スパースコーディングを用いた脳内意味表象推定における BERTの有効性の検証

島　百子 †　　尾崎 花奈 ‡　　小林 一郎 †
† お茶の水女子大学 理学部 情報科学科
‡ お茶の水女子大学大学院 理学専攻 情報科学コース
{g1520517, ozaki.kana, koba}@is.ocha.ac.jp

1　はじめに

Mitchel ら [6] によって言語と脳活動との対応関係を回帰モデルとして捉える手法が提案されて以来、彼の研究アプローチは計算神経言語学と称され、その後も Naselaris ら [7] や Bullinaria ら [1] などヒトの言語活動に対する脳活動の関係を明らかにする研究が進められてきた．このような背景に対して，近年，脳神経科学において自然言語処理技術を導入し，脳活動の解析を行うアプローチが盛んになってきている．Hale ら [3] は、文法解析の曖昧性を beam search で表現したものが実際の脳活動データを説明可能であることを示した．また、Jat ら [4] は、BERT（Bidirectional Encoder Representations from Transformers） [2] によって表現された文がその文を被験者に聞かせた際の脳活動データ (MEG) と強い相関があることを示している．一方、Kawase ら [5, 13, 11, 12] は、脳内において表現される意味表象と脳内活動における情報処理にはスパースコーディングの原理が働いているとの仮説の下、その有意性の検証を行った．Ozaki ら [10] は、word2vec を用いた単語の加法性に基づく言語表象と脳活動データとの同期がとれた行列を辞書学習し、得られた辞書基底を分析することにより、脳活動の特定の単位（辞書基底に相当）に対して、word2vec 空間の言葉を割り当てることを試みている．本研究は、Ozaki らの解析において、BERT を適用した文表象（説明を後述）を用いることで、スパースコーディングによる解析において word2vec と BERT の性能を比較、調査する．

2　実験概要

2.1　データ

使用するデータは，Nishimoto ら [9] が使用したものと同じ動画視聴時の脳活動データと動画説明文である．被験者 A，B，C の併せて 3 人分の脳活動データを用いた．被験者 A と B は訓練データが 4500 サンプル，テストデータが 300 サンプル，C は訓練データが 9000 サンプル，テストデータが 600 サンプルである．脳活動データは、functional MRI(fMRI) を用いて動画視聴時の脳神経活動をボクセル数×サンプル数で記録したものであり，被験者 A と B については 2 秒に 1 サンプル，被験者 C については 1 秒に 1 サンプル記録している。ただし、スパースコーディングを適用するにはボクセル数が膨大なため、二段階で次元削減を行った．まず、解剖学的な見地からの関心領域（ROI）に基づき、全脳から大脳皮質領域のみを取り出した．二段階目として、Nishida ら [8] は word2vec を用いた脳活動の推定モデルを構築し、ボクセルごとにピアソン相関係数を用いた推定精度を示していることから、この値を参考に閾値を設定し閾値以上の推定精度を持つボクセルを抽出した．また、動画説明文は、被験者に見せた動画像から 1 秒ごとに抽出した静止画像に対し、アノテータが想起したことを文章にしたものである．アノテータ 40 人のうちランダムに抽出された 4 人の文章を合わせて動画 1 サンプルに対する説明文としている．動画視聴時の脳活動データとその動画の説明文に対して、同期をとったペアデータ作成し学習を行う．また、本来、脳神経科学の分野では脳内に持つ意味の情報を総称して「意味表象」という用語を使用するが、本研究においては、脳活動データから推定される言語の情報を総称して「意味表象」と呼ぶ．とくに、word2vec によって表現される言語の意味の情報

として「単語表象」、BERT によって表現される意味の情報を「文表象」と定義する．

2.2 推定方法

図 1 に意味表象推定方法についての概要を示す．

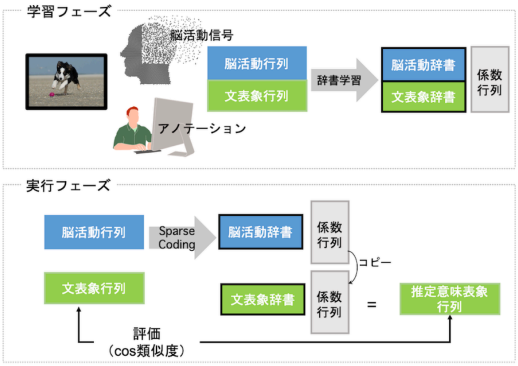

図 1: 意味表象推定方法の概要

以下に，上図で示された概要における各部を説明する．

2.2.1 BERT を用いた辞書学習

まず、訓練用脳活動データとそれに対応する言語データの結合行列を辞書学習し、両データが紐づいた辞書を作成する．学習には Lasso (Least Absolute Shrinkage and Selection Operator) の LARS アルゴリズムを用いる．脳活動データは fMRI で記録した神経信号の値をサンプルごとに 1 列に並べ行列化した．このとき、先行研究 [10] では予測精度 0.55 以上のボクセルのみを利用しているが、本研究では文表象行列の次元数が先行研究で使用されていた 300 から 768 に増えることを考慮し、予測精度の閾値を 0.5 とすることで脳活動データの次元数を増やした。また、言語データについて、川瀬らはサンプル中の名詞・動詞・形容詞に属する単語を skip-gram モデルにおいて日本語ウェブコーパス (NWJC) で学習された 300 次元の分散表現ベクトルを用いて表現し、それらの平均を 1 サンプルのベクトルとしている．本研究では、言語データの表象において言語学習モデル BERT を利用した．本モデルは双方向学習による文脈を捉えた特徴表現抽出を行っており、様々な言語学習タスクにおいて精度向上が報告されている．特に文単位で異なる意味空間を作るため、同じ単語であっても語義に応じてそれぞれ分散表現をもつ．京都大学黒橋・河原研究室が公開している[1]BERT の Whole Word Masking 版日本語事前学習モデル (12-layer, 768-hidden, 12-heads) を用いて、アノテーションデータ 1 サンプル分を 1 シーケンスとして学習し、抽出した 768 次元のベクトルをサンプル数分並べ行列化した．図 2 に言語データの表象方法について先行研究 [10] との比較を示す．

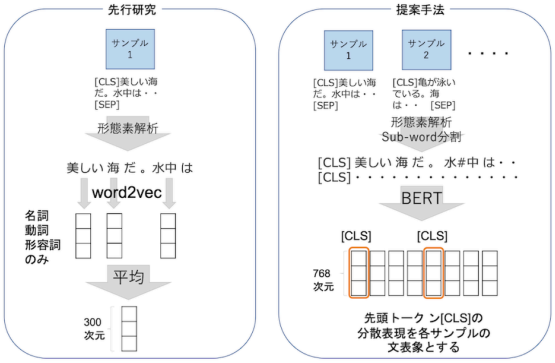

図 2: 言語データの表象方法の比較

最後に、作成した脳活動行列と文表象行列の結合行列を辞書学習する．その際、被験者が動画を見てから fMRI で観測される脳活動に影響が出るまでに生じる時間のずれを考慮し，脳活動データと文表象データを 4 秒または 6 秒ずらして対応づけた．学習により得られた辞書行列は、脳活動と文表象の特徴表現が 1 列になった基底で成り立っており、係数行列は両データで共通である．なお、基底数の設定についてはデータの次元 $\leq$ 基底数 $<$ サンプル数という制約を満たした上で、基底数をできるだけ小さくし、各基底が保持する情報量が多くなるようにすることでスパース性を確保している．また、学習時間を削減する目的と、動画では同様のシーンが数秒続くことを踏まえ、サンプルを数枚に 1 枚間引きしての学習も行った．上記の実験設定については下の表 1 の通りである．参考のため、先行研究での実験設定についても表 2 に示す．

2.2.2 スパースコーディングによる意味表象推定

作成した辞書を用いて脳活動データをスパースコーディングし意味表象を推定する．テスト用脳活動データ 300 または 600 サンプルを訓練データと同様の方法で行列化し、辞書学習で獲得した脳活動辞書行列

[1]http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT 日本語 Pretrained モデル

表 1: データの次元と基底数 (本研究)

被験者	ボクセル数		文表象 (BERT) の次元	結合行列の次元	基底数
	大脳皮質領域	予測精度 0.5 以上			
A	65665	951	768	1719	1800
B	68942	835	768	1603	1700
C	70933	1255	768	2023	2100

表 2: データの次元と基底数 (先行研究 [10])

被験者	ボクセル数		文表象 (word2vec) の次元	結合行列の次元	基底数
	大脳皮質領域	予測精度 0.5 以上			
A	65665	481	300	781	800
B	68942	565	300	865	900
C	70933	782	300	1082	1100

表 3: スパースコーディングにより得られた意味表象行列の推定精度

被験者	サンプル数 訓練/テスト	間引き数	cos 類似度			
			word2vec		BERT	
			刺激と脳活動の時間差		刺激と脳活動の時間差	
			4sec	6sec	4sec	6sec
A	4500/300	1/2	0.138	0.138	0.396	0.384
		1/3	0.143	0.106	0.384	0.355
B	4500/300	1/2	0.695	0.650	0.549	0.587
		1/3	0.482	0.409	0.354	0.278
C	9000/600	1/4	0.187	0.210	0.220	0.177

を用いてスパースコーディングを行った．導出された係数行列と文表象辞書行列により得られる行列を推定意味表象行列とする．

2.2.3 推定精度の評価

先行研究に倣い推定精度の評価には cos 類似度を用いる．テスト用脳活動データに対応する文表象を学習時と同様に行列化し、正解行列とした．推定意味表象行列との cos 類似度をサンプルごと、つまり 1 列ごとに算出し、マクロ平均をとったものを全体の精度としている．

3 結果

3.1 推定意味表象行列の精度

推定行列と正解行列の cos 類似度のマクロ平均を、間引き数・動画と脳活動の観測時間差ごとに表 3 に示す．表から、被験者により精度の増減が観察された．

4 考察

結果の要因として、word2vec による単語表象を用いた Ozaki ら [10] の先行研究に比べ、推定精度の低いボクセルも脳活動データに含めたことが考えられる．また、以下の問題点が挙げられる．第一に、動画における場面の切り替わりを、BERT が文脈を捉える際に利用する文と文の区切りとして扱っている点である．実験で使用した動画は複数の動画クリップを結合したものであるため 1 サンプル前の動画と内容の変化が大きい場合があり、文表象が文脈を捉えたものになっていないと考えられる．さらに、画像サンプル 1 つに対して複数のアノテータが記述した文を集めて 1 シーケンス（つまり一つの文とみなして）BERT で学習したため、アノテータが変わるタイミングで文脈が切り替わることも要因と考えられる．

5 結論

Jat ら [4] は BERT で学習された文表象と脳活動の相関性を示したが、本研究のスパースコーディングを対象にした脳活動の分析においては BERT の有用性を示すことはできなかった．ただし、脳活動に紐付けて学習する文表象について、1 サンプルに対するアノテーションを 1 シーケンスとして BERT に学習させた点や、複数のアノテータの文章を一つにまとめたことで文脈の成立が疑わしいシーケンスとなってしまった点など実験設定に検討の余地があり、BERT の有用性を否定することはできないと考える．今後、1 サンプルに対して一人分のアノテーションもしくは 1 文のみを抽出した文表象を作成することでサンプルのカテゴリを明確にするなど設定を見直して実験を行いたい．

謝辞

本研究を進めるにあたり、科研費（18K19805）からの支援を受けた．また、本研究の実験で使用したデータを NICT CiNet の西本伸志氏より提供を受けると共に、大阪大学山口裕人氏からも有益な助言を頂いた．ここに、深く感謝の意を表す．

参考文献

[1] John A. Bullinaria and Joseph P. Levy. Limiting factors for mapping corpus-based semantic representations to brain activity. *PLOS ONE*, Vol. 8(3), p. e57191, 2013.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2727–2736, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[4] Sharmistha Jat, Hao Tang, Partha Talukdar, and Tom Mitchell. Relating simple sentence representations in deep neural networks and the brain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5137–5154, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Chiaki Kawase, Ichiro Kobayashi, Shinji Nishimoto, Hideki Asoh, and Satoshi Nishida. Semantic representation in the cerebral cortex with sparse coding. *2017 IEEE-SMC*.

[6] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Just MA. Predicting human brain activity associated with the meanings of nouns. *Science*, Vol. 30, pp. 1191–1195, 2008.

[7] Thomas Naselaris, Ryan J Prenger, Kay Kendrick, and Jack Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, Vol. 63, pp. 902–915, 2009.

[8] Satoshi Nishida, Alexander G. Huth, Jack L. Gallant, and Shinji Nishimoto. Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions. *Society Neuroscience Abstract*, Vol. 45, p. 333.13, 2015.

[9] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, Vol. 21, pp. 1641–1646, 2011.

[10] Kana Ozaki, Satoshi Nishida, Shinji Nishimoto, Hideki Asoh, and Ichiro Kobayashi. Analysis of correspondence relationship between brain activity and semantic representation. *2019 Conference on Cognitive Computational Neuroscience*.

[11] 川瀬千晶, 小林一郎, 西本伸志, 西田知史, 麻生英樹. スパースコーディングを用いた脳活動の意味表象推定に関する精度向上への取り組み. 人工知能学会全国大会論文集, Vol. JSAI2017, , 2017.

[12] 川瀬千晶, 小林一郎, 西本伸志, 西田知史, 麻生英樹. 脳活動と分散表現による意味表象へのスパースコーディング適用により獲得された辞書基底の分析. 人工知能学会全国大会論文集, Vol. JSAI2018, , 2018.

[13] 川瀬千晶, 小林一郎, 西本伸志, 麻生英樹, 西田知史. 行列因子分解を用いた動画刺激による脳活動データからの言語表象推定への取り組み. 人工知能学会全国大会論文集, Vol. JSAI2016, , 2016.