従属節が分断された不可能言語を言語モデルは学習するのか

指田 昌樹 [1]　鈴木 彩音 [1]　安田 卓矢 [1]　染谷 大河 [1]　谷中 瞳 [1]
[1] 東京大学
masaki.sashida@weblab.t.u-tokyo.ac.jp
takuya-yasuda@fintech-lab.jp
{suzuki-akane079,taiga98-0809,hyanaka}@g.ecc.u-tokyo.ac.jp

概要

言語モデルの言語獲得能力は人間とどのように異なるかという問いに対して，人間にとって不自然な言語（不可能言語）を人工的に生成し，言語モデルの学習のしやすさを分析した研究がある．先行研究では，GPT-2 モデルにおいて不可能言語は可能言語よりも学習しにくいという結果が報告されているが，言語に共通する性質は単一ではなく，不可能言語についても様々な種類が存在するため，言語モデルがどのような不可能言語においても学習が難しいのかという問題については，依然として多くの疑問が残されている．本研究では，「最下層の従属節は他の節によって分断されない」という言語に共通する性質に着目し，従属節の間に主節の一部を混ぜた不可能言語を作成した．作成した不可能言語を用いて GPT-2 モデルを学習し，可能言語と同様に不可能言語を学習できるかどうか分析を行った．GPT-2 モデルにおいて可能言語と比較して不可能言語の学習は困難であるという結果となり，先行研究と同様の傾向を示した．

1　はじめに

2022 年に ChatGPT [1] が公開され，汎用性の高さから様々な自然言語処理のタスクで応用が進んでおり，並行して言語モデルの言語獲得能力を分析する研究も行われている．

言語学者からは人間と言語モデルの言語獲得能力が異なるという指摘がされており，例えば，Chomsky らは，大規模言語モデル（LLM）は，自然言語と自然言語に見られない性質をもつ言語（不可能言語）を分け隔てなく学習することができると主張している [2, 3]．また，Bolhuis らは，LLM は不可能言語であっても生成することができると主張している [4]．

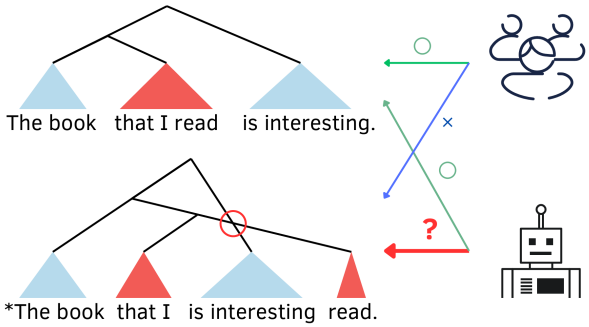

図 1　最下層の従属節が他の節によって分断されたような不可能な言語を，言語モデルは学習するのか

Kallini ら [5] は，不可能言語をいくつかのルールに従って作成し，言語モデルが可能言語と同様に不可能言語を学習できるか実験することで，言語学者の人間と言語モデルの言語獲得能力が異なるという主張を検証した．Kallini らは，トークンを規則的に並べ替えたり，動詞の一部を一定数のトークンの後に配置するなどして不可能言語を作成し，GPT-2 モデルにおいて不可能言語は可能言語よりも学習しにくいという実験結果を得た．しかし，言語に共通する性質は単一ではなく，不可能言語についても様々な種類が存在するため，言語モデルがどのような不可能言語においても学習が難しいのかという問題については，依然として多くの疑問が残されている．

さらに，人間の言語を獲得するうえで言語の構造依存性が重要な役割をしているという言語学者の主張がある [6, 7] が，Kallini らの研究では構造依存性には必ずしも焦点があたっていなかった．

そこで本研究では，Kallini らの実験を参考にして，最下層の従属節は他の節によって分断されないという言語の性質に着目して不可能言語を作成し，言語モデルが可能言語と同様に作成した不可能言語を学習できるかどうか分析を行った（図 1）．

(a) 最下層の従属節は他の節によって分断されない

なお，2 節で詳細は述べるが，一般に言語は (a) の性質を持ち，また人間は言語学習の途中であっても (a) の性質を持たないような文章を生成しないことが知られている [6, 7].

本研究では初めに，BabyLM dataset [8] をもとに人工的に従属節の間に主節の一部を混ぜたような不可能言語とそれと対になる可能言語のデータセットを作成し，可能言語と不可能言語それぞれで GPT-2 モデル [9] の学習を行った．その後，学習したモデルそれぞれにおいて，どの程度可能言語 · 不可能言語の構造を学習できているのか比較した．その結果，GPT-2 モデルにおいて可能言語と比較して不可能言語の学習は困難である傾向がみられた．これは，言語モデルは不可能言語を学習しにくいという Kallini らの論文を支持する結果となった．

2　言語における従属節の非分断性

一般に，言語において最下層にある従属節は他の節の単語が入り込まないという性質がある．この性質は言語が階層構造を持ち，それに依存して言語操作が行われるという，言語の構造依存性から生まれる現象の一つである [10]．関係節を含む文を考えると，関係節は構造的に低い位置に存在しており，関係節外の主節における単語の移動は関係節より構造的に高い位置の操作となるため，関係節は分断されない．

また，Crain&Nakayama は，この性質は人間の言語獲得においてもみられることを経験的に検証した [6]．この実験では，英語を母語として学習している子供に “Ask Jabba if the man who is watching Mickey Mouse is happy.” と尋ねたときどのような回答をしたか分析を行っている．その結果，多くの子どもは (1) のような文を生成し，(2) のよう文は生成しなかったと報告している．

(1)　Is the man [who is watching Mickey Mouse] _ happy?

(2)　*Is the man [who _ watching Mickey Mouse] is happy?

このように，最下層の従属節は他の節によって分断されないという自然言語の性質に着目し，本研究では，それが満たされないような不可能言語を作成した．

3　実験設定

3.1　不可能言語の作成

1 節で述べたように，(a) の性質を持つ可能言語と (a) の性質を持たない不可能言語における言語モデルの挙動を調べるため，表 1 のような，従属節の主語句の直後に主節の述語句を挿入した不可能言語を作成した．

Kallini らに倣って，BabyLM dataset [8] をベースに不可能言語と，それと対になる可能言語を生成した．具体的にはまず，BabyLM dataset の中から，主節が従属節で分断された文を抽出した．次に，主節の内部に含まれない従属節が存在する場合は，その従属節を削除し，この文章を可能言語として取り扱った．この可能言語の従属節の動詞句を文末に移動し，不可能言語を作成した（表 1）．このような作成方法をとっているため，可能言語と不可能言語は対で作成され，また対応する文は語順が異なるだけとなる．可能言語と不可能言語，それぞれ学習データ 4.7 万文，検証データ 5713 文，テストデータ 6191 文を作成した．なお，構文解析には，spaCy を利用した [11].

3.2　言語モデルの学習

3.1 節で作成した可能言語および不可能言語のデータセットを用いて GPT-2 モデルの学習を行い，可能言語と同様に不可能言語を学習できるかどうか分析を行った．本研究では，GPT-2 モデルの軽量な実装として nanoGPT[1] を使用した．パラメータ数 124M の GPT-2 モデルを使用した．学習ステップ数ごとの検証データのクロスエントロピーロスの推移を測定し，学習する速度が可能言語と不可能言語でどの程度異なるのか比較を行った．

3.3　評価方法

本研究では，言語モデルの文の生成確率を用いて，言語モデルが言語の構造を学習しているか確認した．文の生成確率は，次の式で表される：

$$P(\text{sentence}) = \prod_{i=1}^{N+1} P(y_i \mid y_{<i})$$

ここで，

1）　https://github.com/karpathy/ nanoGPT, 2022

表 1　不可能な言語の作成

不可能言語の例	元の文章	The book that I read is interesting.
	不可能言語	The book that I is interesting. read
不可能言語の生成	元の文章	主節が従属節で分断された文章を抽出 The idea which you suggested sounds great though it needs testing.
	可能言語	元の文章の主節の外側にある従属節を削除 下記文章を可能言語のデータとして採用 The idea which you suggested sounds great
	不可能言語	可能言語の従属節の動詞句を文末に移動 下記文章を不可能言語のデータとして採用 The idea which you sounds great suggested

- N は文のトークンの総数（トークン列の長さ）を表す．
- y_i は i 番目のトークンを表す．ただし y_{N+1} は EOS (End of Sentence) トークンを表す．
- $P(y_i \mid y_{<i})$ は，言語モデルが文脈（すなわち y_1 から y_{i-1} までのトークン列）を条件として i 番目のトークン y_i を生成する条件付き生成確率である．

学習したモデルに対して，可能言語と不可能言語の生成確率を計算し，どちらの生成確率が高いかを比較することで，モデルが作成した言語の構造を学習しているか検証した．具体的には，可能言語を学習したモデルに対し，テストデータの 1 文とそれと対になる不可能言語のテストデータの 1 文について，それぞれの生成確率を計算し，その大小を比較した．本研究では，3.1 節のように可能言語と不可能言語をセットで作成しており，比較する際は単語の順番のみが異なる文同士の生成確率を比較している．可能言語および不可能言語のテストデータはそれぞれ 6191 文含まれており，各文について生成確率の比較を行った．このとき，もしモデルが言語の構造を正しく学習しているならば，可能言語を学習したモデルでは可能言語の生成確率が高く，不可能言語を学習したモデルでは不可能言語の生成確率が高くなることが期待される．6191 回の比較のうち，可能言語を学習したモデルが可能言語の生成確率を高く予測した割合，および不可能言語を学習したモデルが不可能言語の生成確率を高く予測した割合を，それぞれ正答率として定義し，モデルが言語の構造をどの程度学習しているかを検証した．

4　結果と分析

可能言語と不可能言語を学習したときの学習ステップ数ごとの検証データのクロスエントロピーロスと正答率の推移を図 4 に示す．学習時の検証データのクロスエントロピーロスについては大きな差がみられなかった．一方，可能言語と不可能言語の正答率の比較については，一貫して可能言語を学んだ言語モデルの方が高い正答率を示した．不可能言語でも一定程度正答率は上昇したものの，可能言語の方が高い正答率を少ない学習エポック数で達成するという結果となった．

可能言語と比較して不可能言語の構造の学習の方が GPT-2 モデルにとって困難であるという結果となった．この結果は，Kallini らの “GPT-2 モデルにとって可能言語と比べて不可能言語を GPT-2 モデルは学びにくい” という実験結果 [**?**] と同様の傾向を示している．可能言語においては，従属節において主語と動詞が隣接しており，GPT-2 モデルが主語と述語の対応関係を学習することが容易であった一方，不可能言語においては，従属節の主語と述語の間に主節の述語句が挿入されており，GPT-2 モデルが従属節の主語と述語の対応関係を学習することが困難であったという要因が考えられる．

5　追加実験

3.2 節では，単一の言語でモデルの学習を行ったが，ChatGPT などの LLM では多言語のテキストで事前学習を行っている．最下層の従属節は他の節によって分断されないという性質は，多くの言語で共通する性質であり，多言語のテキストで事前学習が行われた言語モデルは，未知の言語に対してもこの性質を汎化させるだろうと仮説を立て，それを実証

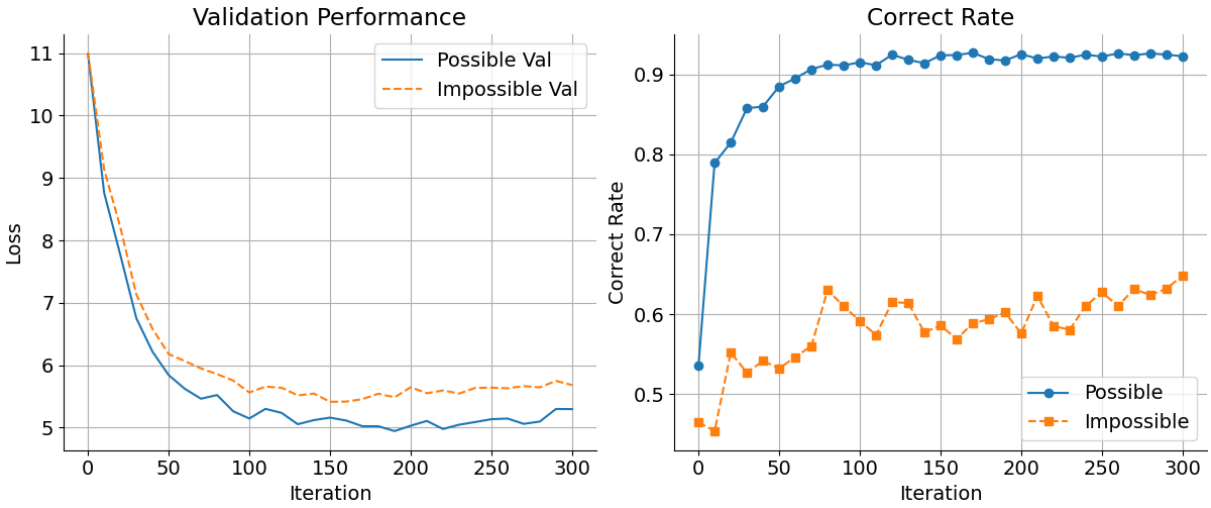

図 2　可能言語と不可能言語学習時の検証データのクロスエントロピーロス（左），および正答率（右）を示す．なお，正答率とは，可能言語を学習したモデルが，不可能言語よりも可能言語の方が生成確率が高いと判断した割合を指す（逆も同じ）

表 2　生成確率の比較結果（平均値）

	可能言語	不可能言語
GPT2	33 / 53	20 / 53
GPT2-large	32 / 53	21 / 53
GPT2-xl	40 / 53	13 / 53
Llama3.2-1B	39 / 53	14 / 53
Llama3.2-3B	34 / 53	19 / 53
Average	35.6 / 53	17.4 / 53

する追加実験を行った．

追加実験では，未知の言語に対する汎化性能を調べるために，英語ベースの可能言語 · 不可能言語を英語と認識されないように単語のダミー化を行っている（付録 A）．単語の品詞などの最低限の情報を学習させるため，主節のみを含む文章を BabyLM dataset から 1000 文抽出し，複数の事前学習済みモデルに事後学習させた．その後，3.1 節で作成した可能言語と不可能言語のうち，事後学習させた文に含まれる単語のみで構成された文を各 53 文抽出した．各 53 文について，3.2 節と同様に，生成確率をそれぞれ測定し，可能言語と不可能言語の生成確率がどちらが高いか比較した．なお，事前学習済みのモデルとしては GPT-2[2)]，GPT-2-large [3)]，GPT-2-xl [4)]，Llama3.2-1B [5)]，Llama3.2-3B [6)] を使用した．

その結果，表 2 のとおり，どのモデルでも，可能言語の生成確率の方が不可能言語の生成確率よりも高い傾向を示した．なお，GPT-2，GPT-2-large，GPT-2-xl では，9 回実験した結果の平均値を，Llama3.2-1B，Llama3.2-3B では 3 回実験した結果の平均値を表 2 に記載している．

これは，多言語のテキストで事前学習された言語モデルは，最下層の従属節は他の節によって分断されないという言語に共通する性質を，未知の言語に対しても汎化させているという仮説をサポートする結果となった．

6　おわりに

本研究では，最下層の従属節は他の節によって分断されないという言語に共通する性質に着目し，その性質を持たない不可能言語を生成し，GPT-2 モデルにおいて可能言語と比較して不可能言語の学習は困難であるかどうか分析を行った．その結果，GPT-2 モデルにおいて可能言語の構造の方が不可能言語よりも学習しやすいという結果となった．また，多言語のテキストで事前学習が行われた言語モデルは，最下層の従属節は他の節によって分断されないという言語に共通する性質を，未知の言語に対しても汎化することを実験で確かめた．今後，より多様な不可能言語を作成し，また，Transformer ベース以外の言語モデルにも対象を広げたうえで，言語モデルの言語獲得能力が可能言語に限定されるのかについて，さらなる分析を進めたい．

2）　https://huggingface.co/openai-community/gpt2
3）　https://huggingface.co/openai-community/gpt2-leargе
4）　https://huggingface.co/openai-community/gpt2-xl
5）　https://huggingface.co/meta-llama/Llama-3.2-1B
6）　https://huggingface.co/meta-llama/Llama-3.2-3B

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] OpenAI. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774**, 2023. version 3.

[2] Noam Chomsky. Conversations with tyler: Noam chomsky, 2023. Conversations with Tyler Podcast.

[3] Noam Chomsky, Ian Roberts, and Jeffrey Watumull. Noam chomsky: The false promise of chatgpt. **The New York Times**, 2023.

[4] Johan J. Bolhuis, Stephen Crain, Sandiway Fong, and Andrea Moro. Three reasons why ai doesn't model human language. **Nature**, Vol. 627, No. 8004, pp. 489–489, 2024.

[5] Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. Mission: Impossible 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. 14691–14714, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] Stephen Crain and Mineharu Nakayama. Structure dependence in grammar formation. **Language**, Vol. 63, No. 3, pp. 522–543, 1987.

[7] Massimo Piattelli-Palmarini, editor. **Language and Learning: The Debate Between Jean Piaget and Noam Chomsky**. Harvard University Press, 1980.

[8] Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors, **Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning**, pp. 1–34, Singapore, December 2023. Association for Computational Linguistics.

[9] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[10] Noam Chomsky. **Aspects of the Theory of Syntax**. MIT Press, Cambridge, MA, USA, 1965.

[11] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2018.

表 3　事後学習用データの語彙をスペシャルトークンに対応付け

元の単語	a	man	likes	pen
対応する文字列	[SPECIAL_TOKEN_0]	[SPECIAL_TOKEN_1]	[SPECIAL_TOKEN_2]	[SPECIAL_TOKEN_3]

表 4　事後学習用データをスペシャルトークンに変換したうえで事後学習を実施

a	man	likes	a	pen
[SPECIAL_TOKEN_0]	[SPECIAL_TOKEN_1]	[SPECIAL_TOKEN_2]	[SPECIAL_TOKEN_0]	[SPECIAL_TOKEN_3]

A　言語のダミー化

5 節の追加実験では，未知の言語に対する汎化性能を調べるために，英語ベースの可能言語 · 不可能言語を英語と認識されないように単語のダミー化を行っている．まず，表 3 のように事後学習用データにおいて出現する語彙を “[SPECIAL_TOKEN_0]” などの文字列と対応付けた．その後，事前学習済みのモデルのトークナイザに新規の単語として “[SPECIAL_TOKEN_0]” などを登録し，トークナイザが “[SPECIAL_TOKEN_0]” という文字列を新たに追加した 1 つの単語として認識することを確認した．その後，表 4 のように事後学習させる文を変換したうえで，事前学習済みのモデルで学習を行った．(3) 文は (4) の文字列に変換される．

(3)　a man likes a pen

(4)　[SPECIAL_TOKEN_0] [SPECIAL_TOKEN_1] [SPECIAL_TOKEN_2] [SPECIAL_TOKEN_0] [SPECIAL_TOKEN_3]

最後に，3.1 節で作成した可能言語と不可能言語のうち，事後学習させた文に含まれる単語のみで構成された文を各 53 文抽出し，“[SPECIAL_TOKEN_0]” などに変換したうえで，事後学習を行ったモデルに入力し，生成確率を算出した．