971 語を対象とした．しかし，事前訓練済みの標準 BERT モデルでは，ほとんど (1961/1971 語) のオノマトペ単語が未登録であったため，事前訓練済モデルでは，オノマトペのほとんどを扱うことができなかった．そこで 1961 語の未登録語をユーザ辞書に登録し，その上で，日本語オノマトペ辞典に掲載されているオノマトペの用例にオノマトペを埋め込む形に変形した文を用いて微調整を行った．微調整では変形した例文中のオノマトペをマスクして，マスクされたオノマトペを予測するように訓練した．

- 辞典の用例そのまま：　真夏の太陽が強く照りつけて：ぎらぎら
- 変形後の用例： 真夏の太陽が [MASK] 強く照りつける

実験では，用例 1741 文中の 70% をランダムに選択して微調整を行い，残りの 30%を用いてテスト文とした．

2.1.2　結果

マスクしたオノマトペそのものを推定する例は多くはなかったが，予測された語をマスクに置換した文は，文として整合性のある文となるものが多く存在し，訓練済みモデルをそのまま用いる場合より良好な推定が可能となった．これは微調整により，マスクの位置に存在しうる単語がどのような性質をもった単語であるかを少ない微調整事例から学習することができたと考えられる．しかし，オノマトペそのものが推定されないのは，訓練済みモデルにオノマトペそのものは存在しておらず，汎化することは困難であると言える．

2.2　文 BERT によるオノマトペの群化

2.2.1　手続き

オリジナル BERT におけるマスク化言語モデルでは，マスク化された語を文内のすべての単語から推定するものである．すなわち，マスクされた語の位置に相応しい語を示す狭い意味での文脈が入力，その文脈においてもっともらしい単語を推定する．このため，オリジナル BERT においては，それぞれの文脈でのもっともらしい単語を推定することはある程度できるようになったが，オノマトペに特に着眼されることはない．一方，本研究は，オノマトペの意味的な類似性や文脈における多義性などを探求すること，および，文脈を表現する AI モデルの仕組みを探ることを目的としている．そこで，ここでは，SBERT により，文の類似性を学習しうるモデルを用いて，BERT の表現しうる文脈と，その文脈を用いた，オノマトペの類似性を示すことを検討する．

ここでは，事前訓練済の日本語 BERT モデルを微調整した日本語 SBERT を用いた．実験に用いた文は，マスク化言語モデルに対する実験と同じく，日本語オノマトペ辞典 [11] に掲載されているオノマトペのうち日本語 wikipedia に出現するオノマトペ 1971 の用例文を用いた（オノマトペを用いる状況を表す文であり，オノマトペ 4500 の用例の表現では，この文のあとにオノマトペを示している（前節の例を参照）．

2.2.2　結果

リード文のみ 2941 文に対して SBERT により得られた各文の分散表現を tSNE により 2 次元で表したものが図 3 である．各点が一文である．図をみてわかることは，小さな集団が散在しており，意味が類似するものが集まっている．例えば,「がたがた」,「がたん」,「どすん」,「ゆっさゆっさ」が一つの群となっている．これは，音韻的な類似性もあるように見えるが，基本的には，大きめの音，揺れたり，落ちたりしたときの音を示すオノマトペである．これらのオノマトペに対応するリード文は，順に,「長いものがだらしなくゆれて」,「重いものが急に転倒・落下し」,「大きく重いものが落ちた衝撃で」,「大きく重そうなものがゆれて」である．

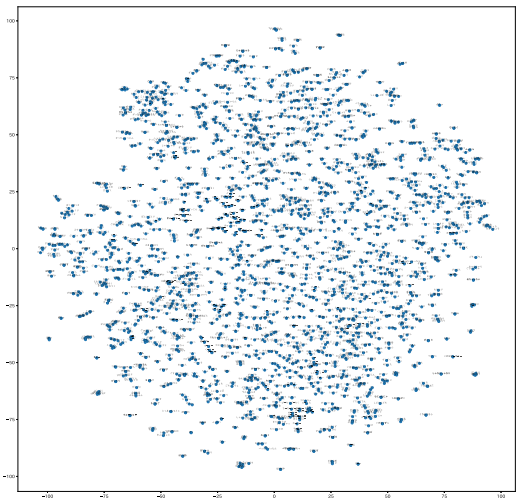

図 3　SBERT によるリード文の tSNE[12] 布置

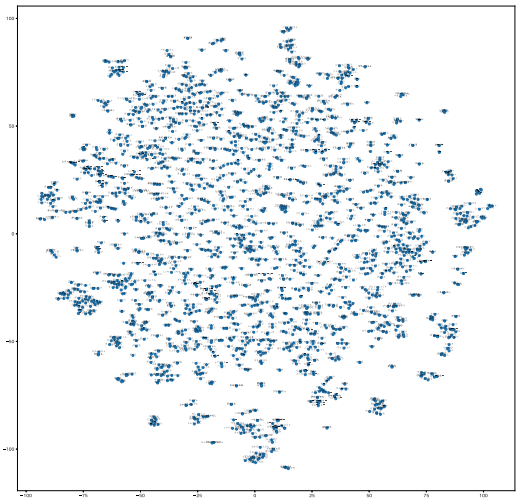

図 4　SBERT によるオノマトペ文の tSNE 布置

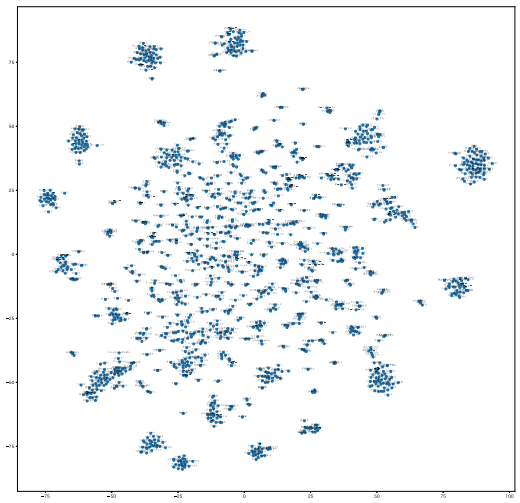

図 5　SBERT によるオノマトペ単語の tSNE 布置

次に，リード文の後に「,」を入れてオノマトペを付加したものを文として SBERT で分散表現を生成させたときの tSNE による解析結果を図 4 に示す．この文の表現は，オノマトペ 4500 の用例の表現と一致するものである．先のリード文のみの図 3 と比較して，図 4 は各文（オノマトペ）の群がかなり異なって見える．例えば，上のリード文だけの場合に「がたがた」,「がたん」と同じ群に入っていた「ゆっさゆっさ」や「どすん」は少し離れた所に付置されている．これは，音韻的類似性が強く影響する群化となっていることが考えられる．

ちなみに，オノマトペ単体の分散表現を上述と同様に SBERT で分散表現を得た場合の tSNE での分析結果を図 5 に示す．この結果は，オノマトペの音韻的類似性によって明確にクラスター化されていることが示された．本題とはずれるが，オリジナルの BERT に存在しないオノマトペが token 化されることによって，音韻的連続の分断化がなされるため，センテンスの類似性が類似した単語の並びを学習しているため音韻の類似性に依存した類似性を表現した分散表現となると考えられる．

さらに，ここで対象としたオノマトペのようなオリジナルの BERT において学習文に含まれてない単語ではない単語の説明文を対象とした場合にはどうなるかを確認してみた．以下の 2 つの例文の類似度を計算した例を示す．それぞれ，運営と経営の辞書（Apple 辞書バージョン 2.3.0）の説明文で試した例を以下に示す．

なお，類似度は，SBERT で得られた 2 つ文の分散表現の cos 類似度を求めた．

例）経営と運営の辞書説明文の類似度上位 5 文

入力文: 組織や機構などを動かし，うまく機能するようにすること、運営

1. 彼にはその仕事をする能力がある．(0.881)
2. 彼は働く人の生活を良くするのに力を使った．(0.877)
3. 彼には事業を経営するのに十分な能力がある．(0.872)
4. この方法は、いろいろなところで使える．(0.871)
5. しっかりした計画と努力のおかげです．(0.871)

入力文: 方針を定め，組織を整えて，目的を達成するよう持続的に事を行うこと、経営

1. 彼は目的を達成するために頑張って働いた．(0.890)
2. 彼は人生の計画を立てるために頑張って働いている．(0.883)
3. 彼はその計画を実行するために力の限りを出した．(0.878)
4. 成功するためには、努力して働かなければならない．(0.874)
5. 彼には事業を経営するのに十分な能力がある．(0.873)

3. 考察

3.1 BERT によるオノマトペ推定と文脈

微調整に用いた訓練マスク化言語モデルは，マスクされた位置の単語を周辺の単語から推測する課題である．この課題は，オノマトペを直接的に推量する課題と言える．

今回用いた事前訓練済 BERT の結合係数は，BPE[13] より正確には sentencepice [14] を用いて部分単語に分割するトークナイザを用いてトークン化したコーパスを用いている．このため，訓練コーパス内に注目するオノマトペが存在しても，適切なトークン化が行われていなかったという可能性が指摘できる．微調整時には，新トークンとして学習することになったため，事前訓練で用いられた文脈情報を適切に利用できなかった可能性が指摘できよう．このため，文表現ベクトルに含まれるオノマトペを処理するために必要な情報を微調整することが困難であった可能性が指摘できるであろう．

3.2 SBERT による文類似度と文脈

上述の，オノマトペ推定課題に比べて，事前訓練として文表現が適切に行われていれば，BPE による部分単語への分割の影響は少なく，結果として良好な群化がなされたとの解釈が可能となる．この意味において，適切な目的関数と微調整とを工夫すれば，文脈情報を取り出すことができるようになると予想される．このことは，最近の Gato などの流れとも矛盾しないと思われる．

4. 結論

オノマトペを題材に，BERT を用いて単語の意味表現と文脈を検討した．BERT では単語の意味的表現が文内の文脈に依存して分散表現として得られているとされている．しかし，マスク化言語モデルにおけるマスク語の推定においては，逆に，同じ文脈においてもさまざまな単語が候補となり得るため，文として整合性のある単語を推定することが可能であってもその分散は大きく，ある特定の領域の語を推定させるために fine-tuning を行っても効果は大きくない．

一方 SBERT の文の分散表現，および，文の類似度を用いれば，辞書のような，ある単語に対する説明文や用例といった，対応関係にある組が存在する場合には，説明文や用例の類似度が対応する語どうしの類似度と考えてよいことが示された．また，オノマトペのようにオリジナル BERT にとって未知語であっても，それと対応する説明文や用例さえあれば，類似の単語を探し出すことが可能である．このような機能は，SBERT による説明文の分散表現が単語の意味表現として利用できることを意味するものであり，文脈とは何か，そもそも意味とは何か，意味はどのように表現されているのか，という疑問に対する答えに一歩近づいているのではないかと期待できる．

文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Łukasz Kaiser. Attention is all you need. *arXiv preprint*, [cs.CL](1706.03762), 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2018.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint*, [cs.CL](arXiv:2005.14165), 2020.

[4] 浅川 伸一 and 近藤 公久. 単語の意味空間を心的操作する＝ 射影. In **第** *38* **回日本認知科学会発表予稿集**, 2021.

[5] 浅川 伸一 and 近藤 公久. Bert による意味表現. In **日本心理学会第** *83* **回大会予稿集**, 立命館大学, 2019.

[6] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint*, [cs.CL](1908.10084), 2019.

[7] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL*, Atlanta, WA, USA, June 2013.

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 3104–3112, Montreal, BC, Canada, 2014.

[9] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *arXiv preprint*, [cs.CV](arXiv:2011.00362), 2020.

[10] Mike Wu, Milan Mosse, Chengxu Zhuang, Daniel Yamins, and Noah Goodman. Conditional negative

sampling for contrastive learning of visual representations. *arXiv preprint*, [cs.LG], 2020.
[11] 小野 正弘, editor. **日本語オノマトペ辞典**. 小学館, 東京, 第一版 edition, 2007.
[12] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
[13] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint*, May 2016.
[14] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*, 2018.