大規模言語モデルに失語症ニューロンは存在するか？

森田早織　原田宥都　大関洋平
東京大学
{msaori6012,harada-yuto,oseki}@g.ecc.u-tokyo.ac.jp

概要

失語症は近年、LLM によって検出や分類が可能であると示唆されているが、LLM が失語症者の言語特徴をどのような内部表現として捉えているのかについては十分に解明されていない。本研究では、LLM 内部に着目し、失語症者の発話に特異的に反応するニューロンの有無を検証した。失語症の有無および 3 種類の言語エラーの検出タスクを設定し、Average Precision に基づいて専門性の高いニューロンを抽出した。その結果、失語症全体の検出では主に最終層付近に、エラータイプ別の検出では音韻、形態、意味の各誤りで異なる層に特異的なニューロン分布が見られ、LLM が言語の異なる側面を機能的に分離して表現している可能性を示唆した。

1　はじめに

1.1　失語症研究の概観

後天的な脳疾患や外傷などで、言葉の産出・理解といった言語運用が障害されることがある。このような言語障害を失語症といい、より具体的には“大脳の損傷に由来する、一旦獲得された言語記号の操作能力の低下ないし消失”と定義される [1]。失語症は患者によって多種多様な発話特徴が見られるが、患者に見られるそれぞれの特徴が発症要因や脳部位とどのように関連し、どう分類・解釈するべきかということについて長く議論されてきた。Lichtheim[2] による提唱によって、脳血管障害によって引き起こされる失語症は Broca 失語症、Wernicke 失語症、伝導性失語症、超皮質性失語症、全失語症という５つのサブタイプ分類がなされ、これは古典分類として現代でも失語症を理解するための枠組みの一つとして使用されている。一方でこの古典分類については、例外的な発話特徴を捉えきれないのではないかという議論もされてきている [3]。

近年では、大規模言語モデル (LLM) の誕生により、LLM を用いて失語症の有無や発話特徴の検出・予測ができるという新たな可能性が指摘されてきている。Cong ら [4] は LLM から算出されるサプライザル (surprisal) 指標に着目し、失語症者の自然発話データと用いて LLM が失語症の診断・重症度評価・下位分類においてどの程度有効であるか検討し、サプライザルが失語症の有無を判別できる指標になりうると提示した。また、Rezaii ら [5] は、神経変性疾患によって引き起こされる進行性失語症を持つ話者の短い発話データから、LLM が 78%の精度でサブタイプを分類できることを示した。一方で、LLM 内部で形成される内部表現が失語症者の発話に含まれる統語的・意味的・語用論的逸脱をどのように解釈しているのかということについては十分に検討されていない。

1.2　LLM の内部表現と解釈可能性

Transformer を基盤とする言語モデルが、内部表現としてどのような言語情報を保持しているのかを明らかにする研究は近年活発に行われてきた。例えば、BERT を対象としたプロービング分析によって、言語的情報が Transformer の層に沿って段階的に符号化されていることを示されている [6, 7]。さらに、層レベルの分析に加えて、ネットワーク内のニューロン単位での分析も進められている。Durrani ら [8] は、言語現象ごとに情報の保持のされ方がネットワークの中で異なっていることを示しており、例えば文脈依存性の低い語は、少数のニューロンで予測可能である一方で、多義的な語は多数のニューロンが寄与していると言及している。さらに、Suau ら [9] は Transformer モデル内部の線形層に含まれる各ニューロンを"概念分類器”として捉え、特定の概念を高精度で識別するニューロン (expert unit) の存在を提唱した。加えて、ごく少数の expert unit を活性化させることである特定の概念を含むテキスト生成が可能になることを示し、これらのニューロンが機能的な役割を担っている可能性を示唆した。

以上のように、LLM の内部表現に着目した分析手法は、LLM が失語症者の発話に含まれる言語的逸脱や誤りをどのように捉え、どの特徴が分類や判別に寄与しているのかを解釈可能な形で検討する上で有効なアプローチであると考えられる。

1.3 本研究の目的

LLM が失語症者の発話やサブタイプを一定の精度で識別できる可能性がある中で、LLM が失語症者の言語特徴を理解する際に LLM 内部ではどのように捉えているのかということは十分に明らかにされてきていないという現状を踏まえ、本研究では以下 2 点の問いに焦点を当てる。

- LLM が失語症者の言語特徴を検出する際、その判断に強く寄与するニューロンは存在するか
- 失語症に見られる音韻障害や意味障害といった言語症状を識別する上で、LLM の内部表現にはどのような傾向が見られるか

本研究は、失語症研究への LLM の応用において、予測や分類性能の評価にとどまるのではなく、モデル内部の表現構造に踏み込んで解釈する試みである。これにより、LLM が失語症者特有の言語特徴をどのように捉えているのかを明らかにするとともに、LLM および失語症者の発話特徴の双方に対して新たな観点から理解を深める手がかりを提供することを目指す。

2 実験内容・手法

2.1 Aphasic expert neuron の検出

本研究では、Suau ら [9] の手法を踏襲し、失語症者の発話パターンを捉えることに特化したニューロン (Aphasic expert neuron) を検出する。Suau ら [9] は事前学習済み Transformer モデルの内部にある各ニューロンを二値分類器とみなし、Average Presicion (AP) によってその専門性を定量化する手法を提案した。AP は情報検索や機械学習の分野で広く用いられる評価指標 [10, 11] であり、適合率-再現率曲線 (Precision-Recall curve) の下側面積として定義されている。二値分類問題において、適合率と再現率は以下のように定義される。

- 適合率 (Precision): 正例と予測したもののうち、実際に正例であった割合
- 再現率 (Recall): 実際の正例のうち、正しく正例と予測できた割合

AP は、異なる閾値における適合率と再現率の関係を統合した指標である。具体的には、再現率を 0 から 1 まで変化させた時の適合率の平均値として計算される。

本研究では、概念 c に対する全文章 ($N_c = N_c^+ + N_c^-$ 個) に対するニューロン m の応答を $\mathbf{u}_c^m \in \mathbb{R}^{N_c}$ とし、対応するバイナリラベルを $\mathbf{b}_c \in \{0,1\}^{N_c}$ とする。各ニューロンをバイナリ分類器とみなし、$\mathbf{u}_c^m$ を予測スコアとして使用することで、AP を計算する:

$$\mathrm{AP}_c^m = \mathrm{AP}(\mathbf{u}_c^m, \mathbf{b}_c) \in [0,1] \tag{1}$$

つまり AP が高いニューロンは、特定の失語症者の言語特徴を含む文章と含まない文章を正確に識別できることを意味する。この方法により、各要素に対して AP 値が最も高い上位 100 個のニューロンを Aphasic expert neuron として特定した。

この分析を適用し、LLM が失語症者の言語的逸脱や誤りを捉える上でどのニューロンが寄与しているか、健常者の言語コーパスとともに比較し分析する。そして、失語症者によって異なる言語特徴について、本研究では音韻の誤り、意味の誤り、形態素の誤りという 3 種類の言語特徴に着目し、それらを LLM が捉える際にどのような傾向が見られるか分析する。

2.2 実験設定・モデル

本研究では、Llama-3.1-8B-Instruct を使用した。Suau ら [9] が Transformer ブロック内の複数の線形層を分析対象としたのに対し、本研究では Feed-Forward Network(FFN) が意味的・概念的な知識を保持している [12] ということを踏まえ、FFN の gate_proj 層に着目した。また、Llama アーキテクチャでは SwiGLU activation を採用しており [13]、gate_proj 層が情報の選択的処理の機能を担うことから、失語症の特徴検出に適していると考え着目した。Llama-3.1-8B-Instruct は 32 層から構成され、各層の gate_proj 層は 14,336 ユニットを持つため、計 458,752 個のニューロンを分析対象とした。

2.3 使用したコーパス

Suau ら [9] の手法では、ある特定の概念を含む文章と含まない文章の集合を用いて Expert neuron の検出を試みている。本研究ではこの枠組みを"失語症者の発話を含むか含まないか"、"ある特定の誤りを含むか含まないか"という 2 段階を設定し、実験

を行った。失語症者および健常者のデータについては、AphasiaBank コーパス [14] にて公開されている英語が母語である失語症者 954 名、同じく英語が母語である健常者 479 名の発話データを使用した。AphasiaBank コーパスには、音韻や意味、形態素などといった言語特徴の誤りを細かく示した独自のエラータグが事前に付与されている。本研究では、前処理としてこれらのエラータグを各発話におけるエラータイプのメタデータとして使用した。また、フィラーや言い淀み、繰り返しなどといった発話中に現れた特徴は、(1) 該当箇所の発話を削除する (2) 該当箇所の発話を残す (3) 発話を残した上でどのような言語特徴なのかアノテーションする、という 3 つの条件を設定し実験を行った。以下の条件ごとに正例と負例に分け、それぞれ 1,000 サンプルずつ取り出して実験を行った。

- **実験 A**：失語症者のテキスト（正例）/ 健常者のテキスト（負例）
- **実験 B-1**：音韻の誤りを含むテキスト（正例）/ 含まないテキスト（負例）
- **実験 B-2**：形態素の誤りを含むテキスト（正例）/ 含まないテキスト（負例）
- **実験 B-3**：意味の誤りを含むテキスト（正例）/ 含まないテキスト（負例）

各サンプルを Llama-3.1-8B-Insturuct に入力し、FFN 層の gate_proj における活性化値を抽出することで、特定の要素に対して活性化する Aphasic expert neuron の分布を分析した。

3 結果・分析

3.1 実験 A: 失語症的特徴の検出

3.1.1 実験結果

図 1 に、実験 A における失語症検出の精度を示す。本実験ではサンプル数が正例と負例ともに同数であることを踏まえると、3 つのパターン全てにおいて Max AP>0.5 を記録したことから、LLM は失語症者の発話と健常者の発話を有意に識別可能であると示された。

特にパターン (2) では Max AP が 0.710 に達し、3 つのパターンの中で最も高い識別性能を示した。パターン (2) は全ての発話の記録を細かく残していたが、フィラーや言い淀みなどをアノテーションしていたパターン (3) よりも識別性能が高いという結果となった。一方、パターン (1) の Max AP は 0.657 と

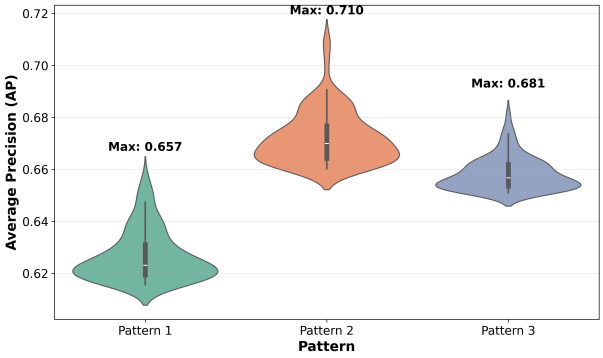

図 1　各パターンの AP スコア

低く、パターン間で検出難易度に差があることが確認された。

さらに、上位 100 個の Aphasic expert neuron の層ごとの分布を図 2 に示す。

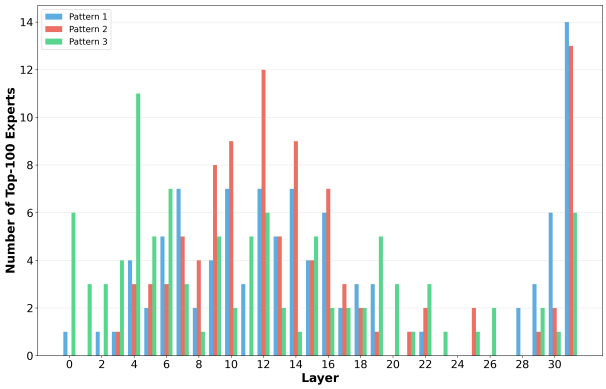

図 2　層ごとの Aphasic expert neuron の分布

まず 3 つのパターンに共通して、最終層付近である 32 層目に Aphasic expert neuron が集中する傾向が見られた。特にパターン 1(青) とパターン 2(赤) では 32 層目に全体の 13-14%の Aphasic expert neuron が検出された。また中間層でも Aphasic expert neuron の集積が確認され、失語症の検出が単一の処理段階ではなく、複数の処理段階において並行的に行われている可能性がある。一方で 20-29 層目はいずれのパターンでも想定的に Aphasic expert neuron の数は少なく、これらの層は失語症検出にあまり寄与していないことが示された。

3.1.2 分析

3 つのパターンにおける上位 100 個の Aphasic expert neuron の重複を分析するため、Jaccard 係数を用いて調べた。Jaccard 係数は 0.081 から 0.282 の間にあり、パターン間での重複は限定的であることが示された。特にパターン (1) と (2) の間で最も高い重複 (Jaccard 係数=0.282, 44 個の共通 Aphasic expert

neuron) が観察された一方、パターン (3) のみ他のパターンとの重複は低かった (Jaccard 係数<0.1)。

3.2 実験 B: 言語エラーの検出

3.2.1 実験結果

実験 B では、失語症者のコーパスのみを用い、特定のエラータイプを含むテキストと含まないテキストを識別させた。図 3 は各エラータイプごとの AP を示している。Max AP は形態素の誤りで 0.924 と最も高く、音韻の誤りで 0.826、意味の誤りで 0.782 と続いた。特に形態素の誤りについては、ニューロンがこの誤りを非常に明確に捉えられていることが示唆された。

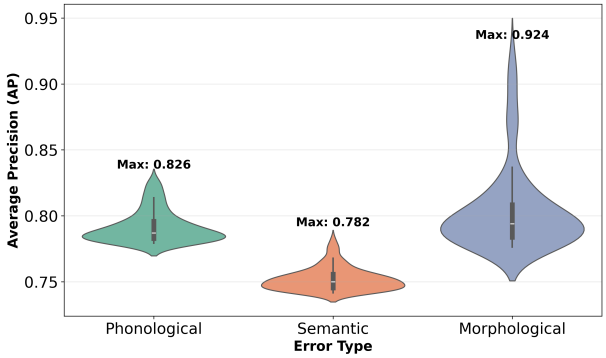

図 3 エラータイプごとの AP

さらに、図 4 にて各エラータイプの上位 100 個の Aphasic expert neuron の層分布を示す。

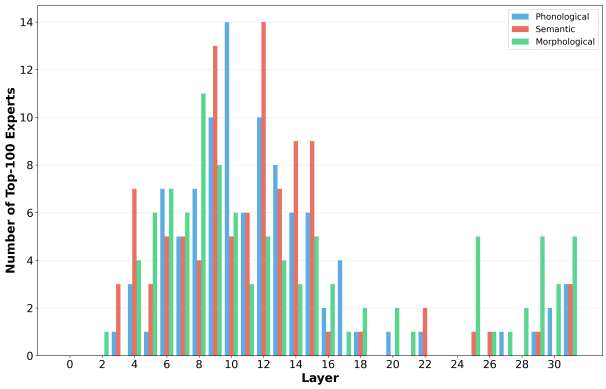

図 4 層ごとの Aphasic expert neuron の分布

3 種類の誤り全てで層分布は有意に非一様な傾向にあることが確認された。また、各エラータイプで最も多くの Aphasic expert neuron が検出される層が異なっており、音韻の誤りでは 10 層目、意味の誤りでは 12 層目、形態素の誤りは 8 層目に集中していた。この傾向は、実験 A で観察された 32 層目への極端な集積とは明確に異なっている。

3.2.2 分析

Kruskal-Wallis 検定により、3 つのエラータイプ間で AP 値の分布に有意な差があることが確認された ($H = 200.16$, $p < 0.001$)。また、Bonferroni 補正を適用した Mann-Whitney U 検定により、全てのペア間で有意差が認められた (全て $p < 0.05$)。特に、音韻エラーと意味エラー間、および意味エラーと形態エラー間では極めて顕著な差が観察された (両方とも $p < 0.001$, $|\text{Cliff's } \delta| > 0.99$)。

3.3 考察

実験 A で提示した失語症全体の検出では、Aphasic expert neuron が 32 層目に最も集中していたが、これは包括的な言語障害の判別には、主に深い層での高次の意味表現や文脈統合が重要になりうるということを示唆している。加えて、4 層目や 10 層目、12 層目といった中間層にも Aphasic expert neuron が分布しており、失語症か否かの識別には複数の段階にわたる統合的な処理を必要とすると考えられる。実験 B では各エラータイプで Aphasic expert neuron が分布する層が明確に異なることが明らかとなった。これは言語処理の階層性を反映していると解釈できる。また、形態処理が比較的浅い層で行われるのに対し、意味処理はより深い層を必要とするということは、意味的な逸脱の判別には文脈全体の統合が必要であることを示唆している。

また、本研究は 3 種類のエラータイプがそれぞれ異なるニューロンによって処理されている可能性を提示した。各エラータイプを識別する上位 100 個の Expert neuron は重複が 0 に近く (平均 Jaccard 係数 =0.014) であり、3 つのエラータイプ全てに共通するニューロンは存在しなかった。これは実験 A の結果とは対照的であるとともに、LLM が言語の異なる側面 (音韻、形態、意味) を独立した表現空間で処理している可能性を示唆している。

4 まとめ

本研究では、LLM が失語症者の言語特徴を捉える際に強く活性化するニューロン (Aphasic expert neuron) が現れるかどうか分析した。失語症全体の検出では、Aphasic expert neuron が主に深い層 (32 層目) に集中した。一方、エラータイプ特異的な検出では、形態エラー、音韻エラー、意味エラーそれぞれで別の層に Aphasic expert neuron が集中しており、処理階層が異なることが示唆された。

謝辞

本研究は，JSPS 科研費 JP24H00087，JST さきがけ JPMJPR21C2，JST CREST JPMJCR2565，JST BOOST JPMJBY24B2 の支援を受けたものです．

参考文献

[1] 山鳥重. 神経心理学入門. 医学書院, 1985.
[2] L. Lichtheim. On aphasia. **Brain**, Vol. 7, No. 4, pp. 433–484, January 1885.
[3] Myrna F. Schwartz. What the classical aphasia categories can't do for us, and why. **Brain and Language**, Vol. 21, No. 1, pp. 3–8, 1984.
[4] Y. Cong, A. N. LaCroix, and J. Lee. Clinical efficacy of pre-trained large language models through the lens of aphasia. **Scientific Reports**, Vol. 14, p. 15573, 2024.
[5] N. Rezaii, D. Hochberg, M. Quimby, B. Wong, M. Brickhouse, A. Touroutoglou, B. C. Dickerson, and P. Wolff. Artificial intelligence classifies primary progressive aphasia from connected speech. **Brain**, Vol. 147, No. 9, pp. 3070–3082, 2024.
[6] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.
[7] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline, 2019.
[8] Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models, 2020.
[9] Xavier Suau, Luca Zappella, and Nicholas Apostoloff. Finding experts in transformer models, 2020.
[10] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In **Proceedings of the 23rd international conference on Machine learning**, pp. 233–240. ACM, 2006.
[11] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. **Introduction to Information Retrieval**. Cambridge University Press, Cambridge, UK, 2008.
[12] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
[13] Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ying Shan, and Ping Luo. Llama pro: Progressive llama with block expansion, 2024.
[14] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. **Journal of Speech, Language, and Hearing Research**, Vol. 54, No. 5, pp. 1286–1296, 2011.