人間の脳と人工知能における短歌の鑑賞に関する神経活動の比較

船井 正太郎[1] 近添 淳一[1] 持橋 大地[2] 浅原 正幸[3]
松井 鉄平[4] 鹿野 豊[5] 川島 寛乃[6] 磯 暁[7]
[1]株式会社アラヤ [2]統計数理研究所 [3]国立国語研究所
[4]岡山大学 [5]群馬大学 [6]慶應義塾大学 [7]高エネルギー加速器研究機構
{funai_shotaro,chikazoe_junichi}@araya.org, daichi@ism.ac.jp, masayu-a@ninjal.ac.jp,
tematsui@okayama-u.ac.jp, yshikano@gunma-u.ac.jp, hirono@ht.sfc.keio.ac.jp, iso@post.kek.jp

概要

短歌を自然言語処理の人工知能（機械学習）に入力したときの振る舞いについて、短歌を読むときの人間の脳神経活動と比較することにより、解析を行った。具体的には、機械学習のモデルとして BERT, 脳神経活動の測定手法として fMRI（機能的 MRI）を用いて、BERT の各層が出力する分散表現と脳の各 voxel における fMRI データとの対応を議論した。その結果、BERT の深い層の振る舞いは、文が詩的であるかを判定する脳部位や、抽象的な情報の認知を行う脳部位と対応していることが示せた。

1 はじめに

近年、人工知能（機械学習）を用いた自然言語処理は急速に発展してきており、文章の生成や要約、多言語間の翻訳など、幅広いタスクをこなせるようになってきた。これは機械学習が言語をより深く理解できるようになってきたことを意味しており、BERT [1]や GPT [2]などの機械学習モデルを中心に、数多くの研究開発がなされている。

そうした研究の多くは Wikipedia などの日常的な文を扱うものである。そこで本研究では、詩的な文を扱い、機械学習の振る舞いを解析する。詩的な文は一般的に、文字通りの意味を超えた内容を持ち、人間を感情的にさせることができる。そうした文に対して、機械学習がどのように言語処理を行うかを議論するのである。

詩的な文として、本研究では短歌を扱う。散文詩のように長い詩だと、様々な形式があって比較が難しい。逆に、俳句のように短い詩だと、読み解くのに前提となる知識が必要とされることが多く、いずれも機械学習には向かないと判断したからである。

また、短歌との比較対象として、一般的な文（本研究では平文と呼ぶ）も扱う。文の長さは短歌と同じく 31 音程度のものに限る。これらの短歌や平文を入力したときの機械学習モデルの振る舞いを調べ、さらに短歌・平文を読むときの人間の脳神経活動とどのくらい共通しているのか解析を行う。

機械学習モデルとしては BERT (base)を用いて、人間の脳神経活動を測定するのには fMRI（機能的 MRI）の手法を用いる。そして、BERT の分散表現から fMRI データへの回帰解析を行うことにより、人工知能と人間の脳神経活動の振る舞いにどのような対応関係があるのかを明らかにする。

特に、BERT の深い層は文の意味的な性質を捉えると考えられている[3]。また、脳科学ではより抽象的な情報を捉える脳の部位がわかってきている[4]。自然言語処理と脳科学、両方の知見に基づいて短歌を扱うことにより、自然言語処理の人工知能が持つ能力をより深く理解できると考えられる。

2 人工知能の振る舞い

2.1 短歌と平文の選定

本研究に用いる短歌と平文は、国立国語研究所の『現代日本語書き言葉均衡コーパス』(BCCWJ)に収録されたものを用いた。このコーパスには短歌が 3603 首、平文（31 音程度の一般的な文）が 5624 文ある。ただ、口語調の短歌が少ないため、現代短歌集である『桜前線開架宣言』[5]と短歌の月刊誌『塔』[6]から著者（持橋・川島）が選んだ 101 首を加えて、本研究で用いる短歌のデータベースとした。

これら合計 9328 文を、BERT の学習済みモデル（cl-tohoku/bert-base-japanese-whole-word-masking, 東北大学モデル）に入力し、最深層の[CLS]分散表現を取り出して、2 次元に PCA（主成分分析）を行うと、図 1 のように描画できる。

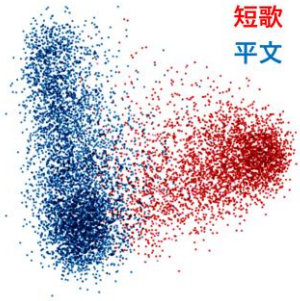

図 1　短歌・平文の BERT 分散表現

この BERT モデルは日本語版 Wikipedia の文を訓練データとしているため、短歌などの詩的な文はほとんど学習していない。それにも関わらず、図 1 は短歌と平文の違いが明確に認識できていることを示しており、興味深い結果だと言える。

これら短歌・平文の中から、人間の脳神経活動を測定する際に用いる文を、以下の方法で選定した。まず、最深層の[CLS]分散表現（768 次元）を入力すると、短歌であれば 0、平文であれば 1 を出力するように教師あり学習を行って、短歌らしさ・平文らしさを予測する機械学習モデルを作成した。隠れ層は 1 層で、すべて全結合層である。この予測モデルを用いて、平文らしい短歌 300 首と短歌らしい平文 300 文を選んだ。その上で、文の読みやすさを考慮して著者（船井・持橋）が短歌 150 首と平文 150 文を選び出した。一目見ただけで短歌か平文か、見分けがついてしまうと、本研究で期待している脳神経活動データが十分に取れない可能性があるため、このような方法で短歌と平文を選定した。

これら 300 文について、先ほどと同様に BERT の分散表現を 2 次元 PCA すると、図 2 のようになる。短歌と平文の分布は重なっているものの、明確にずれており、多くの文は短歌・平文の判別が可能であるように見える。従って、我々が期待したような短歌と平文が選べていることがわかる。

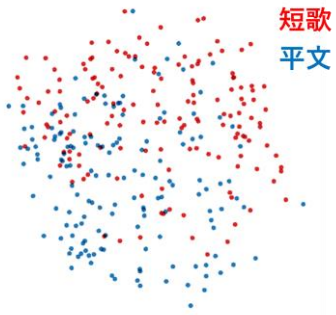

図 2　選定した短歌・平文の BERT 分散表現

3 人間の脳神経活動

3.1　脳神経活動データの測定

2.1 節で選び出した短歌と平文について、人間が読むときの脳神経活動データは、生理学研究所に設置された 3.0 テスラの MRI（シーメンス社製）を用いて測定した。被験者は、生理学研究所の近辺に住む大学生を中心とした、18 歳から 34 歳までの 32 名の健常な日本語母語話者である。そのうち 15 名が男性であり、全員が右利きである。被験者を選ぶ際に、短歌に慣れ親しんでいるかは問わなかった。

この測定実験では、短歌 150 首と平文 150 文が、被験者ごとに異なるランダムな順番で提示される。各々の短歌・平文を 3 行に分けて、最初の 3 秒間は 1 行目のみ、次の 3 秒間は 2 行目まで、次の 3 秒間は 3 行すべてが表示される。最後の 3 秒間は「詩的であると感じますか？」と表示することで被験者に問いかけ、「はい」または「いいえ」をボタンで回答してもらった。このように、1 つの文を 12 秒かけて読み、詩的か否かを判定してもらった。

この作業を短歌 25 首と平文 25 文で 1 セクションとして、セクションが終わるごとに数分間の休憩をはさみながら、全部で 6 セクション行った。

3.2 脳神経活動データの解析

この測定では fMRI の手法を用いて、血流が活発になる脳の部位を時系列データとして取得した。空間の解像度（各 voxel のサイズ）は 2.0mm 立方で、時間の解像度は数秒程度である。この測定で得られた脳活動データを概観したものを、図 3 に示そう。

赤色の領域は、短歌・平文を提示したときに強く

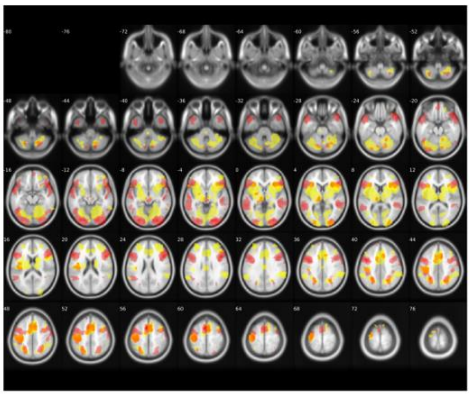

図 3　fMRI 脳活動データの概観

反応した部位を示している。言語野や視覚野に反応が見られ、これは期待通りの結果である。

黄色の領域は、被験者が詩的だと答えたときと詩的ではないと答えたときに、脳活動が大きく異なる部位を示している。被験者が読んだのが短歌か平文かは関係なく、詩的に感じたかどうかを問うていることに注意されたい。特に差異が目立つのは、腹内側前頭前皮質(ventromedial prefrontal cortex), 楔前部(precuneus), 左側の側頭頭頂接合部(temporoparietal junction)である。いずれも認知や感情と関係の深い部位であり、短歌と平文で感情の生じ方に違いがある様子を捉えていると考えられる。特に、腹内側前頭前皮質は、神経美学において中心的な役割を持つ内側眼窩前頭皮質(medial orbitofrontal cortex)を含む領域であり[7]、短歌に美しさなどの価値を見出す様子を捉えていると考えることができる。

4 人工知能と脳神経活動の対応

4.1 機械学習の追加訓練

人間の脳活動データとの対応を見るため、機械学習モデルに以下のような追加訓練を施した。

BERT の最深層に 2 層の全結合層を追加して、まずは被験者全員分の判定データ（詩的か否か）をラベルとして転移学習を行った。次に、同じラベルで BERT 最深層まで訓練し、その次はもう 1 つ浅い層まで訓練し…、このように訓練する層を 1 つずつ増やしながら BERT の全 12 層の追加訓練を行った。その上で、各被験者の判定データのみを使って、同様に訓練する層を 1 つずつ増やしながら、BERT の全 12 層の追加訓練を行った。最初から各被験者の判定データのみを使うと訓練データの量が少ないため、このような方法で追加訓練を行った。

この訓練においては、6 セッションの判定データのうち、4 つを訓練データ(training data), 1 つを検証データ(validation data)として訓練する際のハイパーパラメータを調整するために用いて、5-fold 交差検証(cross validation)を行った。残り 1 つは試験データ(test data)として訓練には用いなかった。

以上により、各被験者の、各セッションを試験データとする、合計 192 個の BERT 追加訓練モデルが得られた。

4.2 脳神経活動データへの回帰モデル

追加訓練を施した BERT モデルを用いて、fMRI（脳活動）データとの対応を議論しよう。

まずは、すべての短歌・平文について、BERT 各層の[CLS]分散表現を計算する。また、各文を 3 行に分けたときの 1 行目まで、2 行目までの分散表現も同様に計算する。その際、各文が試験データに含まれた追加訓練モデルを用いた。

次に、BERT 分散表現と fMRI データの対応関係を解析するために、これらの BERT 分散表現から fMRI データへの回帰モデルを生成する。この回帰モデルの基底となる関数(regressor)は、以下のように定義した。その概略を図 4 に示そう。

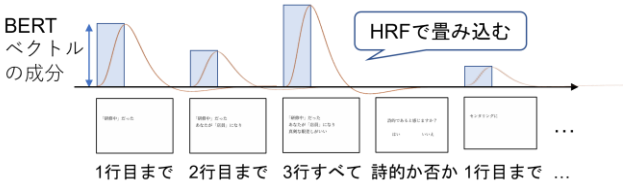

図 4　回帰モデルの regressor となる関数

各被験者が測定実験で提示された短歌・平文の順番（さらに 1 行目まで、2 行目まで、3 行すべての順番）に BERT 分散表現を並べて、各成分の時系列データと見做す。そして、文が提示された瞬間に各成分の値を持つ boxcar 関数を考える（図 4、青色の柱）。詩的か否かを尋ねるときの値はすべて 0 とする。この boxcar 関数に、血流動態応答関数(hemodynamic response function, HRF)を畳み込むと、fMRI データと似た形の関数が得られる（図 4、茶色の線）。この時間の関数を regressor と定義した。

実際の解析では、BERT 分散表現は次元が大きく（768 次元）、overfitting を避ける目的で 100 次元ベクトルに PCA したものを用いた。そのため、被験者ごとに 100 個の regressor が得られ、これらを線形結合することで、各被験者・各 voxel の fMRI 時系列データへの回帰解析を行った。

この解析結果を見ると、BERT のどの層の分散表現が、どの voxel の脳活動とよく対応しているのか、明らかにすることができる。特に、3.2 節で注目した腹内側前頭前皮質(vmPFC)と楔前部(precuneus)付近の結果を図 5 に示そう。これらの脳部位の付近は赤く塗られており、BERT の出力に近い層（深い層）から精度よく回帰ができることがわかる。すなわち、

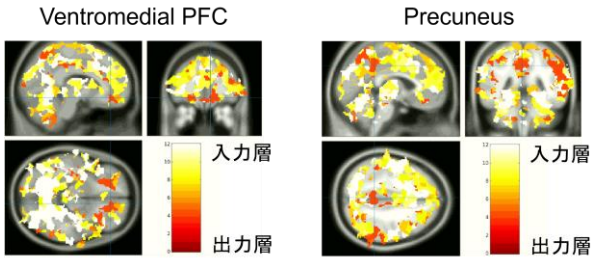

図 5　vmPFC と precuneus 付近の解析結果

BERT の深い層と強く相関している。これらの脳部位は「詩的か否か」の判定と強く関係している部位であった。一方で、BERT は深い層ほど文の意味的な性質を捉えると考えられている[3]。詩的か否かの判定は、意味的な性質を捉えるものであるから、我々の解析結果はこうした主張を支持しているものと考えられる。

4.3 抽象的な認知を行う脳部位との対応

4.2 節で行った回帰解析によって、脳の各 voxel が最も強く相関する BERT の層を同定することができる。一方、脳の各部位が認知する情報がどのくらい抽象的・具体的なものであるかを計算できる手法が提案されている[4]。これは、脳部位間の機能的結合(functional connectivity)のデータを圧縮することによって得られ、認知機能の抽象度を principal gradient score として数値化することができる。

従って、最後に BERT の各層と、その層に強く相関する脳部位の認知機能の抽象度との関係を議論しよう。各部位の principal gradient score を平均して得られた計算結果を図 6 に示す。(被験者 32 名のうち、先に測定を終えた 21 名のみの結果も示した。)

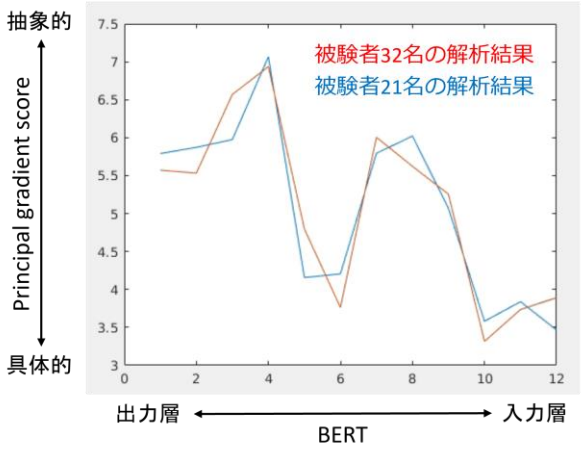

図 6　BERT の各層と認知機能の抽象度の関係

全体的に見れば、BERT の深い層（出力に近い層）はより抽象的、浅い層（入力に近い層）はより具体的な情報を認知する傾向にあることがわかる。BERT は深い層ほど文の意味的な性質を捉え、浅い層ほど統語論（文法的）な性質を捉えると考えられており[3]、それと整合する結果が得られたと言えるだろう。

ただし、BERT の深い層から浅い層へと抽象度が単調に減少しているのではなく、中間層では大きく上下していることにも触れておきたい。この現象は 4.1 節で行った追加訓練が影響して起こっている可能性が考えられる。このような追加訓練を行った BERT モデルでは、中間より浅い層で文の特徴を捉えて、中間より深い層は正しい結果を出力することに重きを置く傾向があるなど、中間層を境に振る舞いが異なるケースが指摘されている。こうした現象は BERT の言語認識能力に密接に関わるものであり、今後の研究で明らかにしていきたいと考えている。

5 おわりに

自然言語処理の人工知能の振る舞いについて理解を深めるために、本研究では詩的な文である短歌を入力データとして扱った。そして、普通の文（平文）を入力したときの人工知能の振る舞いや、短歌・平文を人間が読むときの脳神経活動と比較することにより、自然言語処理と脳科学、両者の知見を活かした解析を行った。

その結果、文が詩的であるか否かを判定する脳部位の神経活動は、BERT の深い層と強く相関していることがわかった。BERT は深い層ほど文全体の特徴を見て、意味的な性質を捉えると考えられているが、我々の結果もそれを支持するものであった。

また、BERT の深い層は、より抽象的な情報の認知を行う脳部位とも強く相関していることがわかった。特に、BERT の深い層と強く相関する脳部位には、美しさなどの価値を捉える、腹内側前頭前皮質が含まれている。

以上のことが本研究にて確認できたことを踏まえ、今後は短歌に慣れ親しんでいる人々に被験者となってもらい、より詳しく鑑賞するときの脳神経活動と、人工知能の振る舞いを比較することで、人工知能が持つ自然言語処理能力をより深く理解していきたいと考えている。

謝辞

本研究は「機構間連携・異分野連携研究プロジェクト」と生理学研究所の共同利用研究からの助成を受けています。

参考文献

1. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv: 1810.04805 [cs.CL], 2018.

2. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, “Improving language understanding by generative pre-training,” https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf, 2018.

3. 例えば、I. Tenney, D. Das, E. Pavlick, “BERT rediscovers the classical NLP pipeline,” arXiv: 1905.05950 [cs.CL], 2019.

4. D. S. Margulies, S. S. Ghosh, A. Goulas, M. Falkiewicz, J. M. Huntenburg, G. Langs, …, J. Smallwood, “Situating the default-mode network along a principal gradient of macroscale cortical organization,” Proceedings of the National Academy of Sciences, 113(44), 12574-12579, 2016.

5. 桜前線開架宣言、**山田航**、左右社、2015 年。

6. 塔、一般社団法人塔短歌会、第 63 巻第 4 号（2016 年 4 月発行）に掲載された短歌から選んだ。

7. H. Kawabata, S. Zeki, “Neural correlates of beauty,” Journal of neurophysiology 91(4), 1699-1705, 2004.