言語モデルの内部表現における文法情報の局所性について

佐藤宏亮 [1]　鴨田豪 [1]　Benjamin Heinzerling[2,1]　坂口慶祐 [1,2]
[1] 東北大学　[2] 理化学研究所
sato.kosuke.s7@dc.tohoku.ac.jp　go.kamoda@dc.tohoku.ac.jp
benjamin.heinzerling@riken.jp　keisuke.sakaguchi@tohoku.ac.jp

概要

言語モデルにおいて，注目する知識をエンコードするニューロンの特定は，モデルの解釈性向上や少数のニューロンの操作による出力の制御の実現に貢献し得る．本研究では，スパースプローブを用いて，隠れ状態における文法情報の局在性を調査した．結果，大部分の層において，隠れ状態の特定の 1 %のニューロンはその補集合からランダムにサンプリングされた 6 倍以上の数のニューロンに匹敵する文法情報を有することが明らかになり，各文法情報が隠れ状態に局在的にエンコードされることが示された．また，一部の文法では，学習したプローブが他の文法に汎化することが明らかになった．

1　はじめに

言語モデル（LM）の内部表現内において，注目する知識をエンコードするニューロンの特定は，モデル解釈を容易にするだけでなく，少数ニューロンへの操作によるモデルの制御に寄与する．本研究では分析対象の知識として，文法に関する知識を扱う．

Transformer ベースの LM の文法に関する研究は盛んに行われてきた [1–6]．Clark ら [1] は特定の注意ヘッドが文法的概念に対応することを示した．Gurnee ら [2] は MLP 層で動詞の活用に反応するニューロンの存在を示した．Hennigen ら [4] は BERT [7] の文脈化された単語埋め込みの少数のニューロンに，単語レベルの文法情報がエンコードされることを示した．He ら [5, 6] は入力全体の文法情報が隠れ状態にエンコードされることを示した．

本研究では，モデルの隠れ状態に対して文法性判定を行う**疎な**プローブを学習し，文全体の文法情報が隠れ状態の少数のニューロンにエンコードされるかを調査する．使用するニューロンの選択には，学習に基づくものと Welch の t 検定に基づく手法の 2 つを採用した．結果，どちらの選択手法でも大部分の層において，25%以下のニューロンを用いたプローブから，全ニューロン使用時と比較して 95%以上の正解率をが得られること，選択されたニューロンは選択手法によらないことがわかった．

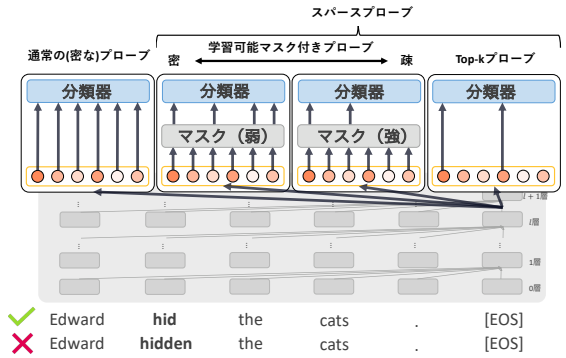

図 1　スパースプロービングの概要を示す．プローブにマスクを付与することでプローブで使用するニューロンの数を削減し，少数のニューロンから文法性判定を行う．少数のニューロンから文法性判定可能である場合，選択された少数のニューロンのみで文法性判定に必要な情報を表現できることを示している．

さらに，全体の 1%に当たる 40 個の選択されたニューロンを用いたスパースプローブと，その補集合からランダムにサンプリングした k 個のニューロンを用いたスパースプローブを比較することで，選択された 40 個のニューロンが有する文法情報を定量的に評価する．その結果，大部分の層において 6 倍以上の数のニューロンを用いたプローブに匹敵する正解率が得られ，このことから，文法情報を多く有する文法ニューロンが存在することが示された．

最後に，特定の文法についての文法性判定で学習されたスパースプローブを用いた異なる文法に関する文法性判定可能性を調査した．結果，汎化するケースが存在することが明らかになった．

これらの結果は，LM の解釈性や制御性の向上につながるだけでなく，低コストな文法性判定器の実現に寄与するものである．

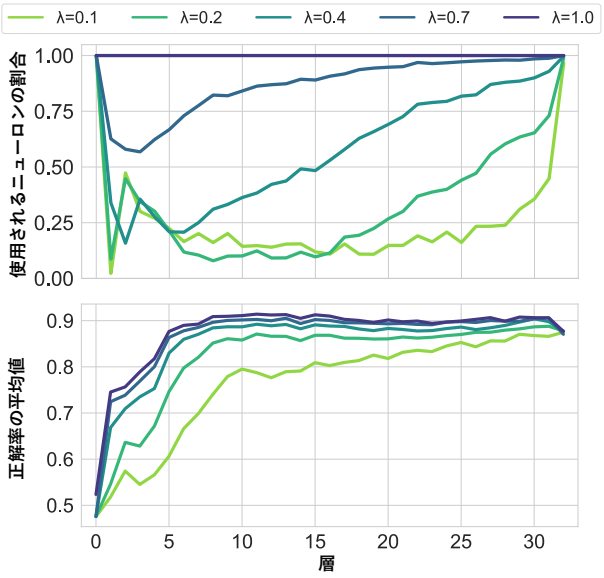

図 2　異なる粗密制御パラメータ λ で学習された学習可能マスク付きプローブにおいて，文法性判定に用いられた隠れ状態内のニューロンの割合を上に示す．また，これらのプローブで文法性を判定した際の正解率を下に示す．少ないニューロンを用いたプローブで文法性判定を行っても，高い正解率を示すことが明らかになった．

2　少数ニューロンでの文法性判定

本節では，入力の文法性を隠れ状態内のどれだけ少数のニューロンから判定することができるかを調査する．

2.1　実験方法

隠れ状態 $x \in \mathbb{R}^{d\times 1}$ 内の少数のニューロンから文法性判定を行うために，マスク付きのプローブを学習するスパースプロービング [8] の技術を用いる。

学習可能マスク付きプローブ　少数のニューロンから文法性判定を行うため，学習可能マスク付きプローブを，重み $w \in \mathbb{R}^{1\times d}$ を持つ線形プローブに対してマスク $m \in \mathbb{R}^{1\times d}$ を付与したもので定義する:

$$\text{Probe(x)} := \text{Sigmoid}\left((w \odot m)x\right) \tag{1}$$

マスクが疎に学習されるように Louizos ら [9] によって提案された L_0 正規化手法を採用し，以下の損失関数を用いて学習を行う:

$$\text{Loss} = \lambda \text{Loss}_{\text{BCE}}(w, m) + (1-\lambda)R(m) \tag{2}$$

ここで，プローブの正解率を上げるためのバイナリクロスエントロピー (BCE) ロス，R は L_0 正則化項を表し，λ はマスクの粗密を制御するハイパーパラメータである。これにより，m の各要素は $[0, 1]$ の連続区間上の値をとるよう学習される．

表 1　BLiMP に収録された文ペアの例

タスク	文
intransitive	[正] Some glaciers are vaporizing.
	[誤] Some glaciers are scaring.
tough vs raising 1	[正] Julia wasn't fun to talk to.
	[誤] Julia wasn't unlikely to talk to.

データセット: BLiMP　BLiMP データセット [10] は 67 のタスクで構成され，各タスクには特定の言語学的知識に関する 1,000 組の正しい文と誤った文が収録されている (表 1).

モデル　事前学習済みの Llama3-8B [11] を用いた．プローブは，Wang ら [12] に倣って文末に付与した EOS トークンの隠れ状態に対して学習した．

2.2　結果

図 2 に各 λ に対応するプローブの分類精度と，Llama3-8B の隠れ状態の 4,096 個のニューロンの何%をプローブに使用したかの割合を示す．全てのニューロンを用いたプローブが文法的正誤を分離できている層においては，使用するニューロンが少ないプローブでも入力の文法性判定が可能であることが明らかになった．10-15 層目においては隠れ状態内の約 10%のニューロンから，全ニューロンを使用した際の約 95%の正解率が得られた．また，L_0 を用いて学習されたマスク付きプローブは，後半層に行くにつれて，文法性判定に使用するニューロンを削減できなくなる傾向が見られた．

2.3　考察

L_0 正則化を行うことでスパースプローブに使用するニューロンが選択されたこと，選択されたニューロンから，入力の文法性判定が可能であったことから，これらのニューロンが他のニューロンと比べて相対的にそれぞれの文法情報を強くエンコードしている可能性が考えられる．また，全ての λ において後半層に行くにつれて L_0 正則化を用いて学習されたマスク付きプローブは，文法性判定に用いるニューロンの数を減らせなくなる傾向があることから，LM は後半層にいくにつれて，文法情報をより多くのニューロンにエンコードしていく可能性が考えられる．

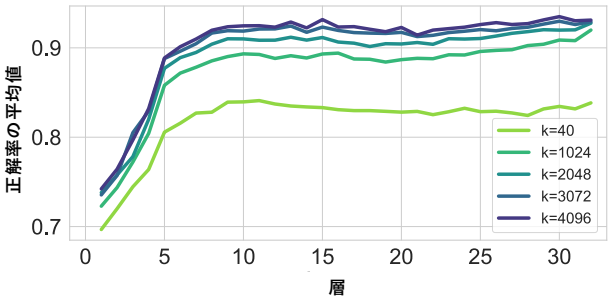

図 3　Welch の t 検定の t 値の絶対値によって選択された k 個のニューロンを用いた Top-k スパースプローブの正解率を示す．t 値を用いて選択された少数のニューロンからも文法性が判定可能であることが確認された．

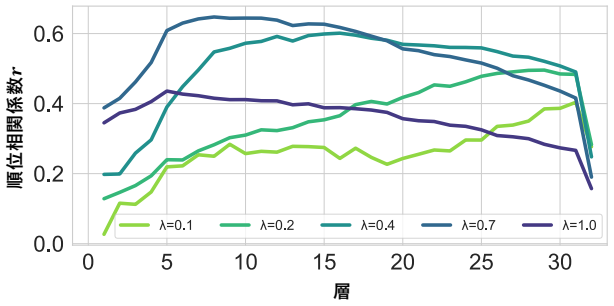

図 4　L_0 正則化によって得られた各 λ でのマスクの値と Welch の t 検定の t 値のスピアマンの順位相関係数を示す．同じ値を多く含むデータの場合，スピアマンの順位相関係数の取り得る最大値は 1 より小さくなる．L_0 正規化によって学習されたマスクが多くの 0 を含みがちであることを考慮すれば，2 つの値は強い相関を示している．

3　文法情報の局在性

この章では前章での考察から，文法情報を強くエンコードする文法ニューロンの存在を示し，文法情報の局在性を明らかにする．

3.1　実験方法

この章では，使用するニューロンの個数を指定して，スパースプローブを学習し，入力の文法性判定を行う．

Top-k スパースプローブ　Welch の t 検定の t 値を用いて，文法性判定に寄与すると推定される Top-k のニューロン [8] のみを用いて文法性判定を行うようスパースプローブのマスク m を定義する。AlKhamissi ら [13] は t 値を用いて，文法処理に寄与するニューロンが多い層を調査した．本研究では、文法的に正しい文の集合と誤った文の集合に対して Welch の t 検定の t 値の絶対値を判定への寄与度として算出し，寄与度が大きい k 個のニューロンに対応する m の要素を 1 に，それ以外を 0 に定める．

3.2　文法ニューロン

本節では，Top-k スパースプローブにおいても，少数のニューロンから文法性判定ができることを確認しする．ニューロン選択に使用した t 値の絶対値と，前章で L_0 正規化を用いて学習したマスクの値の間の相関を調査し，文法情報を多く有する文法ニューロンの存在を予想する．

結果　図 3 に，Top-k スパースプローブを用いた文法性判定の正解率を示す．Top-k スパースプローブでも少数のニューロンから文法性判定が可能であることが示された．全体の 1%のニューロンに当たる Top-40 のニューロンのみから，85%近い正解率が得られ，全体の 4 分の 1 に当たる Top-1024 のニューロンのみから，全てのニューロンを用いた際の正解率の 95%を大きく上回る正解率が得られた．

分析　図 4 に，Top-k ニューロンの選択に用いた t 値の絶対値と，前章で L_0 正則化によって得られた，各ニューロンに対応するマスクの値の間のスピアマンの順位相関係数を示す．L_0 正規化によって学習されたマスクが多くの同じ値 (ゼロ) を含んでいるにも関わらず，2 つの値の間には高い相関が見られ，2 つの異なる選択方法によって選択されたニューロンに類似性があることが明らかになった．この事実は，それぞれの選択方法によって選択されたニューロンが選択方法に依存しないことを表しており，これらのニューロンが文法情報を強くエンコードする文法ニューロンである可能性を示唆している．

3.3　Top-40 ニューロンの定量的評価

本節では，前節で予想された文法情報を多く含む文法ニューロンとして，隠れ状態内の前ニューロンの 1%に当たる 40 個を選択した，Top-40 ニューロンを調査する．Top-40 ニューロンの補集合からランダムに k 個サンプリングしたニューロンを用いたスパースプローブの文法性判定の正解率と比較することで，Top-40 ニューロンが持つ文法情報の定量的な評価を行う．

結果　Top-40 スパースプローブとの補集合からランダムに k 個サンプリングしたニューロンを用いたスパースプローブの文法性判定の正解率の比較を図 5 に示す．Top-40 ニューロンは 15 層目までにおいて，その補集合からランダムにサンプリングした約 6 倍以上の数のニューロンと，15 層目から 25 層目にかけては約 6 倍の数のニューロンと，それ以降

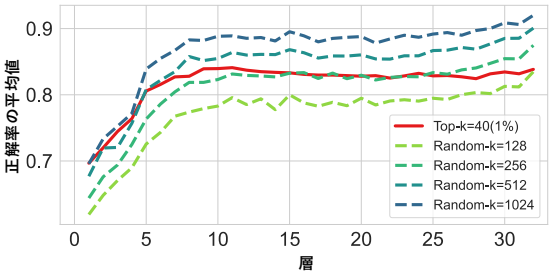

図 5　隠れ状態内の全ニューロンの約 1%に当たる Top-40 スパースプローブと補集合からランダムにサンプリングした k 個のニューロンを用いて訓練したプローブの文法性判定の正解率を示す．Top-40 ニューロンは大部分の層において，ランダムにサンプリングされた 6 倍またはそれ以上の数のニューロンを用いたプローブに匹敵する正解率で文法性を判定できており，隠れ状態内において，文法情報が局在していることを示している．

においては約 3-6 倍のニューロンと同程度の文法情報を持つことが明らかになった．

考察　得られた結果は，選択するニューロンによって，得られる文法情報の量が異なることを示している．隠れ状態において，文法情報は全てのニューロンに均一にエンコードされるのではなく，局在的にエンコードされるていると考えられる．前章において，学習可能マスクを L_0 正規化を用いて訓練した際に，後半層に行くにつれて、マスクが疎にならなくなることが明らかになった．図 5 より，ランダムに選んだニューロンを用いたスパースプローブとの Top-40 スパースプローブの相対的な正解率が低下した．これらのことから，LM が後半層に行くにつれて，特定の一部のニューロンへの依存度を下げていることが考えられる．

4　スパースプローブの汎化性

この章では学習可能マスク付きプローブの汎化性を調査する．一つのタスクで訓練されたプローブを用いて，全 67 タスクの文法性判定を行う．

4.1　結果

図 6 に第 16 層において，$\lambda = 0.2$ で一つのタスクで学習された学習可能マスク付きプローブ（使用ニューロン率 11.5%）を BLiMP の全 67 タスクで評価した正解率と，$\lambda = 1.0$ のプローブ（使用ニューロン率 100%) と比較した正解値の相対値を示す．図の中央に青い帯が観察されることから，学習に使用するタスクによっては，学習に使用したタスク以外のタスクに関する文法性判定でも高い正解率を示すスパースプローブが作成されたことが確認された．この汎化性は，BLiMP での特定のタスクでのファインチューニングによる性能向上の汎化性 [14] より広範なものである．また，作成されたスパースなプローブの正解率は，学習タスクと評価タスクが一致しない場合でも全てのニューロンを用いたプローブに匹敵した．

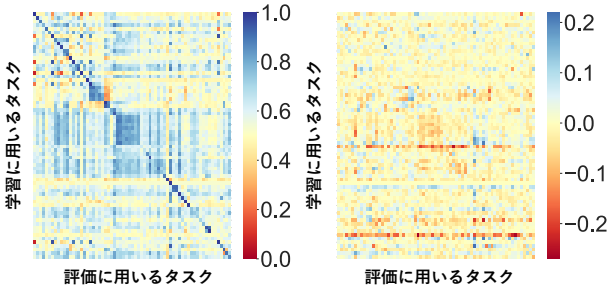

図 6　左に $\lambda = 0.2$ に設定して各タスクで学習した学習可能マスク付きプローブで，全 67 タスクで文法性判定を行った際の 16 層目の正解率を示す．他の多くのタスクに汎化可能なタスクが存在することが明らかになった．右に $\lambda = 1.0$ での対応する正解率との差を示す．$\lambda = 0.2$ のプローブ (使用ニューロン率 11.5%) の正解率は，$\lambda = 1.0$(使用ニューロン率 100%) のプローブの正解率に匹敵した．

4.2　考察

一部のタスクで学習した学習可能マスク付きプローブの汎化性から，これらのプローブのマスクによって選択されたニューロンのみで，文法タスクに依存しない，入力の文法性に関する一般的な情報を表現できる可能性が考えられる．

また，選択されたニューロンに対応する学習可能マスク付きプローブのマスクと重みの要素積は，線形空間内の文法性を表す方向を少数の標準基底の線形結合から構成する係数を示していると考えられる．Burger ら [15] は様々な文の一般的な真実性を表現する方向を発見し，高い精度で真実と嘘を分離できることを示したが，本実験は，文の内容だけでなく，文法性に関しても汎用的な誤り検出器が実現する可能性があることを示唆していると考えられる．

5　おわりに

本研究では，スパースプロービングにより LM の隠れ状態での文法情報の局在性を分析した．結果から，文法情報を多く有する文法ニューロンの存在と，これらの一部が汎用的な文法性判定能力が明らかになった．これらの知見が LM の解釈性や制御性向上に寄与することを期待する．

謝辞

本研究は JSPS 科研費 JP21K21343 の助成を受けたものである.

参考文献

[1] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors, **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[2] Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in gpt2 language models. **CoRR**, 2024.

[3] Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic evaluation of unlearning using parametric knowledge traces. **arXiv preprint arXiv:2406.11614**, 2024.

[4] Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. Intrinsic probing through dimension selection. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 197–216, 2020.

[5] Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R Brennan. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 4488–4497, 2024.

[6] Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. Large language models as neurolinguistic subjects: Identifying internal representations for form and meaning. **arXiv preprint arXiv:2411.07533**, 2024.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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.

[8] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. **arXiv preprint arXiv:2305.01610**, 2023.

[9] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. In **International Conference on Learning Representations**, 2018.

[10] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel Bowman. Blimp: The benchmark of linguistic minimal pairs for english. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[12] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11897–11916, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The llm language network: A neuroscientific approach for identifying causally task-relevant units. **arXiv preprint arXiv:2411.02280**, 2024.

[14] Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. Interpretability of language models via task spaces. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4522–4538, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[15] Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in llms. **arXiv preprint arXiv:2407.12831**, 2024.