言語モデルにおける共通統語メカニズムの分析

九門 涼真 [1,2]　谷中 瞳 [1,2,3]
[1] 東京大学　[2] 理化学研究所　[3] 東北大学
{kumoryo9, hyanaka}@is.s.u-tokyo.ac.jp

概要

言語モデルは振る舞いとしては高い統語処理能力を示すが，その内部機序が言語学的な共通性を反映しているかは自明でない．本研究では，因果的分析手法を用い，フィラー・ギャップ依存関係 (FGD) と否定極性項目 (NPI) を対象に，異なる構文間でのメカニズムの共通性を層・注意機構ヘッド単位で分析した．分析の結果，FGD には序盤から中間層に共通して寄与する限られた数の注意機構ヘッドが存在する一方，NPI には存在しなかった．また，特定した注意機構ヘッドの操作により容認性判断ベンチマークにおけるモデルの精度が向上し，さらに分析に用いた手法が分布外データに対して高い頑健性を示したことから，分析結果の妥当性が支持された．

1　はじめに

言語モデルは，複雑な階層構造を伴う統語処理において高い能力を示しつつある．しかし，その内部機序が言語学の知見とどの程度整合し，人間の処理メカニズムと類似しているかは未解明である．言語学は異なる構文間に共通する普遍的な要素を記述してきたが，言語モデルにおいても，表層的には異なる構文間で共通の処理メカニズムが利用されているかを明らかにすることは，言語の構造および言語モデルのより深い理解に繋がる [1]．

モデルの挙動に内部構成要素が及ぼす因果的影響を特定する手法として，因果的分析手法 [2, 3, 4] が用いられている．先行研究では，この手法を用いて統語メカニズムの分析 [5, 6] が行われており，特に主述の一致 [7] やフィラー・ギャップ依存関係 (FGD) [8] を処理するメカニズムの共通性が分析されているが，課題も存在する．まず，既存の分析はモデルがどのトークンを重要視しているかに関する分析に留まっており，層や注意機構ヘッドの粒度での分析されておらず，どの粒度までメカニズムが共通しているかは示されていない．次に，先行研究 [5, 8] は学習を伴う分析手法を採用していることから，過学習の懸念があり [9]，特定されたメカニズムが分布外にまで汎化するかは明らかでない．

そこで本研究では，学習を伴わない因果的分析手法である activation patching (AP) [2, 3] を用い，層・注意機構ヘッド単位で異なる構文間のメカニズムの共通性を分析する．分析対象として FGD と否定極性項目（NPI）の二つを用い，それぞれにおいてメカニズムが共通しているのか，している場合にはそれが何かを明らかにする．また，特定されたメカニズムの妥当性を検証するため，特定した構成要素の操作による容認性判断ベンチマークにおけるモデルの精度変化を検証し，加えて分布外 (OOD) データによる評価で学習を伴う分析手法である distributed alignment search (DAS) [4] との比較を行う．

分析の結果，FGD には共通のメカニズムが存在する一方，NPI では共通メカニズムは確認されなかった．また，FGD のメカニズムに寄与する構成要素は序盤から中間層に位置するごく少数の注意機構ヘッドに局所化されていることが明らかになった．また，特定した構成要素の活性値を操作することで，容認性判断タスクの精度改善が見られた．さらに，DAS は OOD 設定において分析結果が大幅に変化し，語彙や構文の分布に過学習する可能性が示唆された一方で，AP は一貫した結果を示した．

2　実験設定

2.1　データセット

表 1 に示すように，FGD および NPI のそれぞれについて，異なる構文パターンを含むミニマルペアからなるデータセットを使用する．ミニマルペアは，統語現象に関する重要な要素だけが異なる二文からなり，異なる要素によって正しい出力が定まるように設計する．各言語現象のパターンの網羅性については，先行研究（FGD は [8]，NPI は [10, 11]）に従い，FGD は 7 種類，NPI は 8 種類のパターンを選定

表 1　ミニマルペア（ID テストセット）の例．NC は分析対象としない語彙を表す．

フィラー・ギャップ依存関係 (FGD)								
構文	略称	prefix	filler	NC	the	noun	verb	出力
埋め込み wh 疑問文	EWhK	The man knows	[**who/that**]		the	teacher	liked	[**./her**]
分裂文	Cleft	It was	[**the man/clear**]	that	the	boss	scared	[**./him**]
主題化	Topic	Actually,	[**the kid/ϕ**]		the	guest	hated	[**./them**]
否定極性項目 (NPI)								
構文	略称	prefix	licensor	NC			last	出力
条件節	Cond	The host will sleep	[**if/while**]	the guest			eats	[**any/some**]
否定限定詞	DNeg		[**No/The**]	patient have			liked	[**any/some**]
最上級	Sup	This is the	[**fastest/fast**]	kid that had			liked	[**any/some**]
コントロール								
構文	略称	prefix	filler/licensor	NC	the	noun	verb/last	出力
首都に関する知識	Ctrl		[**Tokyo/Rome**]	is	the	capital	of	[**Japan/Italy**]

した（全てのパターンの例は付録 A の表 2 を参照）．

また，DAS の学習および評価のため，学習セット，分布内 (ID)，分布外 (OOD) テストセットを用意する．学習セットと ID テストセットは同一の語彙セットから生成し，OOD テストセットは異なる語彙セットから生成する．特に断りがない限り，結果は ID および OOD テストセットの平均を報告する．

データセットの構築にあたっては，[12] によるデータ生成スクリプト[1)]を改変して用いた．このスクリプトは，まず文構造のテンプレートを定め，次に広範な語彙セットから適切な単語を挿入することで文を生成する．この手法により，語彙および構文の双方において広い分布を持つデータセットを作成することが可能である．

2.2　モデル

[5, 8] に従い，Pythia-1B, 2.8B, 6.9B [13] を分析に用いる．特に断りがない限り，1B の結果を報告する．

3　共通統語メカニズムの分析

3.1　分析手法

モデルの特定の挙動に対してどの構成要素が因果的に影響を与えているかを特定する手法として，activation patching [2, 3] (AP) を採用する．この手法では，ある入力 $\boldsymbol{b}$ に対する推論において，特定の構成要素 f の活性値 $f(\boldsymbol{b})$ を，別の入力 $\boldsymbol{s}$ から得られた活性値 $f(\boldsymbol{s})$ で置き換え，その介入がモデルの最終的な出力に与える影響を調べる．このとき，置き換えに用いる入力は，元の入力との差分が分析対象とする特定の概念のみに限定されるよう設計することで，当該概念に関わる機序を明らかにする．本研究では，構築したミニマルペアを用いて，モデルのどの構成要素が統語処理を担っているかを分析する．

3.2　評価指標

[5] に従い，特定の構成要素に対する介入が出力に与える影響を評価する指標として，次のように定義される log-odds ratio（以下 Odds）を用いる．この指標の値が大きいほど，モデルの出力に対する構成要素の因果的影響が大きいことを示す．

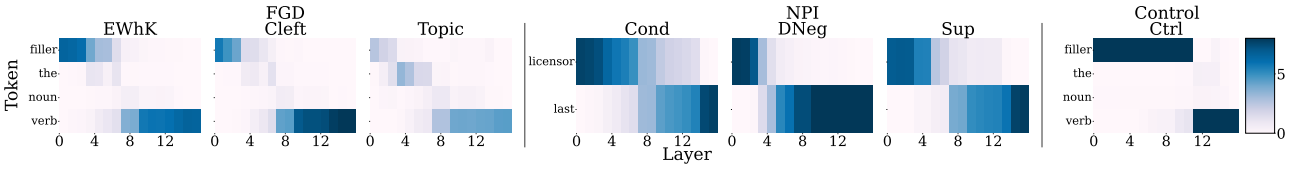
$$\begin{aligned}&\textsc{Odds}(p, p_{f_{\text{interv}}}, T)\\&= \frac{1}{|T|}\sum_{(\boldsymbol{b},\boldsymbol{s},y_b,y_s)\in T} \log\left(\frac{p(y_b|\boldsymbol{b})}{p(y_b|\boldsymbol{s})}\frac{p_{f_{\text{interv}}}(y_b|\boldsymbol{s},\boldsymbol{b})}{p_{f_{\text{interv}}}(y_b|\boldsymbol{b},\boldsymbol{s})}\right)\end{aligned}$$

ここで，y_b, y_s は入力 $\boldsymbol{b}, \boldsymbol{s}$ に対応する正解トークン，T はテストセット，p はモデルの出力確率を表し，$\boldsymbol{b}$ の推論時に $\boldsymbol{s}$ の活性値で f に介入した場合の y_b の出力確率を $p_{f_{\text{interv}}}(y_b|\boldsymbol{b},\boldsymbol{s})$ で示す．

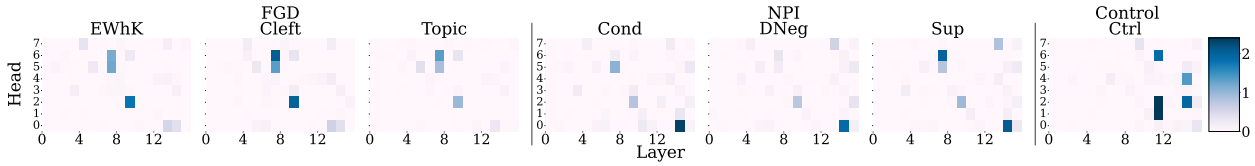
3.3　結果

FGD における各構成要素への AP の結果を図 1 に示す．FGD の異なる構文間において各構成要素のスコア分布は共通している一方で，コントロールパターンとは異なる傾向を示しており，FGD に共通のメカニズムが存在することが明らかになった．また，最終トークンにおいて出力に寄与する注意機構ヘッドは，中間層（ヘッド 7.5, 7.6, 9.2）に疎に存在しており，この局所化された分布も構文間で共通していた．また，最終トークンにおける各層の出力

1）https://github.com/alexwarstadt/data_generation



(a) 残差ストリーム



(b) 注意機構ヘッド

図 1　各構文パターンにおける AP による ODDS スコアの結果．トークン名はおよびパターンの略称は表 1 に対応する．

（残差ストリーム）のスコアの第 7 層での急激な上昇は上述のヘッドの位置と一致する．したがって，これらのヘッドは，先行するトークンから重要な情報を最終トークンへと移動させる役割を担っていると考えられる．FGD の他の構文でも同様の傾向が見られた（付録 B の図 4 を参照）．

一方，NPI に関する構文では異なる傾向が観察された．残差ストリームでは DNEG において ODDS スコアがより早期の層から上昇し，また注意機構ヘッドでは SUP においてヘッド 7.5 のスコアがより高く，構文間で一貫した傾向は見られなかった．この原因として，NPI の処理には構文ごとに異なる統語・意味的な複雑さが伴うためと考えられる．

さらに，モデルのパラメータサイズの違いについては，モデルの層数が増えるほど，FGD の処理がより早期の層で行われる傾向が見られた．一方で，サイズにかかわらず，メカニズムは構文間で共有されており，かつ特定の構成要素に局所化されているという傾向は維持されていた（付録 B の図 5 を参照）．

4　統語メカニズムに基づくモデルの振る舞いの操作

4.1　手法

これまでの分析により，FGD における共通統語メカニズムの存在が明らかになったが，これが統語構造の処理とは関係のないヒューリスティックを捉えたものでないか検証する必要がある．そこで，モデルの推論時に，メカニズムへの寄与が示された注意機構ヘッド (7.5, 7.6, 9.2) の活性値に対し α を乗じ，SyntaxGym [14] および BLiMP [12] における精度の変化を測定する．両ベンチマークはモデルが文法的な文に対して非文法的な文よりも高い確率を割り当てられるかを評価する．なお，BLiMP では，文間のトークン長の違いが確率を歪め，正当な評価を妨げる可能性が指摘されている [15] ため，トークナイズ後の長さが同一のミニマルペアのみを使用した．

4.2　結果

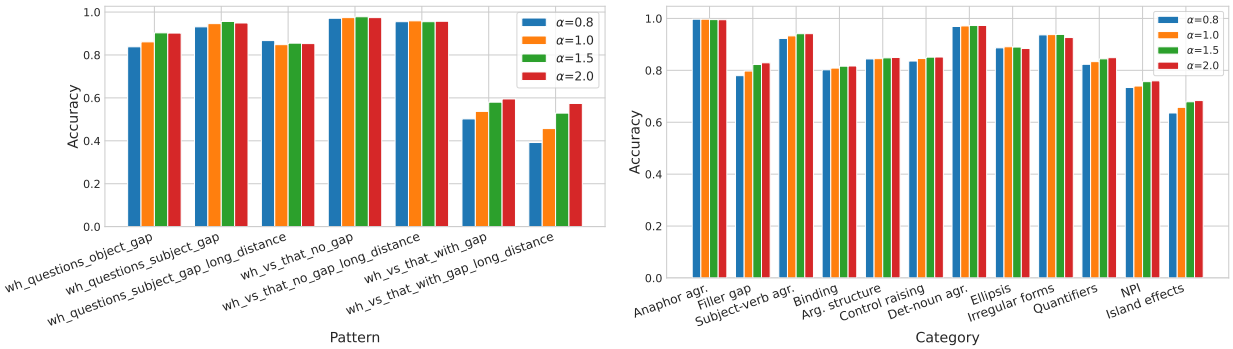
以下では BLiMP の結果を示すが，SyntaxGym でも同様の傾向が得られた．まず，FGD に直接関連するカテゴリの性能変化に注目する．図 2a に示す通り，特定した 3 つの注意機構ヘッドの活性値を $\alpha > 1$ で強化したとき，ほとんどのパターンで性能の向上，あるいは飽和による維持が確認された．この結果は，操作対象とした注意機構ヘッドが，分析に用いた直接目的語ギャップだけでなく，目的語ギャップ，前置詞目的語ギャップ，さらには長距離依存関係を含む FGD 全般において，文法的に正しい文への高い確率の割り当てに寄与することを示す．

次に，FGD 以外のカテゴリに関して（図 2b）は，島の制約，束縛，NPI，数量詞，中央埋め込みなどのカテゴリでも改善が見られた．また，主述の一致に関する改善は，前置詞句や関係節といった介在要素を越えた一致を判定するパターンにおいて生じた．以上の結果は，特定した注意機構ヘッドが，単に FGD に特化した処理を担っているだけでなく，階層的な構造を捉えるためのより広範なメカニズムとして一定程度機能している可能性を示唆する．

5　DAS と AP の比較

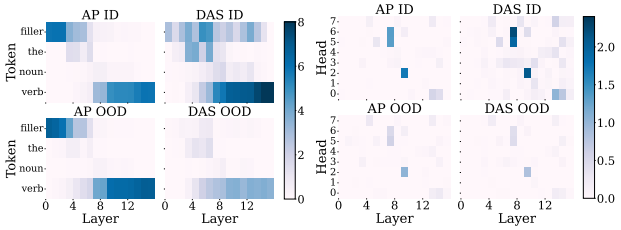
5.1　DAS

先行研究 [5, 8] は DAS [4] を用いて統語メカニズムの分析を行ったため，DAS と AP の比較を行う．DAS は，モデルの振る舞いに因果的に影響を及ぼす



(a) FGD に関するカテゴリ　　(b) 全てのカテゴリ

図 2　特定した注意機構ヘッドの活性値を α 倍した場合の，BLiMP における正答率.



(a) 残差ストリーム　　(b) 注意機構ヘッド

図 3　FGD の EWHK パターンにおける AP と DAS の分布の違いに対する ODDS スコアの傾向の変化.

構成要素を，教師あり学習で得られた部分空間への介入により特定する手法である．本研究では [5, 8] に倣い，出力に因果的な影響を与える方向を学習する一次元の DAS を採用する．この介入操作は，以下の式によって定義される．

$$f_{\text{interv}}(\boldsymbol{b}, \boldsymbol{s}) = f(\boldsymbol{b}) + (f(\boldsymbol{s})\boldsymbol{a} - f(\boldsymbol{b})\boldsymbol{a})\boldsymbol{a}^T$$

$\boldsymbol{a}$ は学習対象となるベクトルである．学習では，以下の目的関数を用い，介入元の入力 $\boldsymbol{s}$ の正解トークン y_s の出力確率を増やす方向 $\boldsymbol{a}$ を学習する．

$$\min_{\boldsymbol{a}} \left\{ - \sum_{(\boldsymbol{b},\boldsymbol{s},y_b,y_s)\in D} \log p_{f_{\text{interv}}}(y_s|\boldsymbol{b},\boldsymbol{s}) \right\}$$

ここで D は学習セットを表す．

本研究では，異なる構文間に共通するメカニズムを分析するため，[8] に従い，leave-one-out により学習および評価を行う．具体的には，構文集合 $\{c_j \mid j \neq i, 0 \leq i, j \leq n\}$ を用いて方向 $\boldsymbol{a}_i$ を学習した後，c_i で介入に適用し評価を行う．他の構文で学習された方向が対象の構文にどの程度汎化するかを分析し，構文間のメカニズムの類似性を評価する．

5.2　結果

FGD の EWHK パターンにおける ODDS スコアの結果を図 3 に示す．なお，以下に述べる傾向は FGD の他の構文においても一貫して観察された．残差ストリームでは，DAS は ID テストセットと比較して OOD テストセットでの ODDS スコアが低下したが，学習を伴わない AP では，ID と OOD の双方で一貫したスコア分布が得られた．この結果は，DAS により学習された方向が過学習により統語構造以外の情報を含んでいた可能性を示唆している．

一方で，注意機構ヘッドを対象とした場合には，DAS と AP のスコア分布は比較的類似しており，DAS は注意機構ヘッドの分析により適している可能性が示唆された．しかし，OOD で ODDS スコアが低下する傾向は依然として存在し，また ID で終盤層の注意機構ヘッドが AP よりも高いスコアを示した．この原因としては，DAS がトークンの出力確率に関する最適化を行うために，終盤層の寄与を過大評価する可能性が考えられる．

DAS を用いた先行研究 [5, 8] では，最終トークンの終盤層の残差ストリームでスコアが急増する結果をもとに，それらの層が最も因果的に影響を持つと結論づけられていた．これに対し本研究の結果は，その急増は OOD にも汎化する傾向ではなく，実際には序盤から中盤の層で統語構造に関する処理が行われていることを示し，学習を伴う解釈手法を用いる際には OOD 評価が重要であることを示唆する．

6　おわりに

本研究では，因果的解釈手法を用い，異なる構文間における言語モデル内の統語メカニズムの共通性を分析した．AP による分析の結果，FGD には共通の局所的な構成要素の寄与が見られた．また，共通統語メカニズムへの寄与が示された構成要素の操作により，モデルの統語処理能力の向上が観察され，分析結果の妥当性が支持された．

謝辞

本研究は JST CREST，JPMJCR2565，JST BOOST，JPMJBY24H5 の支援を受けたものである．

参考文献

[1] Richard Futrell and Kyle Mahowald. How linguistics learned to stop worrying and love the language models. **Behavioral and Brain Sciences**, p. 1–98, 2025.

[2] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 12388–12401. Curran Associates, Inc., 2020.

[3] Atticus Geiger, Hanson Lu, Thomas F Icard, and Christopher Potts. Causal abstractions of neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, **Advances in Neural Information Processing Systems**, 2021.

[4] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez, editors, **Proceedings of the Third Conference on Causal Learning and Reasoning**, Vol. 236 of **Proceedings of Machine Learning Research**, pp. 160–187. PMLR, 01–03 Apr 2024.

[5] Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. 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. 14638–14663, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 160–175, 2021.

[7] Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1828–1843, Online, August 2021. Association for Computational Linguistics.

[8] Sasha Boguraev, Christopher Potts, and Kyle Mahowald. Causal interventions reveal shared structure across English filler–gap constructions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 25032–25053, Suzhou, China, November 2025. Association for Computational Linguistics.

[9] Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to makelov et al. (2023)'s "interpretability illusion" arguments. 2024. arXiv preprint.

[10] Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2877–2887, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11] Deanna DeCarlo, William Palmer, Michael Wilson, and Bob Frank. NPIs aren't exactly easy: Variation in licensing across large language models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi, editors, **Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP**, pp. 332–341, Singapore, December 2023. Association for Computational Linguistics.

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

[13] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, KyleO'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff ほか. Pythia: A suite for analyzing large language models across training and scaling. In **International Conference on Machine Learning**, pp. 2397–2430. PMLR, 2023.

[14] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1725–1744, Online, July 2020. Association for Computational Linguistics.

[15] Naoya Ueda, Masato Mita, Teruaki Oka, and Mamoru Komachi. Token-length bias in minimal-pair paradigm datasets. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 16224–16236, Torino, Italia, May 2024. ELRA and ICCL.

A　データセット

表 2　FGD と NPI のすべての構文パターンのミニマルペア（ID テストセット）の例.

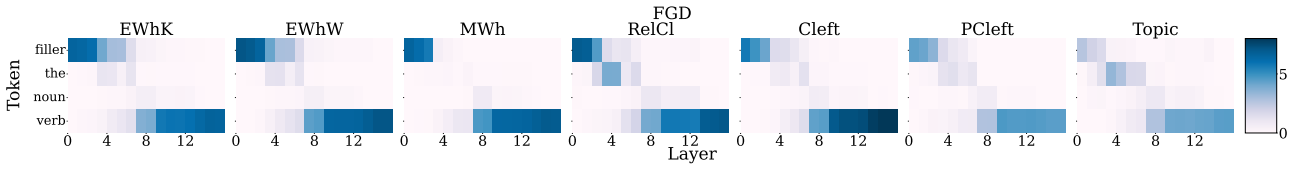
フィラー・ギャップ依存関係 (FGD)

構文	略称	文頭	フィラー	NC	冠詞	名詞	動詞	出力
埋め込み wh 疑問文 (know)	EWhK	The man knows	[**who/that**]		the	teacher	liked	[**./her**]
埋め込み wh 疑問文 (wonder)	EWhW	The boy wondered	[**who/if**]		the	doctor	admired	[**./him**]
wh 疑問文	MWh	Then,	[**who/**ϕ]	did	the	customer	choose	[**?/them**]
関係詞節制限用法	RelCl	The customer	[**who/and**]		the	lady	sounded like	[**./me**]
分裂文	Cleft	It was	[**the man/clear**]	that	the	boss	scared	[**./him**]
擬似分裂文	PCleft		[**Who/That**]		the	dancer	is listening to	[**is/you**]
主題化	Topic	Actually,	[**the kid/**ϕ]		the	guest	hated	[**./them**]

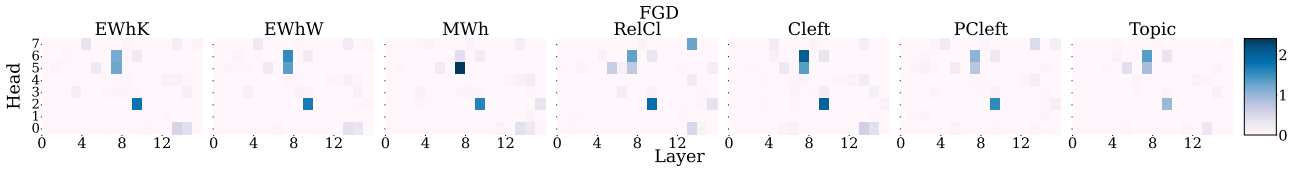
否定極性項目 (NPI)

構文	略称	文頭	認可子	NC	最終	出力
条件節	Cond	The host will sleep	[**if/while**]	the guest	eats	[**any/some**]
否定限定詞	DNeg		[**No/The**]	patient have	liked	[**any/some**]
Only（スコープ）	SOnly		[**Only/Even**]	the boys who	have	[**any/some**]
数量詞	Qnt	These are	[**all/some**]	of the students who	showed	[**any/some**]
埋め込み疑問文	EmbQ	The senators	[**wonder whether /know that**]	the man has	found	[**any/some**]
疑問文	SmpQ		[**Has/**ϕ]	the actor	sold	[**any/some**]
最上級	Sup	This is the	[**fastest/fast**]	kid that had	liked	[**any/some**]
Only	Only	They are the	[**only/upset**]	teachers that	makes	[**any/some**]

B　結果

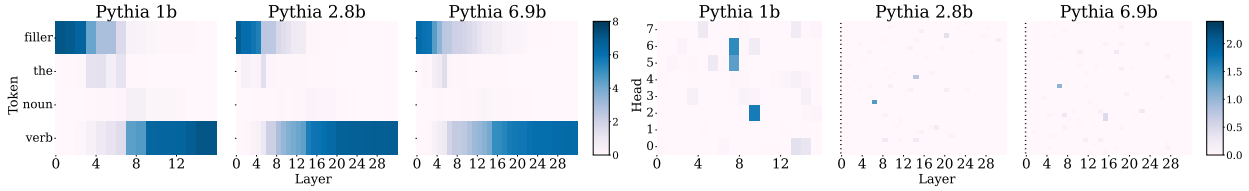


(a) 残差ストリーム



(b) 注意機構ヘッド

図 4　全ての FGD の構文パターンにおける AP による Odds スコアを層・注意機構ヘッドごとに示した結果.



(a) 残差ストリーム　　(b) 注意機構ヘッド

図 5　EWhK における AP による各パラメータサイズのモデルの Odds スコア