構成的汎化における Transformer の内部機序の分析

九門 涼真　谷中 瞳
東京大学
{kumoryo9, hyanaka}@is.s.u-tokyo.ac.jp

概要

ニューラルモデルの構成的汎化能力は，人間のような言語能力の実現に向けた重要な課題の一つである．既存研究は，モデルの出力に着目して構成的汎化能力を評価しており，モデルの構成的汎化における内部機序は明らかでない．そこで本研究では，構成的汎化に寄与するサブネットワークの探索とモデルの統語的特徴の活用に関する因果分析を行い，Transformer の内部機序を分析する．実験結果から Transformer は構成的汎化において統語的特徴に一定程度依存することが示された．一方で，モデル全体よりも優れた汎化性能を持つサブネットワークは，統語的特徴を用いた構成的な解法に加え，非構成的な解法にも依存することも示された．

1　はじめに

既知の単語や統語構造に基づいて，新しい言語表現の意味を理解する能力である構成的汎化は，未知の言語データに対する頑健性における重要な要素の一つである．ニューラルモデルの構成的汎化能力を評価するために，既存の研究のほとんどは，ベンチマークにおけるモデル出力の評価に焦点を当ててきた [1, 2, 3, 4]．しかし，ベンチマークで良いパフォーマンスを示すモデルが必ずしも構成的な解法によって正しい出力に至っているとは限らない．また，近年モデルの解釈可能性に関する研究が活発に進められており，Transformer [5] ベースのモデルの内部機序が調査されてきた [6, 7] が，構成的汎化を対象とした研究はほぼ存在しない．[8] と [9] は，構成的汎化における内部機序を分析したが，構成的汎化においてどのような言語的特徴が重要な役割を担うのかは分析していない．したがって，構成的汎化におけるモデルの内部メカニズムは未だに明らかでない．

本研究では，構成的汎化を必要とするタスクにおいて Transformer がどのように統語的特徴に依存するかに注目し，内部機序を分析する．分析手法は，

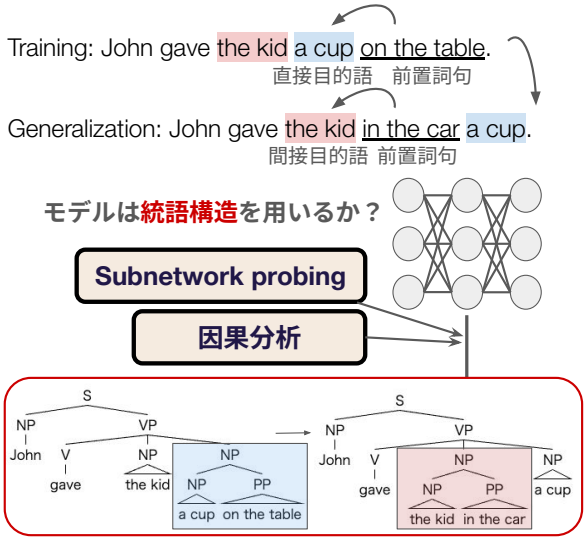


図 1　本研究では構成的汎化を必要とするタスクにおける Transformer の内部機序を分析する．

図 1 に示すように，(i) モデル内で構成的汎化で高精度を示すサブネットワークを特定すること，および (ii) 元のモデルとそのサブネットワークにおける統語的特徴の因果効果を評価することからなる．因果分析では，モデルから統語的特徴を除去し，除去の前後で汎化性能を比較する．厳密に構成的汎化能力を評価・分析するために，事前学習済みのモデルではなく，スクラッチから学習した Transformer モデルを用いる．事前学習データには，モデルの汎化能力を評価する際に対象とする統語構造が含まれている可能性があり，事前学習済みのモデルは構成的汎化能力の正確な評価が難しいためである [10]．また，構成的汎化の評価で一般的に用いられる機械翻訳と意味解析の二つのタスクを用い，それぞれのタスクに対し二つの構成的汎化のパターンで評価を行う．

実験結果から，モデルとその構成的汎化の精度に寄与するサブネットワークは，構成的汎化において統語的特徴を活用することが示唆された．一方で，このサブネットワークは，統語的特徴を用いない解法にも依存していることが明らかになった．そのため，モデルは部分的に非構成的な解法で構成的汎化

表 1　本実験で評価に用いる二つの構成的汎化のパターン．

パターン	学習データ	分布外評価データ
前置詞句 in 間接目的語 (PP-IOBJ)	The child gave **the pen on the table** to Liam.	The friend gave **the girl in the room** a hat.
前置詞句 in 主語 (PP-SUBJ)	The child broke **a cup on the table**.	**The friend in the room** broke a cup.

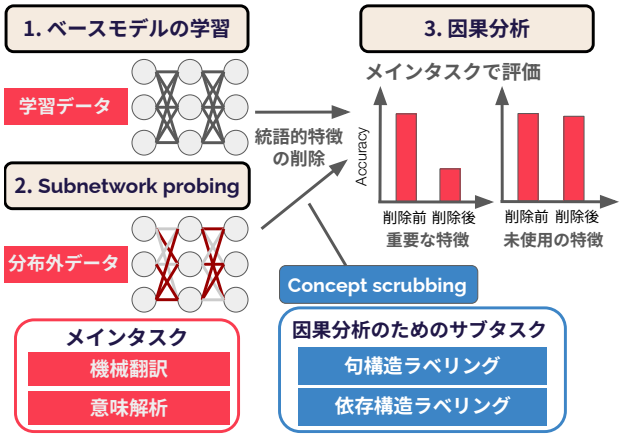


図 2　分析手法の概要．

を要するタスクを解くことが示された．また，非構成的な解法は学習の初期段階で習得され，学習が進むにつれて徐々に構成的な解法が習得されることも示された．

2　関連研究

ニューラルモデルの構成的汎化能力の評価は，意味解析 [1, 4] や機械翻訳 [2, 11] におけるモデルの出力の分析により行われてきた．これらの研究により，一般に既存のモデルの構成的汎化能力には改善の余地があることが示された．また，近年様々なモデルの内部機序の分析手法も提案されている．その一つに，特定の特徴がモデルの振る舞いに及ぼす影響をモデルへの介入により分析する手法がある [12, 13]．[13] は特定の概念をモデルから消去し，消去前後での精度を比較することで，モデルがどの程度その概念に依存するかを調べる手法を提案した．モデルの分析の異なる方向性として，モデル内から特定のタスクを解くことができるサブネットワークを特定する手法も提案されている [14, 15]．[16, 17] はこれらの手法に基づき，統語的汎化を担うサブネットワークを分析したが，本研究では構成的汎化を担うサブネットワークを分析する．

3　分析手法

図 2 に構成的汎化における内部機序の分析手法の概要を示す．分析手法はベースモデルの学習，subnetwork probing，因果分析の三段階からなる．

3.1　ベースモデルの学習

まずモデルの学習と評価に用いるデータセットを構築する．データセットは学習データ，分布内評価データ，分布外評価データからなる．分布外評価データには，学習データに含まれる統語構造の組み合わせである未知の統語構造が含まれており，正しい出力のためにはモデルはギャップを埋めて汎化することが求められる．このデータセットは，SGET [11] で使用されたルールベースのパイプラインに基づいて構築する．SGET では，確率文脈自由文法を活用して文を生成することで，学習データと分布外評価データ間のギャップを厳密に制御し，構成的汎化能力の正確な評価を可能にしている．

次に，学習データを用いて Transformer をスクラッチから学習する．学習タスクとして，構成的汎化の評価で広く用いられている機械翻訳と意味解析を採用する．タスクを二つ用いることで，タスクの出力形式がモデルの内部処理に与える影響を調べる．機械翻訳は英日翻訳を対象とし，意味解析の論理式は，[18] が提案した変換手法に基づいて作成した．また，表 1 に示した二つの構成的汎化のパターン（PP-IOBJ, PP-SUBJ）を評価の対象とする．学習データにおいて前置詞句 (PP) に修飾される名詞句 (NP) は全て直接目的語であり，分布外評価データにおいて PP に修飾される間接目的語や主語の NP を含む文への汎化能力を評価する．

3.2　Subnetwork Probing

次に，構成的汎化のタスクを高精度で解くサブネットワークを特定するために，subnetwork probing を学習済みのベースモデルに適用する．subnetwork probing は，枝刈りに基づいた probing により，特定のタスクで高い精度を示すサブネットワークを発見する手法である．本分析では，分布外評価データと同じ分布から生成されるが重複のないデータを用いて枝刈りのためのマスクを学習し，サブネットワークを得る．なお，学習する対象はマスクのみで，モデル自体の重みは固定されており，モデルは新たに分布外評価データに関する学習をしないことに注意

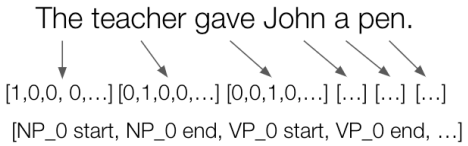


図 3　句構造に関する多ラベル分類タスクの例.

する．subnetwork probing の詳細は，付録 A に示す．

3.3　因果分析

最後に，ベースモデルと発見されたサブネットワークを，機械翻訳と意味解析においてどの程度統語的特徴に依存しているかという観点で分析する．ベースモデルとサブネットワークの両方から構成的汎化に関わる統語的特徴を削除し，その前後で機械翻訳と機械翻訳と意味解析の分布外評価データにおける精度がどのように変化するかを評価する．これによりベースモデルとサブネットワークの統語的特徴への依存を評価できる．特徴の削除には，concept scrubbing [13] を用いる．concept scrubbing は，モデルの内部表現への影響を最小限にしながら特定の概念を消去する手法である LEACE [13] をモデルの各層の内部表現に順に適用する．具体的には，線形分類器が定数関数よりも特定の概念に関するラベルを正確に予測できなくなり，かつ他の概念はモデルに保存されるように内部表現を更新する．

concept scrubbing では，対象とする特徴の消去に分類タスクを使用する．そこで図 3 に示すように，[12] に従い，統語的特徴に関する多ラベル分類タスクを定義する．削除する対象の特徴として，句構造と依存構造を考え，それぞれに分類タスクを定める．また，詳細な分析のために，部分的な統語的特徴の削除の影響も検証する．例えば，PP に修飾される間接目的語の NP のみに関する句構造の特徴を削除した場合を考える．このとき，ラベルは注目している構造に関わるトークンにのみ割り当てられる．句構造は Penn Treebank [19] の定義を，依存構造は Universal Dependencies [20] の定義をもとにラベルを割り当てる．PP-IOBJ の因果分析では，句構造と依存構造全体に加えて間接目的語（PP-SUBJ では主語）の NP，直接目的語の NP，全ての NP の PP 修飾のみに関する句構造と依存構造の影響を分析する．

4　実験設定

データセット　3.1 項で述べた手法により，機械翻訳と意味解析それぞれのデータセットを構築する．両データセットともに，学習データに 80,000 件，分布内評価データに 10,000 件，分布外評価データに 12,000 件の文からなる．分布外評価データは構成的汎化のパターンごとにそれぞれ構築し，subnetwork probing においてもパターンごとに高精度を示すサブネットワークを探索する．

モデル　三層のエンコーダとデコーダ，四個の注意機構ヘッドからなる Transformer を学習し，分析する．学習はランダムに選んだシードで三回行い，その平均の値を報告する．5.1 項と 5.2 項で報告する結果は学習の最後のチェックポイントのものである．学習に用いたハイパーパラメータは付録 B に示す．

評価指標　構成的汎化の評価の先行研究 [1, 11] に従い，機械翻訳と意味解析ともに完全一致率を評価指標に採用する．ルールベースでのデータセット構築により，構成的な規則に従うモデルの出力が一意に定まるような設定となっているため，本研究では完全一致率は構成的汎化能力の評価に適する．

表 2　機械翻訳（MT）と意味解析（SP）における完全一致率 (%). Base はベースモデル，PP-IOBJ Sub. は PP-IOBJ のサブネットワークを指し，PP-IOBJ の列は PP-IOBJ の分布外評価データでの精度を示す．PP-SUBJ に関しても同様.

タスク	モデル	分布内	PP-IOBJ	PP-SUBJ
MT	Base	$99.9_{\pm 0.0}$	$47.0_{\pm 2.2}$	$0.0_{\pm 0.0}$
	PP-IOBJ Sub.	$99.8_{\pm 0.0}$	$91.4_{\pm 3.0}$	—
	PP-SUBJ Sub.	$96.5_{\pm 3.5}$	—	$57.0_{\pm 18.8}$
SP	Base	$99.8_{\pm 0.0}$	$55.3_{\pm 4.1}$	$0.1_{\pm 0.0}$
	PP-IOBJ Sub.	$94.3_{\pm 0.8}$	$91.2_{\pm 5.5}$	—
	PP-SUBJ Sub.	$98.3_{\pm 0.0}$	—	$12.4_{\pm 7.4}$

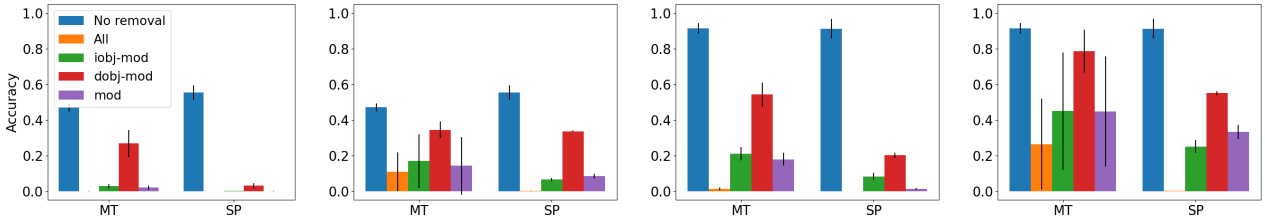
5　結果

5.1　出力評価

concept scrubbing を行う前のベースモデルとサブネットワークの結果を表 2 に示す．ベースモデルでは，既存研究と同様に，分布外評価データの方が分布内評価データより著しく低い精度となった．一方で，subnetwork probing で得られたサブネットワークはベースモデルより高い汎化精度を示し，特に PP-IOBJ では 90%を超えた．そのため，何らかの方法で構成的汎化のタスクを解くことができるサブネットワークが存在することが示唆される．

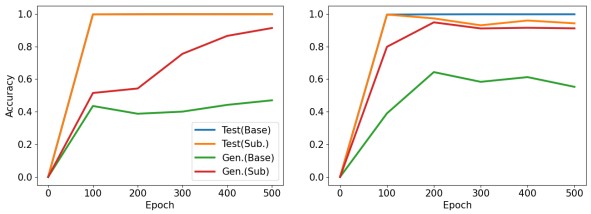
5.2　因果分析

PP-IOBJ の因果分析の結果を図 4 に示す．以降の PP-SUBJ の結果は付録 C に示す．ベースモデル・サブネットワークともに統語的特徴が完全に削除され



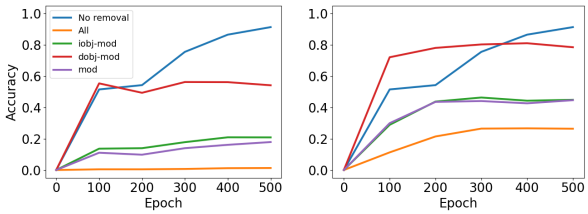
(a) 句構造 (Base)　(b) 依存構造 (Base)　(c) 句構造 (Sub.)　(d) 依存構造 (Sub.)

図 4　PP-IOBJ の因果分析の結果．各棒は対応する特徴の削除後の分布外評価データでの完全一致率を示す．Base と Sub. はそれぞれベースモデルとサブネットワークを指す．All は統語的特徴の完全な削除，iobj-mod は間接目的語の NP への PP 修飾に関する統語的特徴のみの削除（dobj-mod も同様），mod は NP への PP 修飾に関する統語的特徴の削除を指す．



(a) 機械翻訳　(b) 意味解析

図 5　PP-IOBJ における学習中のモデルの完全一致率の推移．Test と Gen. はそれぞれ分布内，分布外評価データでの精度に対応する．



(a) 句構造（機械翻訳）　(b) 依存構造（機械翻訳）

図 6　PP-IOBJ における各エポック数のサブネットワークの因果分析の結果．各線は対応する特徴の削除後の分布外評価データの完全一致率を示す．

た場合には汎化精度が極端に低下し，機械翻訳や意味解析における統語的特徴への依存が示された．また，サブネットワークが構成的な規則に基づいて汎化をしていれば，PP-IOBJ においては間接目的語の NP の PP 修飾に関する統語構造を用いており，削除後には汎化精度は 0%近くまで低下することが期待される．しかし，間接目的語の NP の PP 修飾に関する句構造の特徴のみを削除したときには，サブネットワークの汎化精度は低下したが，依然として 10%から 40%程度の完全一致率を維持した．そのため，構成的汎化のタスクにおいてサブネットワークは統語的特徴は用いている一方，非構成的な解法にも同時に依存していることが示された．

5.3　学習中の推移

次に，PP-IOBJ でモデルの汎化精度が学習を通じてどのように変化したかを分析する．図 5 の通り，機械翻訳ではベースモデルの精度は 200 エポック付近で上昇が止まり，サブネットワークは 500 エポックまで上昇し続けた．意味解析では同様の傾向は見られず，サブネットワークも比較的少ないエポック数で精度が向上し，収束した．この傾向の違いは，意味解析の方が構造がより明示されている出力形式であることが原因と考えられる．

また，機械翻訳・PP-IOBJ におけるサブネットワークの内部機序の推移を調べる．図 6 に各エポック数のサブネットワークに対して因果分析を行った結果を示す．統語的特徴を削除したサブネットワークはいずれも 200 エポック以降は汎化精度に大きな変化がなかった．意味解析や PP-SUBJ でも同様の傾向が見られた．この結果は，非構成的な解法は学習の初期の段階で形成され，学習が進み構成的汎化のタスクの精度が向上する段階でも失われなかったことを示す．さらに，統語的特徴を削除しないサブネットワークの精度は 200 エポック以降も向上したことと合わせると，徐々に統語的特徴を用いる構成的解法を獲得したことが示唆される．

6　おわりに

本研究では，Transformer の構成的汎化のタスクにおける動作原理をサブネットワークの探索や因果分析により調べた．実験の結果から，構成的汎化のタスクを解く際に Transformer は一定程度統語的特徴を用いているが，非構成的な特徴にも依存していることが示された．構成的汎化のタスクで優れた精度を出すサブネットワークでさえも，構成的な規則に基づいて汎化をしているとは限らないことがわかった．今後は，構成的汎化の他のパターンや他の言語的特徴への依存の分析を行い，Transformer の構成的汎化のメカニズムをより詳細に明らかにしたい．

謝辞

本研究は JST さきがけ JPMJPR21C8 の支援を受けたものである.

参考文献

[1] Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9087–9105, Online, November 2020. Association for Computational Linguistics.

[2] Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. On compositional generalization of neural machine translation. 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. 4767–4780, Online, August 2021. Association for Computational Linguistics.

[3] Verna Dankers, Elia Bruni, and Dieuwke Hupkes. The paradox of the compositionality of natural language: A neural machine translation case study. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4154–4175, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[4] Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. SLOG: A structural generalization benchmark for semantic parsing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3213–3232, Singapore, December 2023. Association for Computational Linguistics.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, 2017.

[6] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A primer on the inner workings of transformer-based language models, 2024.

[7] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2024.

[8] Yuekun Yao and Alexander Koller. Structural generalization is hard for sequence-to-sequence models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 5048–5062, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[9] Michael Y. Hu, Chuan Shi, and Tal Linzen. Compositional cores: Persistent attention patterns compositionally generalizing subnetworks. In **Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP**, 2024.

[10] Najoung Kim, Tal Linzen, and Paul Smolensky. Uncontrolled lexical exposure leads to overestimation of compositional generalization in pretrained models, 2022.

[11] Ryoma Kumon, Daiki Matsuoka, and Hitomi Yanaka. Evaluating structural generalization in neural machine translation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 13220–13239, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[12] 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.

[13] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. In **Advances in Neural Information Processing Systems**, 2023.

[14] Steven Cao, Victor Sanh, and Alexander Rush. Low-complexity probing via finding subnetworks. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 960–966, Online, June 2021. Association for Computational Linguistics.

[15] Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1513–1528, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[16] Adithya Bhaskar, Dan Friedman, and Danqi Chen. The heuristic core: Understanding subnetwork generalization in pretrained 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. 14351–14368, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

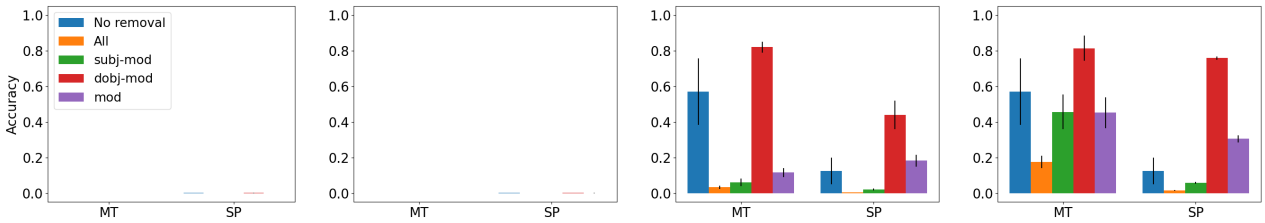
[17] Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A. Smith, Navin Goyal, and Yulia Tsvetkov. Learning syntax without planting trees: Understanding when and why transformers generalize hierarchically, 2024.

[18] Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. Universal semantic parsing. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 89–101, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[19] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. **Computational Linguistics**, Vol. 19, No. 2, pp. 313–330, 1993.

[20] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. Universal Dependency annotation for multilingual parsing. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 92–97, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[21] Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. The devil is in the detail: Simple tricks improve systematic generalization of transformers. 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. 619–634, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.



(a) 句構造 (Base)　(b) 依存構造 (Base)　(c) 句構造 (Sub.)　(d) 依存構造 (Sub.)

図 7　PP-SUBJ の因果分析の結果．各棒は対応する特徴の削除後の分布外評価データでの完全一致率を示す．All は統語的特徴の完全な削除，subj-mod は主語の NP への PP 修飾に関する統語的特徴のみの削除（dobj-mod も同様），mod は NP への PP 修飾に関する統語的特徴の削除を指す．

A　Subnetwork probing の詳細

subnetwork probing [14] は対象とするタスクを解くサブネットワークを探すためにマスクを学習する．$\phi \in \mathbb{R}^d$ をモデルの重み，$Z_i \in [0,1]$ を重み ϕ_i のマスクとする．Z_i は温度 β_i と確率変数 θ_i をパラメータに持つ hard concrete function に従う．

$$U_i \sim \mathrm{Unif}[0,1]$$
$$S_i = \sigma\left(\frac{1}{\beta}\left(\log\frac{U_i}{1-U_i} + \theta_i\right)\right)$$
$$Z_i = \min(1, \max(0, S_i(\zeta-\gamma)+\gamma))$$

$\zeta = 1.1, \gamma = -0.1$ に固定されている．

subnetwork probing はマスクのパラメータである θ を以下の損失関数を最小化することで最適化する．

$$\frac{1}{|D|}\sum_{(x,y)\in D} \mathbb{E}_{U_i\sim\mathrm{Unif}[0,1]} L(f(x;\phi * z(U,\theta)), y) + \lambda \mathbb{E}|\theta|_0$$

第一項は $Z_i = z(U_i, \theta_i)$ でマスクされたモデル f の損失関数で，第二項は値が 0 でないマスクに対するペナルティ項である．推論時には，マスク Z_i は閾値に基づいて $\{0,1\}$ に二値化する．

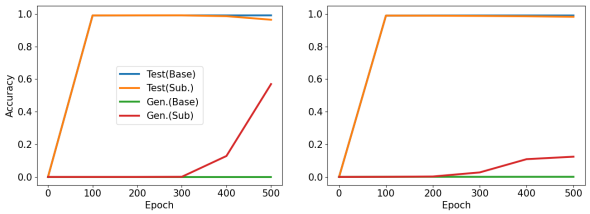
B　学習の詳細

ベースモデルの学習では，学習率を 0.0001，エポック数を 500，バッチサイズを 256，weight decay を 0.1 とした．[21] に従い，early stopping は用いなかった．Subnetwork probing におけるマスクの学習では，学習率を 0.0005，エポック数を 300，バッチサイズを 256 として，early stopping は用いなかった．

C　PP-SUBJ の結果

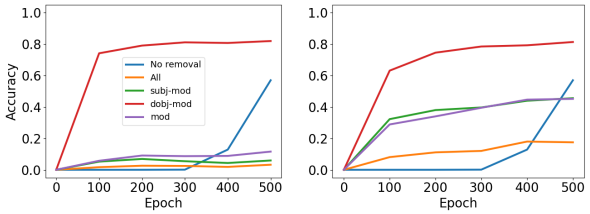
PP-SUBJ における因果分析の結果を図 7 に示す．まず，PP-IOBJ とは異なり，サブネットワークでさえ汎化精度は 100%から程遠いため，常に正しい構成的な解法が学習されているとは考えにくい．一方で，主語の NP の PP 修飾に関する統語的特徴を削除したときには，ほとんどの場合で 0%近くまで精度が低下した．そのため，少なくとも正しい出力をしているときには，サブネットワークは構成的な解法を用いることがわかる．また，直接目的語の NP の PP 修飾の統語的特徴を削除したときには，精度が向上したため，過学習が汎化精度のさらなる向上を阻んでいたことが示唆される．

PP-SUBJ におけるモデルの精度の推移と各エポック数のサブネットワークの因果分析の結果を図 8, 9 に示す．PP-IOBJ と同様に，統語的特徴を削除したサブネットワークは 200 エポック経過後には精度が大きく変化しないのに対し，元のサブネットワークは 200 エポック以降も精度が向上し続けた．



(a) 機械翻訳　(b) 意味解析

図 8　PP-SUBJ における学習中のモデルの完全一致率の推移．Test と Gen. はそれぞれ分布内，分布外評価データでの精度に対応する．Base と Sub. はそれぞれベースモデルとサブネットワークの精度に対応する．



(a) 句構造（機械翻訳）　(b) 依存構造（機械翻訳）

図 9　PP-SUBJ における各エポック数のサブネットワークの因果分析の結果．各線は対応する特徴の削除後の分布外評価データの完全一致率を示す．