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　(指導教員：小林一郎)

1　はじめに

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いる．Jat ら [1] は，BERT（Bidirectional Encoder Representations from Transformers）[2] により表現された文がその文を聴いた被験者の脳活動データ (MEG) と強い相関があることを示している．一方，Kawase ら [3] は，動画を視聴した際の脳活動データと，動画説明文を word2vec によって表現したベクトル（本研究では「単語表象行列」と呼ぶ）との対応関係をとった結合行列に対してスパースコーディングによる辞書学習を行い，その辞書を用いて脳活動データの行列から脳内の意味表象行列を推定した．その結果，直接脳活動行列から単語表象行列への Ridge 回帰を推定した意味表象行列に比べ精度が高かったことから，脳内意味表象におけるスパースコーディングの原理の成立を仮説している．本研究では，言語情報の表象を word2vec から BERT に変更したもの（本研究では「文表象行列」と呼ぶ) を利用し，スパースコーディングによる解析における word2vec と BERT の性能を比較，調査する．

2　意味表象推定実験

一般に，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意味情報として「単語表象」，BERT によって表現される意味情報を「文表象」と呼ぶ．

2.1　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．

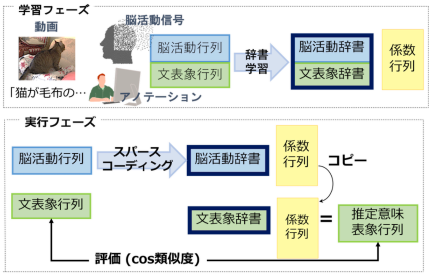

図 1: 意味表象推定方法の概要

2.2　脳活動データと言語データ

使用するデータは，VideoBlocks [5](以下 VB) および NishidaVimeo [6](以下 NV) データセットの動画視聴時の脳活動データと動画説明文である．VB データは動画の中心を注視するよう指示された被験者 A，B，C の 3 人分の脳活動データを保持し，訓練データが A と B は 4500 サンプル，C は 9000 サンプル，テストデータが A と B は 300 サンプル，C は 600 サンプルである．一方，NV データは視線を自由にした状態で計測した被験者 D，E，2 人分のデータを使用し、両者とも訓練データが 7200 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，fMRI(functional MRI) を用いて神経活動と相関があるとされる BOLD(Blood Oxygenation Level Dependent) 信号をボクセル数×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，その他の被験者については 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 [6] 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，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．アノテータは被験者とは別に用意した 40 人 (VB) または 48 人 (NV) で、このうちランダムに抽出された 4 人 (VB) または 4~7 人 (NV) 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．ただし，NV データのテストデータについては，動画視聴後の被験者がアノテーションを作成し，2 人分の文章を 1 サンプルの動画説明文としている．

2.3　BERT を用いた辞書学習

まず、訓練用脳活動データと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 BOLD 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 [4] 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，本研究では言語データの分散表現が 300 次元から 768 次元に増えることを考慮し，脳活動データも同程度の次元数になるよう閾値を 0.48~0.57 に設定した．また，言語データについて川瀬らは word2vec を用いて出現単語を 300 次元の分散表現ベクトルで表現し、その平均を 1 サンプルのベクトルとしている．本研究では，言語情報の表象において言語学習モデル BERT（Bidirectional Encoder Representations from Transformers）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，多義語に対応できるという特徴をもつ．京都大学黒橋・河原研究室が公開している 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 [4] との比較を示す．

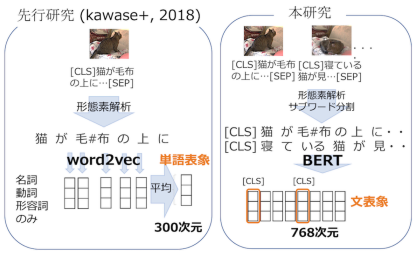

図 2: 言語データの表象方法の比較

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 が脳活動を観測するまでに生じる時間のずれを考慮し，脳活動と文表象を 4 秒または 6 秒ずらして対応づけた．学習で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することでスパース性を確保している．上記の実験設定については下の表 1 の通りである．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 枚間引きして学習を行った．

表 1 データの次元と基底数

被験者	ボクセル数		結合行列の次元	基底数
	大脳皮質領域	削減後 (閾値)		
A	65665	951 (0.50)	1719	1800
B	68942	835 (0.50)	1603	1700
C	70933	1255 (0.50)	2023	2100
D	70933	929 (0.57)	1607	1700
E	61127	1006 (0.48)	1774	1800

2.4 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 サンプル (被験者 A,B) または 600 サンプル (被験者 c) 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．

2.5 推定精度の評価

先行研究 [3] 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の動画説明文を学習時と同様に行列化し，正解行列とした．正解行列と推定意味表象行列の cos 類似度をサンプルごとに算出し，そのマクロ平均を全体の精度としている．

3 結果と考察

3.1 実験結果

推定精度を表 2 に示す．word2vec を用いた実験結果と比較すると精度の増減にばらつきがあった．

3.2 考察

精度が向上しなかった要因として，1 サンプルのアノテーションが複数人の文章を並べたもので作成されているため，シーケンス内で文脈を捉えられなかった可能性がある．そこで各サンプルのアノテーションをアノテータ 1 人の文章のみで作成し同様に実験を行なった．結果は表 3 の通りアノテータが複数の場合の方が精度が高く，アノテータの切り替わりによる文脈への影響よりも情報量の多さの方が推定精度に貢献すると考えられる．

表 2 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
		1/6	0.134	0.131	0.124	0.157
D	7200/600	1/3	0.152	0.139	0.155	0.184
E	7200/600	1/3	0.146	0.144	0.110	0.147

表 3 1 サンプル中のアノテータの人数別の推定精度

被験者	サンプル数 訓練/テスト	間引き数	cos 類似度			
			刺激と脳活動の時間差			
			4sec		6sec	
			アノテータ 4 人	アノテータ 1 人	アノテータ 4 人	アノテータ 1 人
A	4500/300	1/2	0.396	0.386	0.384	0.342
B	4500/300	1/2	0.549	0.541	0.587	0.569
C	9000/600	1/4	0.220	0.206	0.177	0.175

4 まとめと今後の課題

Jat ら [1] 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用いた脳活動の分析においては BERT の有用性を明確に示すことはできなかった．要因として，単語表象を用いた Ozaki ら [4] 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ためノイズが増えたことが考えられる．また追実験から 1 シーケンス内の厳密な文脈に依るものであるとは言い難いが，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同士の区切りとして扱っており，1 サンプル前の動画と内容の変化が大きい場合，文表象が文脈を捉えたものになっていない可能性が挙げられる．このように実験設定に検討の余地があり，BERT の有用性は否定できないと考える．今後は自然な文脈が成立するようデータを抽出するなど設定を見直して実験を行いたい．

参考文献

[1] Jat, Sharmistha and Tang, Hao and Talukdar, Partha and Mitchell, Tom. (2019) Relating Simple Sentence Representations in Deep Neural Networks and the Brain. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5137-5154, Florence, Italy, July

[2] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. (2019) 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), pages 4171-4186

[3] 川瀬 千晶 and 小林 一郎 and 西本 伸志 and 西田 知史 and 麻生 英樹. (2018) 脳活動データからのスパースコーディングによる意味表象推定と基底の分析. 言語処理学会 2018 論文集, pages 133-135

[4] Kana Ozaki and Satoshi Nishida and Shinji Nishimoto and Hideki Asoh and Ichiro Kobayashi. (2019) Analysis of Correspondence Relationship between Brain Activity and Semantic Representation. In Proceedings of the 2019 Conference on Cognitive Computational Neuroscience

[5] Shinji Nishimoto and An T. Vu and Thomas Naselaris and Yuval Benjamini and Bin Yu and Jack L. Gallant. (2011) Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. The journal of Current Biology, volume 21, pages 1641-1646

[6] Satoshi Nishida, Shinji Nishimoto. (2018) Decoding naturalistic experiences from human brain activity via distributed representations of words. The journal of NeuroImage, volume 180, part A, pages 232-242