ニューラル機械翻訳モデルにおける構成的汎化能力の評価

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

概要

構成性は，複合的な表現の意味がその部分の意味と全体の構造をもとに定まるという言語の性質である．ニューラルモデルの構成的汎化能力の評価は主に上流タスクで行われており，機械翻訳のような下流タスクにおける評価は十分でなく，特に文の構造に関する汎化についての研究は少ない．そこで本研究では，様々な種類の構成的汎化能力を問う翻訳データセットを構築し，機械翻訳タスクにおけるニューラルモデルの構成的汎化能力を評価する．結果から，ニューラルモデルの構成的汎化能力には改善の余地があることを明らかにした．

1　はじめに

構成性は，全体の意味は部分の意味とその統語的な組み合わせ方により定まるという，言語の重要な性質の一つである [1]．見たことのない複合表現を既知の単語や統語構造から理解し，生成することができるという構成的汎化能力は，言語理解において重要な能力の一つである [2]．

近年，ニューラルモデルが様々な自然言語処理タスクを高い精度で解く能力を持つことが示されているが，構成的汎化能力を持つかどうかに関しては盛んに議論がされている．これまでに，様々な種類の構成的汎化の評価のためのベンチマークや評価手法が複数提案されている [3, 4, 5]．ニューラルモデルの構成的汎化能力に関するこれらの既存の研究は，主に意味解析のような上流タスクで行われており，上流タスクに限らない多角的な評価のために，機械翻訳をはじめとする下流タスクにおける評価手法が求められている．一方で，既存の機械翻訳タスクを用いた構成的汎化の評価手法のほとんどは，既知の単語の未知の組み合わせ方に関する汎化である，**語彙的汎化**のみに注目している [6, 7, 8]．そのため，既知の統語構造の未知の組み合わせ方に関する汎化である，**構造的汎化**を必要とする文を機械翻訳モデルがどの程度正しく翻訳できるかは明確でない．

そこで本研究では，語彙や文の構造をコントロールした上，構造的汎化を中心に様々な種類の構成的汎化能力を問う日英対訳データセットを構築する．[1)] また，構築したデータセットを用いて，ニューラル機械翻訳モデルの構成的汎化能力を評価する．評価結果から，機械翻訳における構成的汎化は現在のニューラルモデルには難しいことを示す．

2　関連研究

ニューラルモデルの構成的汎化能力の研究の多くは，主に意味解析タスクで行われてきた．Lake と Baroni [3] は，限られた文法で生成された文章を行動の指示の単語列に変換するタスクを対象とする SCAN というデータセットを用いて構成的汎化を評価した．Kim と Linzen [4] は，SCAN よりも広範囲の語彙と統語構造をカバーした文章を意味表現に変換するタスクを対象とする COGS というデータセットを提案した．COGS では，意味解析タスクにおける語彙的汎化と構造的汎化を評価したが，LSTM [9] と Transformer [10] の両方とも構造的汎化の精度は低かった．また，COGS の構造的汎化パターンの拡張 [5]，データセットの構築手法 [11] や実験設定 [12] の改善に取り組んだ研究も存在する．

一方で，複数の研究が機械翻訳タスクで構成的汎化を評価を行っている．Li ら [6] は，学習データ中に含まれる語彙からなる未知の組み合わせが含まれる文の翻訳精度を調べた．Dankers ら [7] は，文の要素の入れ替えや文同士を連結したものの翻訳精度を調べることで，モデルの体系的汎化能力を評価した．また，Dankers らはテンプレートで合成的に構築したデータセットでの評価に加え，自然言語に含まれる語彙や統語構造の多様さを反映するために，コーパスを用いた評価も行った．Moisio ら [8] は，

1）　データセットは後日研究利用可能な形で公開する予定である．

Dankers らと同様に，実世界の自然言語のデータでの評価のために，分布外評価データの翻訳に構成的汎化を必要とするような翻訳コーパスの学習データと分布外評価データへの分割を，単語の組み合わせの出現分布を制御して行った．しかし，これらの研究は，ニューラルモデルの語彙的な汎化能力のみを評価しており，COGS や SLOG [5] に見られるような統語構造に関する汎化，すなわち構造的汎化を調べていない．また，構造的汎化を評価する上では，学習データに含める語彙や統語構造の厳密なコントロールが必要であるため，分布に基づくコーパスの分割は適していない．そのため，本研究では形式文法を用いた手法（詳細は次節）でデータセットを構築する．

3 提案手法

ニューラルモデルの機械翻訳タスクにおける構成的汎化能力の評価のために，英語および日本語の対訳からなるデータセットを構築する．データセットは，学習データ，検証データ，分布内評価データ，分布外評価データの 4 つからなる．分布外評価データには，学習データ，検証データ，分布内評価データに含まれない語彙や統語構造を持つ文が含まれ，学習データ中の語彙や統語構造からの構成的汎化によってのみ，正しい翻訳が得られる．モデルの学習後に，検証データを用いたバリデーションで最も高い性能を示したチェックポイントに関して，分布内評価データと分布外評価データのそれぞれで性能を評価し，比較する．

例えば，表 1 の「主語 → 目的語（普通名詞）」に注目する．学習データでは，*goat* は主語にのみ出現し，分布外評価データに *goat* を目的語に持つ文が含まれる．このとき，目的語に *goat* を持つ文を正しく翻訳できるかを調べることで，既知の単語が未知の文中の文法関係を持つ場合の汎化を評価できる．

データセットの構築は，汎化パターンの定義 (3.1 項)，確率文脈自由文法（Probabilistic Context-Free Grammar; PCFG）を用いた英語文生成 (3.2 項)，ルールベース機械翻訳による対訳の作成 (3.3 項) により行った．また，PCFG を用いて文を生成する際に，文の一部に関して選択制約を考慮する (3.4 項)．

3.1 汎化のパターン

COGS や SLOG が意味解析タスクで用いたものを参考に，翻訳タスクで評価対象とする汎化パターンを 76 種類設計した．本研究で評価対象とする汎化のパターンの一部を表 1 に示す．また，既存研究にない汎化パターンとして，形容詞の再帰の深さに関する汎化と 2 種類の時制に関する汎化を評価対象に加える．形容詞の再帰の深さに関する汎化は，文中の形容詞の反復回数が学習データと異なる文への汎化である．時制に関する汎化の一つ目は，特定の動詞の過去形から現在形への汎化である．学習データに異なる動詞の現在形が含まれるため，動詞の接尾辞に関する汎化が要求される．二つ目は，特定の動詞の過去形と，異なる文型における現在形から，特定の文型での現在形への汎化である．学習データに該当の動詞の現在形が含まれるため，一つ目の種類より易しいが，文型に関する汎化が要求される．

また，ReCOGS [11] に従い，一部の汎化のパターンをより公平に評価するために，学習データに 2 つの変更を加える．一つ目は，学習データにおいて，直接目的語を修飾する前置詞句または関係詞節を含む文の 5%を主題化する．これは修飾句・節が先頭に出現する文を学習データに含めるためである．二つ目は，学習データから，連結することで分布外評価データの最も長い文より長くなるような文を 2 つランダムに取り出し，連結したものを 1000 件追加する．これは学習した文より長い文が分布外評価データに含まれると，モデルは長さに関する汎化を必要とするので，構成的汎化のみに関する正確な評価ができないためである．

3.2 PCFG を用いた英語文生成

データセット中の英語の文は，COGS や SLOG と同様に，PCFG を用いて生成する．PCFG を用いることで，あらかじめ定めた生成規則と語彙のみから文が生成されるため，データセットに出現する文の統語構造や語彙をコントロールすることができる．

我々のデータセットでは，語彙と生成規則は，COGS と SLOG で用いられているものに，新たに評価の対象とする汎化パターンに対応する語彙と生成規則を追加し，誤った用法がなされている語を削除した．また，分布外評価データ中の文の翻訳に構造的汎化を要求するために，分布内データと分布外評価データで異なる生成規則を用いる．

3.3 翻訳データの生成

データセット中の日本語の対訳は，PCFG で生成した英語の文をルールベースで機械翻訳すること

表 1 本研究の評価対象の汎化パターンの一部

種類	学習データ	分布外評価データ
既知の単語とその文中の文法関係の未知の組み合わせ		
主語 → 目的語（普通名詞）	The **goat** ate the apple.	The dogs found the **goat**.
修飾句と文法的役割の未知の組み合わせ		
直接目的語 → 主語	The child offered **the book in the corner** to the princess.	**The book beside the ladder** fell.
再帰の深さの増加・減少		
that 節補文の増加	The kid admired **that** Liam dreamed **that** the friend was helped by the horse.	Samuel believed **that** Liam thought **that** the men knew **that** the women packed the fig.
形容詞の増加	The teacher liked a **big blue** table.	The teacher liked a **rare big round blue wooden** table.
動詞の項の変更		
能動態 → 受動態	The man **moved** the car.	The tool was **moved**.
既知の統語構造の未知の組み合わせ		
間接目的語を修飾する関係詞節	The teacher found the student that ate an apple. / The teacher found the student that Liam liked.	The teacher found the student that Liam gave the book to.
既知の動詞と時制の未知の組み合わせ		
過去時制 → 現在時制	The boy **offered** the girl the game.	The boy **offers** the girl the game.
過去/現在/二重目的語 → 現在&二重目的語	The king **showed** the book./ The king **shows** the book./ The king **showed** a child the book.	The king **shows** a child the book.

で生成する．ルールベースの機械翻訳は，Wang と Hershcovich [13] の手法を用いる．英語の文の生成に用いた PCFG を反映したルールベースの翻訳により，汎化パターンに応じて語彙や統語構造をコントロールした日本語の対訳を得ることができる．

Wang と Hershcovich の手法では，原言語（翻訳前の言語）の生成規則，原言語の各生成規則から目的言語（翻訳語の言語）の生成規則への変換規則，そして原言語の語彙に対応する辞書を用いて，翻訳する．例えば，*The teacher ate an apple*．の翻訳を考える．言語間の生成規則の変換規則は，以下のように定められ，「先生がリンゴを食べた.」と訳される．

- S → NP VP: S→ NP が VP
- VP→ V NP: VP→ NP を V

我々のデータセットでは，原言語の生成規則は，3.2 項で定義した生成規則を用い，言語間の生成規則の変換規則と辞書に関しては，新たに作成する．

3.4 語彙項目の選定

PCFG では語彙項目同士の関係を考慮していないため，自然言語では出現しない不自然な文が生成される．そこで PCFG で生成された文を選択制約を考慮してより自然なものへと変換する．

我々のデータセットでは，無生物主語と動詞，動詞と直接目的語の組のみに関して選択制約を考慮する．選択制約を満たさない組が存在する場合には，選択制約を満たすように，無生物主語または目的語の名詞を入れ替える．また，選択制約を満たす語彙の組み合わせは，京都大学格フレーム辞書 [14] から日本語の動詞と名詞のペアを抽出することによって作成する．抽出する際，格はそれぞれの動詞に関して適切なものを選択する．

4 実験

4.1 実験設定

モデル 評価の対象とするモデルは，Vanilla Transformer と事前学習済みモデルである Llama2 [15] とする．Transformer は OpenNMT-py [16] を用いてスクラッチから学習を行う．また，日本語の単語ごとの分割に Sudachi [17] を用いる．Llama2 は `meta-llama/Llama-2-7b-hf`[2] を LoRA [18] でファインチューニングをしたモデルを用いる．また，Transformer, Llama2 ともにランダムに選んだシードで学習を 5 回行う．学習のハイパーパラメータは付録 A に示す．

データ 各データに含まれる文の数は，学習データに 43240 個，検証データに 5280 個，分布内評価

2) https://huggingface.co/meta-llama/Llama-2-7b-hf

表 2 データセット全体の評価結果．Exact は完全一致 (%) を意味する．

モデル	データセット	Exact	BLEU
Transformer	分布内評価	99.7	99.8
	分布外評価	36.8	77.6
Llama2	分布内評価	99.5	99.8
	分布外評価	74.0	93.0

データに 5280 個，分布外評価データは各汎化パターンごとに 1000 個で合計 76000 個である．

評価指標 モデルの翻訳の評価には，3 つの指標を用いる．一つ目は，モデルの翻訳文のうち，ルールベース翻訳文と完全一致する割合である．二つ目は，BLEU であり，これを SacreBLEU [19] を用いて求める．三つ目は，一部の汎化のパターンのみを対象とする評価で，翻訳文中の汎化対象の部分がルールベースの翻訳文中の該当部分と一致する割合（部分一致）である．この指標では，該当部分の文字列が一致しているだけでなく，汎化部分の文中の文法関係の一致も求められる．文中の文法関係の一致は，GiNZA [20] を用いた句構造解析を基に調べた．また，that 節補文の汎化のように，文全体が汎化の対象のときには，部分一致の評価は行わない．まとめると，完全一致と BLEU は，翻訳文の全体の精度を評価するための指標で，部分一致は，汎化の精度を評価するための指標である．

4.2 結果と分析

データセット全体の評価結果を表 2 に示す．Transformer と Llama2 ともに，完全一致と BLEU の両方で分布内評価データの方が分布外評価データよりも高い精度となった．これらの結果は，ニューラルモデルの機械翻訳における構成的汎化の性能に改善の余地があることを示す．また，分布外評価データでの性能では，Llama2 が Transformer を上回った．しかし，Kim ら [21] が指摘したように，事前学習モデルの学習に用いられたデータは語彙や統語構造がコントロールされたものではない．そのため，分布外評価データにおいて，Llama2 の方が Transformer より高い精度を出しているが，この結果から必ずしも Llama2 の方が構成的汎化能力が高いという結論には至らない．さらに，評価指標ごとの違いとして，BLEU における Transformer と Llama2 との差は完全一致における差よりも小さく，BLEU は構造の一致を反映できていないためと考えられる．

表 3 汎化パターンごとの評価結果の一部．CP recursion は that 節補文の再帰の深さ，PP recursion は前置詞句の再帰の深さ，Adj recursion は形容詞の再帰の深さを意味する．Exact は完全一致 (%)，Partial は部分一致 (%) を意味する．

パターン	モデル	Exact	BLEU	Partial
CP recursion	Transformer	7.4	82.8	—
1, 2, 4 → 5, 6	Llama2	69.1	96.3	—
CP recursion	Transformer	82.0	97.5	—
1, 2, 4 → 3	Llama2	97.0	99.5	—
PP recursion	Transformer	19.3	84.0	19.6
1, 2, 4 → 5, 6	Llama2	84.4	97.6	85.0
PP recursion	Transformer	85.5	97.6	88.6
1, 2, 4 → 3	Llama2	96.3	99.4	97.1
Adj recursion	Transformer	98.9	99.7	99.1
1, 2, 4 → 5, 6	Llama2	98.3	99.7	98.7
Adj recursion	Transformer	99.1	99.7	99.4
1, 2, 4 → 3	Llama2	99.0	99.7	99.3

また，再帰の深さの汎化パターンの評価結果を表 3 に示す．[3] that 節補文，前置詞句，形容詞の再帰の深さに関する汎化のうち，最も精度が高かったパターンは形容詞で，次に前置詞句が高かった．最も精度が低かったパターンは，that 節補文の再帰の深さの汎化だった．このように，再帰が深くなるにつれて文がより長くなるパターンほど，精度が低い傾向が見られた．似た傾向は意味解析タスクにおいて SLOG でも確認されている．一方で，新たに追加した形容詞の再帰の深さの汎化では，他のパターンと異なり，深さが増加するパターンでも精度が高かった．これは形容詞の再帰は語数が少ないことに加え，英語と日本語の翻訳の語順が一致するためと考えられる．これらの結果は，ニューラルモデルは任意の深さの汎化能力が欠けているわけではなく，汎化の度合いは再帰の種類に依存することを示す．

5 おわりに

本稿では，構造的汎化を中心に様々な構成的汎化を要求する翻訳データセットを構築し，機械翻訳タスクにおいてニューラルモデルの構成的汎化能力を評価した．実験結果から，ニューラルモデルの構成的汎化能力に改善の余地があることが示唆された．また，意味解析タスクでは見られなかった，モデルの汎化の傾向を見ることができた．今後は，英日翻訳以外の別の言語間でどのような現象が観察されるかについて調べ，機械翻訳タスクにおけるニューラルモデルの構成的汎化能力を改善する手法を研究する予定である．

3) その他の汎化パターンの評価結果は付録 B に示す．

謝辞

本研究は JST さきがけ JPMJPR21C8，JSPS 科研費 JP20K19868 の支援を受けたものである．

参考文献

[1] Barbara Partee, et al. Compositionality. **Varieties of formal semantics**, Vol. 3, pp. 281–311, 1984.

[2] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. **Cognition**, Vol. 28, No. 1-2, pp. 3–71, 1988.

[3] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy and Andreas Krause, editors, **35th International Conference on Machine Learning, ICML 2018**, pp. 4487–4499. International Machine Learning Society (IMLS), 2018.

[4] Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9087–9105, Online, 2020. Association for Computational Linguistics.

[5] 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.

[6] 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.

[7] 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.

[8] Anssi Moisio, Mathias Creutz, and Mikko Kurimo. On using distribution-based compositionality assessment to evaluate compositional generalisation in machine translation. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, and Elia Bruni, editors, **Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP**, pp. 204–213, Singapore, December 2023. Association for Computational Linguistics.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, p. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[11] Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. ReCOGS: How incidental details of a logical form overshadow an evaluation of semantic interpretation. 2023.

[12] 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.

[13] Zi Wang and Daniel Hershcovich. On evaluating multilingual compositional generalization with translated datasets. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1669–1687, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] 河原大輔, 黒橋禎夫ほか. 高性能計算環境を用いた web からの大規模格フレーム構築. 情報処理学会研究報告自然言語処理 (NL), Vol. 2006, No. 1 (2006-NL-171), pp. 67–73, 2006.

[15] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[16] Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. Opennmt: Neural machine translation toolkit, 2018.

[17] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Paris, France, may 2018. European Language Resources Association (ELRA).

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations (ICLR)**, 2022.

[19] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[20] Hiroshi Matsuda, Mai Omura, and Masayuki Asahara. 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. 言語処理学会第 25 回年次大会, 2019.

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

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **International Conference on Learning Representations (ICLR)**, 2015.

表 4　その他の汎化パターンの評価結果（that 節内に汎化対象を入れた汎化パターンは除く）．Exact は完全一致 (%), Partial は部分一致 (%) を意味する．

		Transformer			Llama2		
カテゴリ	パターン	Exact	BLEU	Partial	Exact	BLEU	Partial
単語 × 文法関係	主語 → 目的語 (普通名詞)	99.6	99.8	99.9	98.3	99.4	98.6
	主語 → 目的語 (固有名詞)	99.6	99.8	100.0	93.7	97.5	93.8
	目的語 → 主語 (普通名詞)	99.5	99.7	100.0	99.1	99.6	99.9
	目的語 → 主語 (固有名詞)	99.5	99.7	100.0	98.8	99.6	99.7
	単語 → 主語 (普通名詞)	0.0	0.1	0.0	62.7	94.4	63.1
	単語 → 主語 (固有名詞)	0.0	0.2	0.0	98.0	99.5	98.8
	単語 → 目的語 (普通名詞)	15.1	47.3	15.3	68.5	88.2	68.6
	単語 → 目的語 (固有名詞)	0.2	34.4	0.2	66.3	88.3	66.6
修飾句 × 文法関係	前置詞句 in 主語	0.0	66.1	0.2	35.7	78.5	38.9
	前置詞句 in 間接目的語	22.4	70.4	27.3	91.2	96.9	93.5
	関係詞節 in 主語	0.0	46.2	0.6	22.7	69.7	27.1
	関係詞節 in 間接目的語	6.9	62.2	16.8	56.9	84.8	59.7
	形容詞 in 主語	46.3	70.5	48.6	92.5	98.2	92.7
	形容詞 in 間接目的語	49.1	81.1	54.8	97.8	99.4	98.4
再帰の深さ	中央埋め込みの再帰の深さ 1, 2, 4 → 5, 6	6.1	76.8	—	59.9	91.7	—
	中央埋め込みの再帰の深さ 1, 2, 4 → 3	85.7	97.4	—	95.7	98.8	—
動詞の項	能動態 → 受動態	0.7	37.2	0.7	94.0	96.9	94.2
	受動態 → 能動態	0.0	58.0	0.0	71.1	87.3	71.8
	目的語省略 他動詞 → 他動詞	78.8	91.0	80.0	54.1	77.4	59.9
	非対格動詞 → 他動詞	18.0	70.2	18.1	50.0	79.8	50.5
	二重目的語 → 前置詞句	99.9	100.0	100.0	97.4	99.3	97.8
	前置詞句 → 二重目的語	98.8	99.5	100.0	99.5	99.8	99.8
未知の部分構造	間接目的語の引き抜き（関係詞節）	71.8	88.1	71.9	46.0	84.8	46.2
	間接目的語の引き抜き（wh 疑問文）	0.0	46.4	—	0.0	75.9	—
wh 疑問文の構造	能動態の主語（wh 疑問文）	75.9	91.7	—	85.9	94.6	—
	受動態の主語（wh 疑問文）	54.1	84.2	—	53.6	80.6	—
	直接目的語（wh 疑問文）	52.6	80.8	—	81.6	94.5	—
	主語 with 前置詞句（wh 疑問文）	0.1	45.6	—	67.8	85.0	—
	長距離移動（wh 疑問文）	0.0	43.5	—	0.0	74.7	—
時制	現在形（二重目的語）	4.2	55.4	4.6	98.8	99.7	99.7
	現在形（不定詞）	0.0	60.0	1.4	32.2	76.1	32.5
	現在形（that 節）	2.5	68.0	3.4	96.5	99.0	97.6
	現在形（他動詞）→ 現在形（二重目的語）	78.2	90.3	100.0	99.3	99.7	100.0
	現在形（他動詞）→ 現在形（不定詞）	5.5	65.9	6.0	61.8	84.5	61.9
	現在形（他動詞）→ 現在形（that 節）	86.9	96.0	93.9	98.1	99.5	98.8

A　学習のハイパーパラメータ

Transformer　Transformer は 6 層のエンコーダー，6 層のデコーダーと 8 個の注意機構ヘッドを持つモデルを用いる．さらに，Csorda ら [12] に従い，相対位置埋め込みを入れた Transformer を用い，early stopping と label smoothing は用いない．最適化手法として Adam [22] を採用し，学習のハイパーパラメータは，学習率を 1e-4，バッチサイズを 256，学習ステップ数を 70000 とする．

Llama2　Llama2 のファインチューニングのハイパーパラメータは，学習率を 1e-4，LoRA ランクを 8，α を 32，ドロップアウトを 0.1，エポック数を 8 とする．

B　結果

4 節に示したパターン以外の評価結果は表 4 に示す．