小規模言語モデルによる子供の過剰一般化のモデリング

芳賀あかり [1]　菅原朔 [2]　深津聡世 [3]　大羽未悠 [1]
大内啓樹 [1]　渡辺太郎 [1]　大関洋平 [3]
[1] 奈良先端科学技術大学院大学　[2] 国立情報学研究所　[3] 東京大学
{haga.akari.ha0,oba.miyu.ol2,hiroki.ouchi,taro}@is.naist.jp
saku@nii.ac.jp　{akiyofukatsu,oseki}@g.ecc.u-tokyo.ac.jp

概要

近年，言語モデルの学習データを子供の入力に近づけることで学習効率向上が示されるなど，人間の言語獲得過程を模倣することの有効性が示唆されている．一方で子供の一般化の特徴を機械で説明しようという研究は長らくされているが，その多くは語の屈折を直接学習するなど子供の学習環境とは異なるものである．本研究は，子供が言語獲得過程で起こす誤りを模倣することで言語モデル (LM) が人間らしい学習過程を得られるという仮説のもと，子供の入力に近いデータで学習した LM の誤りの選好を分析した．その結果，子供の入力に近い学習データを用いるのみでは人間の学習過程の誤りの特徴は十分に捉えられない可能性が示唆された．

1　はじめに

近年の大規模言語モデル (LLM) の精度は飛躍的に向上しているものの，いまだに様々な誤りを起こす．しかしニューラルモデルのブラックボックス性から，誤りがどうして起きるのかは現在自明でないことがほとんどであり，その誤りを制御するための確立された知見は少ない．さらに，現在の LLM は学習に膨大なデータを必要とするため，データ収集コストや学習時間の観点から問題視されている．一方で乳幼児は低リソースで母語を獲得できることが知られており，LLM の一つである GPT-3[1] は学習に約 2,000 億語必要とするのに対し，乳幼児は 1 億語程度の学習量しか必要としないと言われている [2]．近年の研究では人間の言語獲得を模倣することによる効果が示されており，子供向けの語彙や Child-directed-speech (CDS) を学習データに用いることで学習効率が向上することがわかっている [3][4]．

他方で，心理言語学や認知言語学の分野では子供の誤用やそれに伴って U 字を描く学習曲線についての研究が多くされており，これらが子供の言語獲得プロセスについて示唆を与えると言われている [5]．

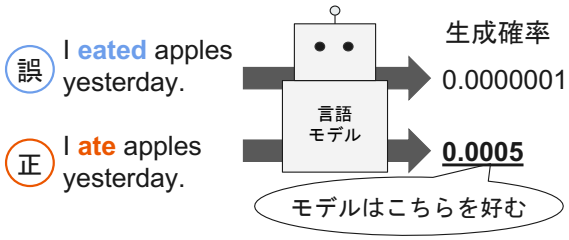

図 1: ミニマルペアデータを用いた誤りの選好の評価

以上の背景から，本研究では LM で人間の誤りを模倣することで人間らしい学習過程を得られる可能性があるという仮説を立てる．また，人間らしい学習過程の実現に付随して，既存研究で示されているような学習効率の向上も期待できる．仮説の検証に向けた第一段階として，本論文では，既存の LM が人間の誤りの特徴をどの程度捉えているか分析するため，(1) 人間らしい学習曲線を描くか，(2) 間違いの種類は人間らしいかを観察する．

子供の言語獲得過程に見られる代表的な特徴の一つとして，得た言語知識を過剰に一般化することが知られており，例えば英語の動詞の過去形の屈折の学習は過剰な一般化を観察するためによく取り上げられる現象である．本研究では，英語の動詞の過去形の過剰な一般化に注目し，CDS を学習データに用いたニューラル LM の，英語の動詞の屈折の学習過程を観察する．具体的には，過剰な一般化が含まれる文と含まれない文のペアを用意し，図 1 に示すようなミニマルペアデータを用いた LM の選好の観察によって各ステップでモデルがどのような誤りを好むか評価を行った．その結果，CDS で学習されたモデルよりも比較対象として用意した Wikipedia で学習されたモデルの方がより子供に近い U 字の学習曲線を示した．さらに，誤りの種類の選好はどの学習データでも子供とは異なる傾向を示し，言語モデル

の学習過程での誤りの選好は人間の特徴を十分に捉えていない可能性が示唆された．

2 関連研究

人間の学習過程の模倣による効果 近年の研究では人間の学習過程の模倣による効果が示されており，Huebner らは，人間が 6 歳までにふれるテキストを時系列順並べたデータで学習するモデル BabyBERTa を提案し，学習効率向上に成功している [3]．Eldan らは，GPT-3.5 および GPT-4 によって生成された，3〜4 歳児が通常理解できる単語のみを含む短編小説のデータセット TinyStories でモデルを学習することで学習の効率向上に成功している [4]．

子供の言語獲得過程で見られる一般化の特徴 子供の言語獲得過程の代表的な特徴の一つは過剰に一般化を行うことである．例えば英語には規則変化動詞と不規則変化動詞があるが，子供は学習段階のある時点で多くの不規則変化動詞の過去形を規則変化形に一般化してしまう．つまりほとんど全ての動詞に-d, -ed を加えるという過剰な一般化が見られる．このような特徴から，子供の言語獲得には (1) 暗記を行う (2) 過剰な一般化を行う (3) 不規則/規則変化形両方を正しく使えるようになる，という 3 つのステージがあると言われている [6]．重要なのは，ステージ (2) では新規単語に限らず，今まで正しく産出できていた単語も間違えるようになることである．そのため，子供の学習曲線は U 字を描くと言われている [7]．本研究では LM も同様に U 字の学習曲線を描くか分析を行う．

子供の言語獲得のモデリング 子供の言語獲得，特に過剰な一般化が見られる代表的な現象である英語の過去形の獲得を機械でモデリングしようという試みは長らくされてきている [8]．1986 年には Rumelhart らが英語の原形から過去形への変換をニューラルモデルで学習し，子供のようなエラーの傾向や U 字の学習曲線が観察された [7]．これに対し，Pinker らはこのモデルに多くの欠陥があることを示した [9]．しかし，Kirov らは近年のニューラルモデルを用いることで Pinker らの反論のほとんどが解消されることを主張している [10]．

最近のニューラルモデルの進化により，ニューラルネットワークが子供の言語獲得過程をどの程度捉えることができるかに再び関心が集まっており，2019 年以降も近年のニューラルモデルを用いて子供の言語獲得を模倣する研究はいくつかされてきているが，どの研究も，動詞の原形から過去形への変換を直接学習しており，これは人との対話や大人同士の会話から学習する子供の言語獲得とは全く異なる設定である [11]．本研究では，子供の自然な言語獲得過程をモデリングするために，語の屈折の直接の学習は行わず，CDS で学習した言語モデルを用いる．

3 提案手法

本研究では，CDS を学習データに用いてニューラル LM を学習し，学習の各ステップでどのような一般化を好むか観察する．先行研究では動詞の原形から過去形への直接の変換を学習しているため，生成された過去形を評価するだけで十分だった．しかし，本研究では過去形の生成を直接学習しないため，別の評価方法が必要である．よって，図 1 に示すような，生成確率の比較によってモデルの選好を評価する，ミニマルペアに対する relative acceptability judgments を採用する．本研究では子供の誤りの選好を評価するため，過剰汎化を起こしている文と，そうでない文のペアを含む，子供の誤りのミニマルペアデータを作成する．モデルにこのペアを強制的に生成させたとき，過剰汎化を含む文の生成確率がより高い場合は過剰汎化を好む，低い場合は過剰汎化を好まないと評価する．このような評価方法は，数年前に LM プロービングに初めて適用されて以来，広く普及している評価手法である [12]．本研究で作成するデータは構文評価のためミニマルペアデータセットである BLiMP[13] を参考にしたものであり，BLiMP の文法生成手法をベースに過剰汎化用のスクリプトを実装し，一つの現象につき 1,000 ペアの自動生成を行う．

本論文では，子供の過剰な一般化が見られる代表的な現象である「英語の動詞の過去形の屈折」に着目する．不規則変化動詞について，まず過剰汎化形 (write→writed など) と正しい過去形 (wrote) のペアを作成する．次に作成した過剰汎化形を含んだ文と，正しい過去形を含んだ文のペアを作成する．過剰汎化形とは全ての動詞の原形に「-d」または「-ed」を加えた形とし，ルールベースで作成する．

4 実験設定

本実験では，CDS で学習可能なことが示されている BabyBERTa を採用し，CDS を学習データとして実験を行う．学習データの比較のため，Wikipedia

のデータを追加した以下の 3 パターンの小規模データで学習を行う．

(i) AO-CHILDES [14]
(ii) Wikipedia [3]
(iii) AO-CHILDES+Wikipedia

(i)，(ii) は Huebner の論文で使用されている訓練データと同じ設定を採用しており，(i) の AO-CHILDES とは子供と大人の会話を記録したデータセットである CHILDES[15] から，約 500 万語のアメリカ英語による子供に向けた発話が年齢順に書き起こされているデータである．(ii) は，英語の Wikipedia コーパスから無作為に 50 万文収集したデータである．(iii) は (i) と (ii) を結合しシャッフルしたものである．各コーパスでは全ての文を小文字化し，3 単語より短い文は除外した．全ての実験で 4 つの異なるシードで学習を行い，その平均の結果を報告する．

5 実験結果

正しい過去形と過剰汎化形の選好 正しい過去形を含んだ文と過剰汎化形を含んだ分のうち，正しい過去形が選択されたペアの割合を，動詞の過去形の学習曲線として図 2 に示す．横軸は学習のステップを示し，縦軸は 0.5 より大きい場合正しい過去形を好む，0.5 未満の場合過剰汎化形を好むことを示している．実験の結果，どの学習データを用いた場合も，学習後半まで過剰汎化形を好むことが示された．しかし，学習コーパスに多く含まれる正しい過去形 (実在語) よりも過剰汎化形 (非実在語) が学習後半でも好まれるのは非直感的である．このような結果となった原因として，今回の実験ではサブワードレベルでのトークナイザを使用しており，過去形の屈折形がそもそも分割されず，規則変化の過去形は末尾に-ed がつくというような規則をモデルが学習できていない可能性が考えられる．今回使用したトークナイザの挙動を確認したところ，多くの動詞の過去形は分割されず 1 トークンとみなされていることがわかった．この問題の影響を確かめるため，過去形の屈折形が分割されるようにトークナイザを改変し，再実験を行った結果を図 3 に示す．その結果，0-50,000 ステップ付近まではモデルは過剰汎化形を好み，その後正しい過去形をより好むことが示された．学習後半では，Wikipedia や AO-CHILDES+Wikipedia で学習したモデルは 7 割以上の正解率を示した．この正解率の妥当性を確かめるため，RoBERTa [16] で同じ実験を行ったところ，学習済み RoBERTa-base, large モデル両方において正解率は 8 割程度であった．BabyBERTa での実験は RoBERTa と比較してパラメータ数や学習データが大幅に少ないことを考慮すると，この実験結果で得られた正解率は妥当であると考えられる．

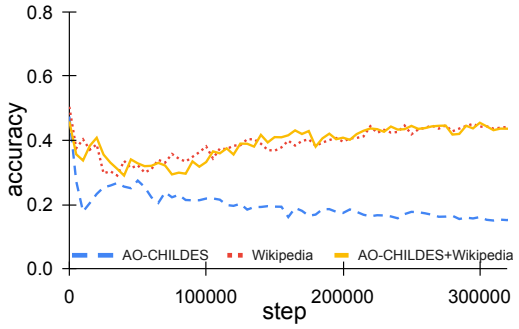

図 2: 通常のトークナイズでのモデルの学習曲線

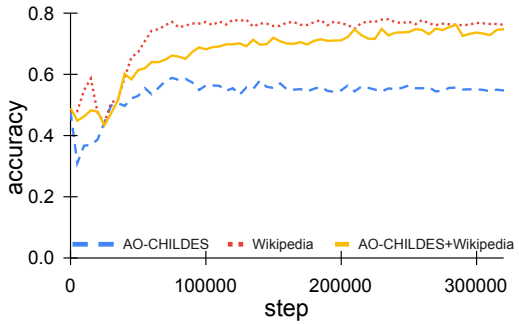

図 3: 屈折形が分割されるようにしたときの学習曲線

base+ed と past+ed の選好 子供の動詞の過剰な一般化には，動詞の原形の末尾に-ed を付加する形 (以下，base+ed) の他に動詞の過去形の末尾に-ed を付加する形 (以下，past+ed) が稀に見られる [17]．Kuczaj ら [17] は，子供の成長に伴って base+ed が増えるにつれて past+ed も増加し，学習後半では past+ed が base+ed より多く産出されることを観察している．また，past+ed が base+ed の産出を上回るのは，子供がほとんど過剰な一般化を起こさなくなったタイミングであるということが示されている．Rumelhart らは，彼らが学習した動詞の原形→過去形変換モデルがこの傾向を捉えていると主張した [7]．しかし，past+ed は，子供が動詞の過去形を原形だと思っているために産出されると考えられている [9]．そのため，Rumelhart らのモデルは入力として動詞の原形を明示的に与えており past+ed を生成するのは適切ではないとして Pinker らが批判を行っている [9]．しかし，本研究はモデルに動詞の原形を明示的に与えないため base+ed と past+ed の選好の比較実験は有用と考えられる．

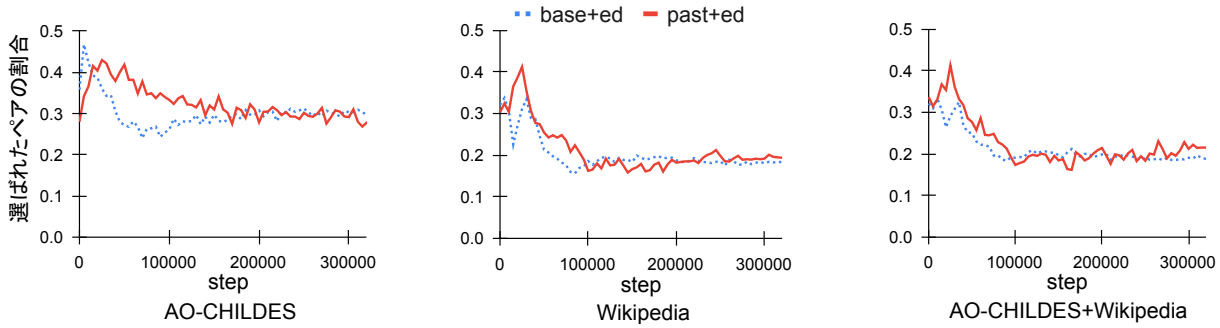

図 4: base+ed と past+ed の選好

したがって，本実験では past+ed と base+ed について CDS で学習したモデルが同様の選好を示すか観察する．図 4 は，AO-CHILDES で学習したモデルが正しい過去形，base+ed，past+ed のうち base+ed を好んだペアの割合と，past+ed を好んだペアの割合を示す．トークナイザは，規則を学習できない問題を回避するため過去形の屈折形が分割されるように改変したものを用いた．その結果，どの学習データで学習した場合でも，学習初期には base+ed よりも past+ed が好まれる傾向が見られた．さらに，AO-CHILDES で学習したモデルは学習後半で base+ed が past+ed を上回り，子供の傾向とは異なる結果を示した．

6 考察

図 2 より，通常のトークナイザを用いた場合では，Wikipedia と AO-CHILDES+Wikipedia で学習したときの学習曲線は 0-100,000 ステップ付近で正解率が低くなり，その後緩やかに上昇している．これは，一見子供の U 字の学習曲線を捉えているように見えるが，全体的に正解率が低く最後まで過剰汎化をより好むという点では子供の特徴とは相反している．この結果は対象とした全ての動詞の平均値あるため，一部の動詞で子供のような U 字の学習曲線が見られる可能性があり，更なる分析が必要である．

図 3 より，Wikipedia で学習したモデルの正解率は，0-15,000 ステップ付近まで 5%ほど上昇し，その後は 25,000 ステップ付近まで減少し，その後は上昇している．この学習曲線は子供の U 字の学習曲線とよく似ている．他の学習データで学習したモデルでも学習前半に正解率が減少しているが，減少する以前の正解率は約 0.5 とランダムに近いため，不規則動詞においても正しい形を答えられるステージ 1 には対応しないと考えられる．

AO-CHILDES のみで学習したモデルの正解率が最も低い原因として，本実験では評価データを作成する際，BLiMP の語彙を用いており，AO-CHILDES に含まれない語彙が評価データに存在することが考えられる．ただし，評価データの語彙のうち，学習データには含まれているが学習初期ではまだ学習されていない語彙が存在するため全体的に正解率が低くなっている可能性がある．よって，学習データに存在する語彙のみを含む評価データを用意し，学習の各ステップで既に学習済みの動詞についてのみ評価を行う必要がある．図 4 より，Wikipedia データを学習に用いたモデルでは学習後半で past+ed が base+ed の選好を上回っているがその差はごく僅かである．

本論文では CDS で学習可能なことが示されているモデルを採用したが，人間の言語獲得に近い学習を目指すためには，順方向に入力を受け取るモデルや音韻ベースのトークナイザを用いた実験を行うのが自然である．今後は対象のモデルを変更した実験を検討している．

7 おわりに

本研究では，動詞の過去形の屈折の一般化に注目し，CDS を学習データに用いて学習したニューラル LM が学習過程でどのような一般化を好むかを分析した．正しい過去形の学習曲線を観察したところ，CDS で学習されたモデルよりも Wikipedia で学習されたモデルの方がより子供に近い U 字の曲線を描くことが示された．モデルが過剰な一般化を選択した際の誤りの種類を分析したところ，どの学習データでも子供とは異なる傾向を示した．実験の結果は，言語モデルを CDS で学習するのみでは，人間の学習過程の誤りの特徴は十分に捉えられない可能性を示唆している．今後は異なるモデルでの実験や，複数の誤りの種類にわたっての評価を行う予定である．

謝辞

本研究は JSPS 科研費 JP21H05054, JST さきがけ JPMJPR21C2, JPMJPR20C4 の助成を受けたものです.

参考文献

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020.

[2] Noam Chomsky. A review of B. F. Skinner's verbal behavior. pp. 26–58, 1959.

[3] Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. BabyBERTa: Learning more grammar with small-scale child-directed language. In **Proceedings of the 25th Conference on Computational Natural Language Learning**, pp. 624–646, Online, November 2021. Association for Computational Linguistics.

[4] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? **arXiv preprint arXiv:2305.07759**, 2023.

[5] Melissa Bowerman. Starting to talk worse: Clues to language acquisition from children's late speech errors. In **U shaped behavioral growth**, pp. 101–145. Academic Press, 1982.

[6] Roger Brown. **A first language: The early stages**. Harvard University Press, 1973.

[7] David E. Rumelhart and James L. McClelland. On learning the past tenses of english verbs. **Parallel Distributed Processing: Explorations in the microstructure of cognition**, p. 216–271, 1986.

[8] Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. Are we there yet? encoder-decoder neural networks as cognitive models of English past tense inflection. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3868–3877, Florence, Italy, July 2019. Association for Computational Linguistics.

[9] Steven Pinker and Alan Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. **Cognition**, Vol. 28, No. 1-2, pp. 73–193, 1988.

[10] Christo Kirov and Ryan Cotterell. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 651–665, 2018.

[11] Charles Yang. **The price of linguistic productivity: How children learn to break the rules of language**. MIT press, 2016.

[12] Tal Linzen and Brian Leonard. Distinct patterns of syntactic agreement errors in recurrent networks and humans. **In Proceedings of the 40th Annual Conference of the Cognitive Science Society**, p. 692–697, 2018.

[13] 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.

[14] Philip A Huebner and Jon A Willits. Using lexical context to discover the noun category: Younger children have it easier. In **Psychology of learning and motivation**, Vol. 75, pp. 279–331. Elsevier, 2021.

[15] Brian MacWhinney. **The CHILDES project: Tools for analyzing talk, Volume II: The database**. Psychology Press, 2014.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[17] Stan A Kuczaj II. The acquisition of regular and irregular past tense forms. **Journal of verbal learning and verbal behavior**, Vol. 16, No. 5, pp. 589–600, 1977.