言語モデルを用いた言語刺激下の脳内情報処理解読

理学専攻・情報科学コース
2340662
佐藤 杏奈

1　はじめに

近年，機械学習技術の高度化により，大規模言語モデルなどをはじめとした深層学習モデルを用いて，ヒト脳内の情報処理プロセスの解明や脳情報解読技術を行う研究が盛んになっている．この背景を踏まえ本研究では，デコーディング分析を用いたヒト脳内の情動処理領域の解明，また，言語モデルを用いた脳データからの文の意味的再構成を目指す．

2　短歌固有の属性に対応する脳内情報表現

言語芸術を刺激として与えられた際のヒト脳内状態を解析し，文を読んだ際にヒトが感じる**詩的である**という情動について調査する．1) 詩的という情動が脳内でどう表現されるのか，2) 文に含まれるどのような特徴が詩的感覚を構成するのかについて調査する．

2.1　脳活動データ

機能的磁気共鳴画像法 (fMRI) により血中酸素濃度依存 (BOLD) 反応を取得する．日本人被験者 32 名に MRI 内で文を呈示し，その文が詩的と感じるか否かを回答させる．刺激には,『現代日本語書き言葉均衡コーパス』(BCCWJ),『桜前線開架宣言』[1],『塔』[2] から抽出された短歌 150 首と, BCCWJ に含まれる短歌と同じ 31 文字程度の平文 150 文を用いた．9 秒間 1 つの短歌または平文が呈示された後 3 秒間でそれが詩的か否かを右手に持つボタン押しで解答する．

2.2　マルチボクセルパターン分析

詩的感覚と関係のある領域を調査するため，本研究ではサーチライト解析 [3] を用いる．これは事前に決められた半径の球状サーチライトに含まれるボクセルを用いたデコーディング分析手法で，対象範囲内でサーチライトを動かしデコーディング精度を測り，各サーチライトの精度を中心のボクセルに挿入することで，刺激と関係した情報を持つ脳領域を調査する．

2.3　特徴量の設定

サーチライト解析で用いられる被説明変数を設定する．ある文がどれほど詩的であるかを表す (a) 文の詩的さ に加え，詩的感覚の構成要素の解明のために，(b) 文の珍しさ，(c) 繰り返しと韻，(d) 文法的違い の三点について特徴量の設定を行った．

(a) 文の詩的さ

感情には個人差があるとされるが，ここでは文の詩的さは MRI 実験 (2.1 節) のタスク結果に基づき，詩的と感じる (+1)，感じない (−1) の回答を集計し，全 32 名の被験者の結果からスコアを算出した．

(b) 文の珍しさ

日常会話であまり使われない語彙や表現を含む文は，文学的な趣を感じさせることがある．そこで，言語モデル GPT-NeoX を用い，文中の全トークンの交差エントロピー誤差の平均を計算し，文の珍しさと定義した。

(c) 繰り返しと韻

繰り返しや韻を用いることは，詩的な文の構築のためによく使われる技法の一つである．本研究では，撥音や促音を省略し，長音記号は直前の音を繰り返すよう前処理を行った．さらに，全てを母音に変換した上で，3 字以上で韻を踏む文字数を加算し，スコア化した．

(d) 文法的違い

俳句，短歌といった言語芸術は，字数制限のための語の調節や体言止めなどの表現技法より，特徴的な文法を持つことがある．ここでは初めに，BCCWJ,『桜前線開架宣言』,『塔』から抽出された実験で未使用の短歌と平文を使用し，短歌特有の文法的特徴を調査した．GINZA による形態素解析で，i) 各品詞の出現回数と，ii) 品詞から品詞への接続確立を全組み合わせ調査し，短歌と平文の 2 群に有意差 ($p < 1 \times 10^{-10}$) が見られた 42 個の項目を抽出した．これらの項目で文をベクトル化し，cosine 類似度を算出することで，文法的な短歌らしさをスコア化した．

2.4　実験結果

図 1 にスコア化された各要素のデコーディング精度を示す．評価には予測値と実測値のピアソン相関係数を用い，値は被験者毎に FDR 補正 ($q < 0.01$) された後全被験者で平均された．グラフは，青：(a) 詩的さ，橙：(b) 珍しさ，緑：(c) 韻，赤：(d) 文法 の平均スコアを，グレーの各点が各被験者のスコアを示す．

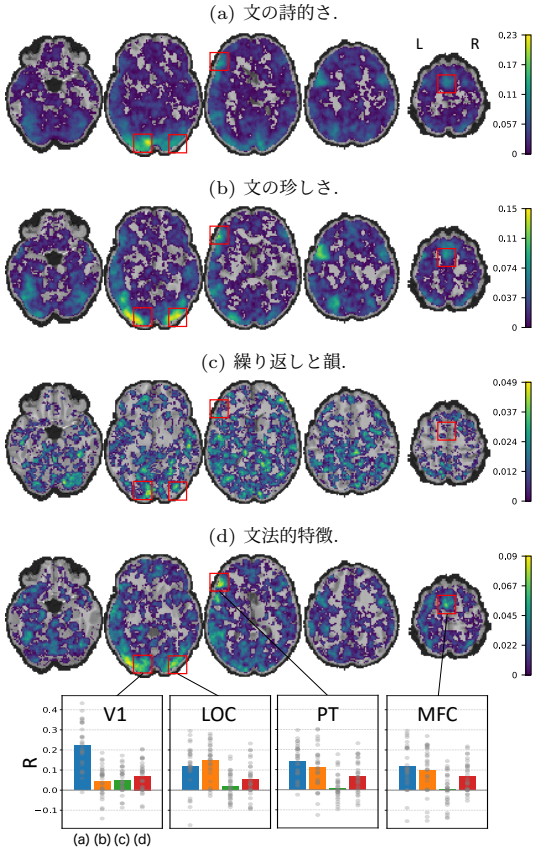

図 1: サーチライト解析結果 ($z = -20, -4, 11, 37, 59$)．グラフは各サーチライトにおけるデコード精度を示す (V1: pericalcarine cortex, LOC: lateral occipital cortex, PT: pars triangularis, MFC: middle frontal cortex).

2.5 考察

刺激文の詩的さを予測できる脳領域は大脳皮質上に広く散布しており，なかでも後頭葉やブローカ野の精度が高いとわかった (図 1(a))．また，詩的感覚の構成要素のデコーディングでは，文の珍しさ，繰り返しと韻のスコアと関係のある領域は，(a) とは独立して機能的局在している可能性を示唆した (図 1(b,c))．一方でどのスコアにおいても $R = 0.2$ 程度に収まっていることから，詩的感覚の構成要素そのもの，またそのスコア化方法は引き続き議論する必要があるだろう．

3 大規模言語モデルを用いた言語刺激下の脳内意味表象解読

本研究では，言語モデルを活用した脳内意味解読を提案した Tang ら [4](図 2) の研究の拡張を行う．より高精度な解読を目指し，先行研究で使われた Fine-tuned GPT に加え新たに 3 つの言語モデルを導入，精度の比較を行い，高精度な解読に重要な要素を探究する．

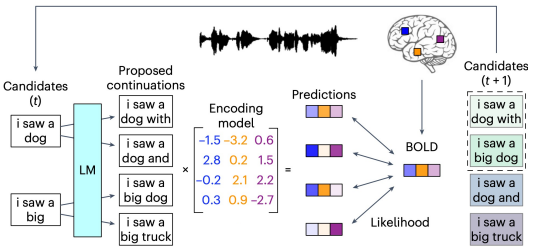

図 2: 言語刺激下の脳データによる文の再構築 ([4] より引用)．言語モデルで次に来る可能性のある単語を出力し，符号化モデル (3.3 節参照) で候補の文から誘起される脳反応を予測．候補の内，実測脳反応と近い k 個の候補が次のタイムステップへ保持される．

3.1 MRI データ

先行研究 [4] と同じデータセット [5] を使用する．MRI データは健康な大人 3 名より取得された．刺激データは，*The Moth Radio Hour* と *Modern Love* から抽出された 82 のストーリーで，各ストーリーでは一人の話し手が自伝的な物語を語る音声刺激である．

3.2 言語モデル

先行研究で使用された Fine-tuned(FT) GPT に加え，Hugging Face Hub で公開されている事前学習済モデルの，GPT，Llama3，OPT モデルを使用した (表 1)．Fine-tuned GPT は，Reddit のコメントと，MRI 実験と同じコーパスで実験では未使用の，自伝的物語のコーパスで訓練された．

表 1: 使用した 4 種の言語モデル．

Model	Size	Training Data
FT GPT	120M	Reddit posts, autobiographical stories
GPT	120M	Unpublished books in various genres
Llama3	8B	Large public text datasets
OPT	6.7B	Books, story-like data, news, web text

3.3 符号化モデル (Encoding model)

刺激単語から抽出された特徴から，正則化線形回帰で各特徴が被験者の脳活動にどのような影響を与えるかを予測する重みを学習する [6]．被験者に与えた刺激文を言語モデルの入力として与えた際の，言語モデル内の隠れ状態が，その刺激文の特徴量として抽出され，脳活動の予測に使用される．

3.4 実験結果

被験者が想起する文を再構築できているか評価するため，デコーダで再構築された文と実際の刺激文の類似度を BERTScore [7] で測る．また生成された文が有意に高いスコアを持つか確かめるため，脳活動を用いずに言語モデルに出力させた 300 文で同様に BERTScore を測り，偶然レベルを示す帰無分布を構築した．図 3 にデコーダの結果を示す．(a) は文全体の類似度を示す Story similarity であり，箱ひげ図は帰無分布，*は帰無分布より有意に高い (q(FDR)<0.05) スコアであることを示す．(b) は 20 秒の window 内における類似度を示す Window similarity であり，上部の線はそのタイムポイントにおいて帰無分布より有意に高い (q(FDR)<0.05) スコアであることを示す．

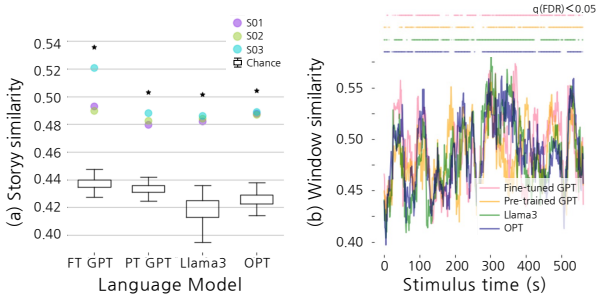

図 3: デコーダによる再構成文の結果．

3.5 考察

図 3(a) より，どの言語モデルを用いた際も有意に刺激文と類似した文を再構成できた一方で，Llama3-8B、OPT-6.7B などの大規模モデルよりも，FT GPT モデルの方が高いスコアを出す傾向を確認した．これは，FT GPT が今回のテストデータと同じコーパスで追加訓練されていたため，実際の刺激文と似ていた出力がされやすいことが原因のひとつだと考える．

4 まとめ

本研究では，言語芸術を刺激として与えられた際のヒト脳内状態を解析し，文を読んだ際にヒトが感じる詩的であるという情動について調査を行った．また，言語モデルを用いた意味解読手法の拡張を行い，その技術の有用性を確かめると同時に，より良い解読につながる要因の追求を行った．

参考文献

[1] 山田航. 桜前線開架宣言. 左右社, 2015.

[2] 塔. 第 63 巻第 4 号. 一般社団法人塔短歌会, 2016.

[3] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, Vol. 103, No. 10, pp. 3863–3868, 2006.

[4] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, Vol. 26, No. 5, pp. 858–866, 2023.

[5] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. "an fmri dataset during a passive natural language listening task", 2024.

[6] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fmri. *NeuroImage*, Vol. 56, No. 2, pp. 400–410, May 2011.

[7] Tianyi Zhang, Varsha Kishore*, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.