小規模言語モデルによる統語パラメータの獲得

山田裕真　染谷大河　大関洋平
東京大学
{mercury50620, taiga98-0809, oseki}@g.ecc.u-tokyo.ac.jp

概要

本研究では、言語モデルが統語パラメータを獲得できるか検証することを目的とする。統語パラメータとは、CoLA/BLiMP 等のベンチマークで評価される様な統語現象毎の統語能力では無く、統語現象間の相関・クラスター効果に関する統語能力であり、理論言語学において生得的に賦与されると仮定されている。具体的には、人間の言語獲得を量・質共に近似した小規模言語モデルである BabyBERTa が、冠詞の有無と 2 種類の移動の可否の間の相関・クラスター効果を説明する NP/DP パラメータを獲得できるか検証する。結果として、BabyBERTa が NP/DP パラメータを部分的に獲得できることが示され、統語パラメータが生得的に賦与される必要は無く、経験的に獲得される可能性が示唆された。

1　はじめに

近年、OpenAI の GPT [1] に代表される大規模言語モデルの発展が著しく、特に Transformer [2] に基づく言語モデルは多くの下流タスクで高い性能を実現している。このように、言語を理解している様に見える言語モデルだが、内部がブラックボックスであるため、言語モデルがどれほど主述の一致などの統語能力を獲得しているのか検証する研究が活発になりつつある [3, 4]。また、様々な統語現象を包括的にまとめた CoLA (Corpus of Linguistic Acceptability) [5] や BLiMP （Benchmark of Linguistic Minimal Pairs for English) [6] 等のベンチマークが構築され、言語モデルの統語能力を評価する研究が進展した。

しかしながら、これらの研究は統語現象毎の統語能力を評価しているため、記述的な評価に終始してしまっているという問題がある。例えば、ある言語モデルが統語現象 A と B に正解し、統語現象 C と D に正解しない場合、なぜその様な結果になっているのか説明が得られない。従って、統語現象毎の統語能力だけでは無く、統語現象間の相関・クラスター効果に関する統語能力も評価する必要がある。この統語現象間の相関・クラスター効果は、理論言語学において**統語パラメータ**と呼ばれ、主に ON と OFF の 2 値を取る「スイッチ」の様なものであり、統語現象間の相関・クラスター効果を説明するものとして提案されている。これまで、統語パラメータが生得的に賦与されると仮定した上で、統語パラメータの値が学習できるか検証する研究は存在したが [7, 8]、統語パラメータ自体が言語モデルによって獲得できるか検証する研究は存在しない。

そこで、本研究では、言語モデルが統語パラメータを獲得できるか検証することを目的とする。具体的には、人間の言語獲得を量・質共に近似した小規模言語モデルである BabyBERTa [9] が、冠詞の有無と 2 種類の移動の可否の間の相関・クラスター効果を説明する NP/DP パラメータ [10] を獲得できるか検証する。特に、NP/DP パラメータによって説明される統語現象と、NP/DP パラメータによって説明されない一般的な統語現象（Zorro [9] に含まれる統語現象）を、言語モデルによる統語獲得が収束するタイミングに関して比較する。

2　統語パラメータ

2.1　NP/DP パラメータ

NP/DP パラメータによって説明される統語現象は、(i) 冠詞の有無、(ii) 名詞句からの付加詞移動、(iii) 名詞句からの左枝移動の 3 つである。冠詞は、NP 言語には見られず、DP 言語には見られる。名詞句からの付加詞移動とは、名詞を後ろから修飾する前置詞などの付加詞が前方に移動する現象で、NP 言語の一部で見られるが、DP 言語では認められない現象である。名詞句からの左枝移動は、名詞を修飾する前方から修飾する形容詞などが前方に移動する現象で、これも NP 言語の一部で見られるが、DP 言語では認められない現象である。これら 3 つの統語現象に対して、2,000 の正文と対応する非文

を作成しペアにすることで、それぞれ 2,000 ミニマルペア（4,000 文）を生成した。その際、適切なミニマルペアを自動的に生成するため、各統語現象に対するテンプレートを (1-3) の様に作成した。これらのテンプレートは変数を含んでおり、形態的、統語的、意味的な制約を満たす語彙[1)]を挿入した。また、これらのテンプレートの変数に挿入する語彙は、BabyBERTa の学習に利用するコーパスである CHILDES [11] から抽出した。以上の手順で生成されたミニマルペアの例は表 1 に示す。

(1) 冠詞の有無
  a. *{PN(NOM)_or_NAME} broke window .
  b. {PN(NOM)_or_NAME} broke {DET} window .

(2) 名詞句からの付加詞移動
  a. *from which {PLACE} did {PN(NOM)_or_NAME} {VB_not_with_For} {NN_pl} ?
  b. in which {PLACE} did {PN(NOM)_or_NAME} {VB_not_with_For} {NN_pl} ?

(3) 名詞句からの左枝移動
  a. *{JJ} {PN} {VBD} {NN_pl} ?
  b. {PN} {VBD} {JJ} {NN_pl} ?

表 1 NP/DP パラメータのミニマルペアの例

現象	例文
定冠詞の有無	She broke that window. *She broke window.
名詞句からの付加詞移動	In which library did thomas need keys? *From which library did thomas need keys?
名詞句からの左枝移動	They saw white dogs. *White they saw dogs.

2.2 一般的な統語現象

一般的な統語現象は、BLiMP [6] を BabyBERTa の検証のために改変した Zorro [9] を利用する。Zorro は主述の一致や否定極性項目など言語モデルの統語能力を評価するためのベンチマークであり、BabyBERTa の学習に利用するコーパスである CHILDES [11] にある語彙のみを用いている。これらの統語現象は NP/DP パラメータによって説明されない一般的な統語現象であり、NP/DP パラメータによって説明される統語現象と、3.2.2 節で定義する統語獲得の収束点に関して比較する。

1) PN(NOM) は主格代名詞、NAME は人物を表す固有名詞、DET は指示代名詞を含む冠詞、PLACE は場所を表す名詞、VB_not_with_For は '*for*' と共起しない動詞（'*for*' 句が動詞補部ではなく付加詞であることを保証する）、NN_pl は複数名詞、JJ は形容詞、VBD は過去動詞を意味する。

3 実験

3.1 言語モデルと学習データ

本研究では、RoBERTa [12] のパラメータ数を縮小した BabyBERTa [9] を採用した。ハイパーパラメータは先行研究 [9] を踏襲し、unmasking の割合は 0、学習は 2,600,000 ステップまで進めた。RoBERTa と BabyBERTa の比較については付録 A に示す。

学習データは、CHILDES [11] から米国英語話者の子供に向けられた発話を抽出したコーパスである AO-CHILDES (Age-Ordered CHILDES) [13] を採用した。このコーパスは、子供が 6 歳までに経験する学習データを量・質共に近似している。

3.2 評価尺度

3.2.1 正答率

言語モデルの統語能力を評価するため、文の長さや単語の頻度に影響を受ける単純な尤度では無く、SLOR [14] を正答率の計算に使用した。SLOR は、文の長さや単語の頻度による影響を統制することが可能なため、統語能力の評価尺度に適している。SLOR は以下の様に定義される。

$$\mathrm{SLOR} = \frac{\log p_m(\zeta) - \log p_u(\zeta)}{|\zeta|}$$

ここで、$p_m(\zeta)$ は BabyBERTa によって推定される文の尤度、$p_u(\zeta) = \prod_{w \in \zeta} p_u(w)$ はユニグラム言語モデルによって推定される文の尤度、$|\zeta|$ は文の長さを表す。そして、正文の SLOR が非文の SLOR より高い場合は正解、そうでない場合は不正解として、各統語現象の正答率を計算する。

3.2.2 収束点

また、言語モデルによる統語獲得の収束点を使用した。言語モデルによる意味獲得を検証した先行研究 [15] を踏襲し、収束点は「そのステップにおいて正答率が任意の閾値を超え、且つそれ以降の正答率の平均が閾値を超えるようなステップ数」として定義する。本研究では閾値を 65%に設定した。[2)]

そして、NP/DP パラメータによって説明される統語現象 A の収束点が、NP/DP パラメータによって説明されない一般的な統語現象 B の収束点と比べて、

2) 閾値は最終ステップの正答率の平均を参考に設定した。

NP/DP パラメータによって説明される別の統語現象 C の収束点に近いということを示すため、ウェルチの t 検定を利用した。この統計分析のため、言語モデルは 10 個のシードで学習・評価した。

4 結果

NP/DP パラメータによって説明される統語現象および一般的な統語現象 (Zorro) の学習曲線を、それぞれ図 1 および図 2 に示す。横の破線は閾値、縦の破線は収束点をそれぞれ表す。まず、正答率に関しては、平均が 70%を超え、NP/DP パラメータによって説明される統語現象 2 つおよび一般的な統語現象 14 つ、計 16 現象が閾値である 65%を超えた。特に、NP/DP パラメータによって説明される統語現象に関しては、名詞句からの左枝移動は正答率が約 80%となり、先行研究 [16] と一貫する結果が得られた。また、冠詞の有無は正答率が約 100%となった一方で、名詞句からの付加詞移動はほとんど学習が進まなかった。

また、収束点に関しては、正答率と同様、NP/DP パラメータによって説明される統語現象 2 つおよび一般的な統語現象 14 つ、計 16 現象が 3.2.2 節の定義で収束した。特に、NP/DP パラメータによって説明される現象に関しては、名詞句からの付加詞移動は収束しなかった一方で、冠詞の有無および名詞句からの左枝移動は収束し、冠詞の有無と名詞句からの左枝移動の収束点の差は統計的に有意では無かった （$t = 1.80, p = 0.10$）。一方、一般的な統語現象に関しては、冠詞の有無の収束点との差および名詞句からの左枝移動の収束点との差の両方が統計的に有意（$p < 0.05$）であったものは、収束した 14 つの統語現象のうち 10 現象であった。

従って、NP/DP パラメータによって説明される統語現象のうち、収束した 2 つの統語現象は収束点が比較的「近い」一方で、これら 2 つの統語現象と一般的な統語現象は収束点が比較的「遠い」と結論付けることが出来る。統計分析の詳細は付録 B に示す。

5 考察

以上の結果から、言語モデルが NP/DP パラメータを部分的に獲得できることが示された。一方、この示唆を必ずしも支持しない結果や、言語獲得のモデルとして子供の言語獲得との差分がみられた。

まず、冠詞の有無および名詞句からの左枝移動に関する正答率が、比較的「近い」収束点を持っているという結果は、モデルが学習において NP/DP パラメータで説明される現象間のクラスター効果を示したといえる。この結果は、その背後にある NP/DP パラメータ自体とその値を獲得したということを示唆する。さらにこの示唆からは、こうした統語パラメータが生得的でないという示唆を与える。

一方、名詞句からの付加詞移動が「収束」しなかった原因としては、NP/DP パラメータが獲得されなかった可能性だけでなく、モデルにおいて NP/DP パラメータは獲得されたが動詞とコロケーションを作る／作らない前置詞の知識が欠落していた可能性も考えられる。

また、名詞句からの左枝移動についても学習が安定しているわけではない。これも、NP/DP パラメータではない要素（これらの 2 現象についていえば、理論言語学において、反局所性やフェーズ不可侵条件 [10, 17] といった性質が知られている）が獲得されなかったことで、名詞句からの左枝抽出が適切に学習されなかった可能性がある。

冠詞の有無および名詞句からの左枝移動に関する正答率が比較的早い段階で「収束」している点は、ヒトの言語獲得の傾向と一致しない。ヒトの言語獲得においては、冠詞の獲得は 4 歳前後であると言われており、これは言語獲得の時期としては比較的遅めである [18, 19]。ただし、この結果自体がモデルがヒトの言語獲得の順序に従っていないことは NP/DP パラメータの獲得の反証にはならない。

名詞句からの左枝移動などのいくつかの現象において、正答率が途中で一度大きく落ちている点は、明確な説明が得られなかった。これらの現象については、理論言語学の観点から新たな共通性の提案が待たれる。

これらの考察から、BabyBERTa が学習において NP/DP パラメータによるクラスター効果を見せたことから、NP/DP パラメータを部分的に学習できることが示された。したがってヒトにおいては統語パラメータが生得的に賦与される必要は無く、経験的に獲得される可能性が示唆された。一方、明確な結論を導くためには、更なる検証が必要である。

6 おわりに

本研究では、言語モデルが統語パラメータを獲得できるか検証することを目的とした。具体的には、人間の言語獲得を量・質共に近似した小規模言語モ

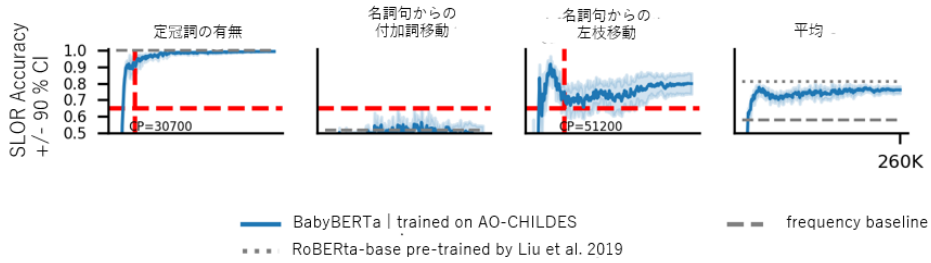

図 1　NP/DP パラメータによって説明される統語現象の学習曲線

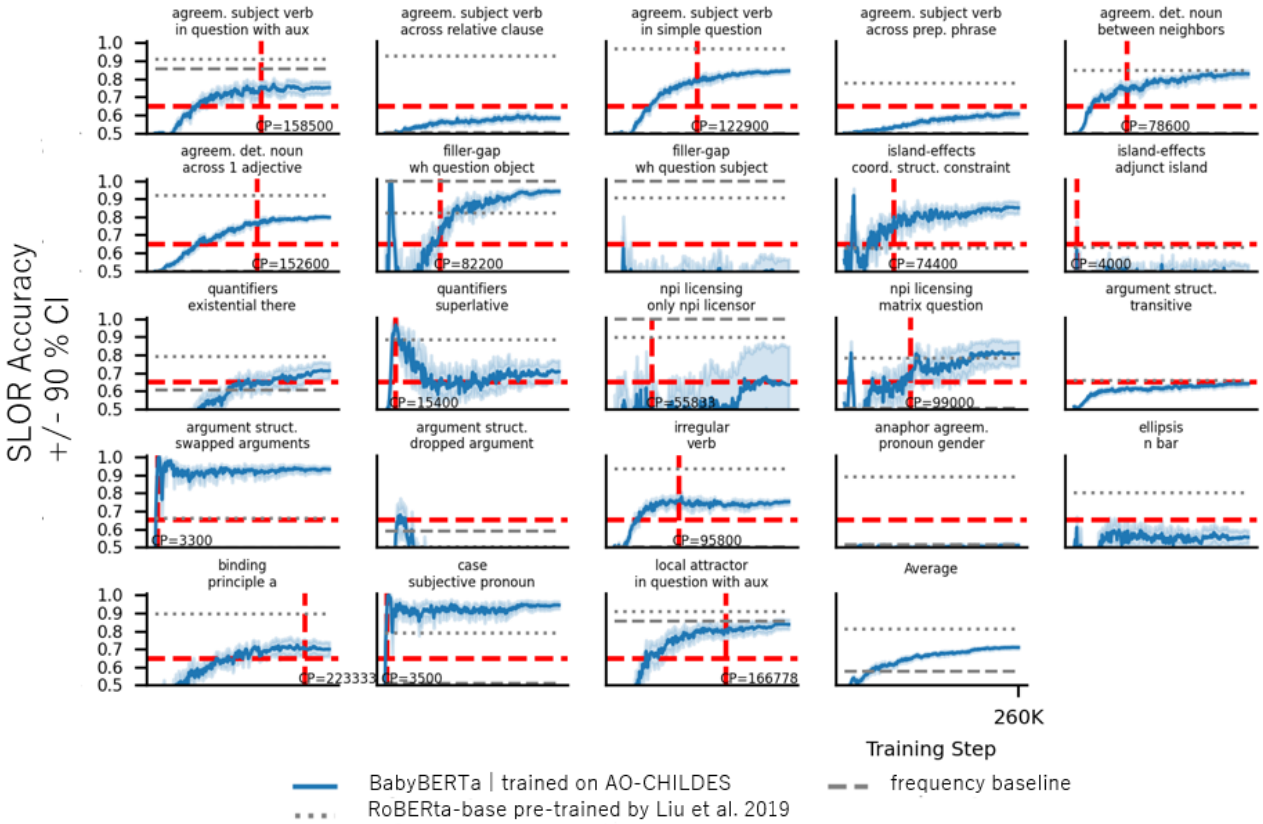

図 2　一般的な統語現象の学習曲線

デルである BabyBERTa [9] が、冠詞の有無と 2 種類の移動の可否の間の相関・クラスター効果を説明する NP/DP パラメータ [10] を獲得できるか検証した。特に、NP/DP パラメータによって説明される統語現象と、NP/DP パラメータによって説明されない一般的な統語現象（Zorro [9] に含まれる統語現象）を、言語モデルによる統語獲得が収束するタイミングに関して比較した。結果として、BabyBERTa が NP/DP パラメータを部分的に獲得できることが示され、統語パラメータが生得的に賦与される必要は無く、経験的に獲得される可能性が示唆された。

本研究では、AO-CHILDES のような特定のコーパスが利用可能であった、DP 言語である英語のみを対象として実験を行ったが、日本語のような NP 言語についても訓練データとして利用することで、本研究のより深い検証が可能だと考えられる。また、モデルについても、本研究で利用した BabyBERTa のような小規模言語モデルだけでなく、近年特に人間並みの性能を発揮している GPT [1] をはじめとする大規模言語モデルによる検証を行うことで、言語モデルの言語理解に関して更なる洞察が得られると考えられる。こうした方向性については、今後の研究の課題としたい。

謝辞

本研究は、JST さきがけ JPMJPR21C2 の支援を受けたものです。

参考文献

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. June 2017. arXiv: 1706.03762.

[3] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 521–535, 2016.

[4] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1192–1202, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[5] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 625–641, 2019.

[6] 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, 07 2020.

[7] Charles D Yang. **Knowledge and learning in natural language**. Oxford University Press, USA, 2002.

[8] William Gregory Sakas and Janet Dean Fodor. Disambiguating syntactic triggers. **Language Acquisition**, Vol. 19, No. 2, pp. 83–143, 2012.

[9] Philip A Huebner, Elior Sulem, Cynthia Fisher, and Dan Roth. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. Technical report, 2021.

[10] Željko Bošković. What will you have, dp or np? In **Proceedings-Nels**, Vol. 37, p. 101, 2008.

[11] Brian MacWhinney. **The CHILDES Project: Tools for analyzing talk. transcription format and programs**, Vol. 1. Psychology Press, 2000.

[12] 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, July 2019. arXiv:1907.11692 [cs].

[13] Philip A. Huebner and Jon A. Willits. **Using lexical context to discover the noun category: Younger children have it easier**, pp. 279–331. Psychology of Learning and Motivation - Advances in Research and Theory. Academic Press Inc., United States, January 2021. Publisher Copyright: © 2021 Elsevier Inc.

[14] Jey Han Lau, Alexander Clark, and Shalom Lappin. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. **Cognitive Science**, Vol. 41, No. 5, pp. 1202–1241, 2017.

[15] Shane Steinert-Threlkeld and Jakub Szymanik. Learnability and semantic universals. **Semantics and Pragmatics**, Vol. 12, pp. 4:1–39, November 2019.

[16] Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. Using Computational Models to Test Syntactic Learnability. **Linguistic Inquiry**, pp. 1–44, 04 2023.

[17] Željko Bošković. On NPs and clauses. **Discourse and grammar: From sentence types to lexical categories**, Vol. 179, p. 245, 2012.

[18] Don Nilsen and Roger Brown. A first language: The early stages. **The Modern Language Journal**, Vol. 58, p. 268, 09 1974.

[19] David A. Warden. The influence of context on children's use of identifying expressions and references. **British Journal of Psychology**, Vol. 67, No. 1, pp. 101–112, 1976.

A　RoBERTa [12] と BabyBERTa の比較 [9]

	RoBERTa-base	BabyBERTa
parameters	125M	5M
data size	160GB	0.02GB
words in data	30B	5M
batch size	8K	16
max sequence	512	128
epochs	>40	10
max step	500	260
hardware	1024x V100	1x GTX1080
training time	24hours	2hours
accuracy	81.0	80.5

B　ウェルチの t 検定による統計分析の詳細

表 2　冠詞の有無と、その他の現象の収束点の統計的比較。p 値が 0.05 以下になっているものは、その現象と冠詞の有無に関しての収束点の差分が有意であることを示す。

現象	t 値	p 値
agreement subject verb in question with aux	-8.1667	7.49e-05 **
agreement subject verb in simple question	-16.2285	2.02e-08 **
agreement determiner noun between neighbors	-9.7809	1.76e-06 **
agreement determiner noun across 1 adjective	-26.2794	9.36e-1 1**
filler gap wh question object	-1.9950	0.0770
island effects coordinate structure constraint	-3.1347	0.0118 *
island effects adjunct island	nan[1]	nan[1]
quantifiers superlative	9.5192	2.08e-08 **
npi licensing only npi licensor	-0.8718	0.4230
npi licensing matrix question	-2.0990	0.0805
argument structure swapped arguments	19.6946	5.42e-12 **
irregular verb	-6.9337	5.80e-05 **
binding principle a	-32.0163	0.0006 **
case subjective pronoun	20.1129	1.37e-11 **
local attractor in question with aux	-9.4240	1.20e-05 **
left branch extraction out of a nominal phrase	-1.8008	**0.1045**

表 3　名詞句からの左枝移動と、その他の現象の収束点の統計的比較。p 値が 0.05 以下になっているものは、その現象と名詞句からの左枝移動に関しての収束点の差分が有意であることを示す。

現象	t 値	p 値
agreement subject verb in question with aux	-5.5660	8.13e-05 **
agreement subject verb in simple question	-5.6857	7.26e-05 **
agreement determiner noun between neighbors	-2.2318	0.0453 *
agreement determiner noun across 1 adjective	-8.3275	2.86e-06 **
filler gap wh question object	-1.1007	0.2920
island effects coordinate structure constraint	-1.2947	0.2124
island effects adjunct island	nan[1]	nan[1]
quantifiers superlative	3.1479	0.0115 *
npi licensing only npi licensor	-0.1497	0.8855
npi licensing matrix question	-1.3882	0.2051
argument structure swapped arguments	4.2225	0.0022 **
irregular verb	-3.0424	0.0072 **
binding principle a	-13.4845	3.74e-08 **
case subjective pronoun	4.2066	0.0023 **
local attractor in question with aux	-6.3123	1.15e-05 **
definite article definite article	1.8008	**0.1045**

1）　1 つのシードのみが収束したため、ウェルチの t 検定の計算に必要な自由度が計算できない。