どれほどの統語的教示が必要十分なのか？

能地 宏
LeapMind 株式会社
noji@leapmind.io

大関 洋平
東京大学
oseki@g.ecc.u-tokyo.ac.jp

概要

本研究では、言語モデルをより人間らしくするために、どれほどの統語的教示が必要十分なのか検証する。具体的には、新しい手法である**統語的アブレーション**（syntactic ablation）を提案し、統語的言語モデルである再帰的ニューラルネットワーク文法（Recurrent Neural Network Grammar; RNNG）における NP、VP、PP、SBAR およびそれらを組み合わせた非終端記号を段階的に削除することで、完全な統語的教師あり（純正 RNNG と等価）から完全な統語的教師なし（単方向 LSTM と等価）まで 17 つの RNNG を構築した上で、SyntaxGym における 6 つの統語的サーキットで評価する。結果として、部分的な統語的教師あり RNNG が SyntaxGym で世界最高性能を達成し、言語処理に対する必要十分（good enough）アプローチが人間らしいことを示唆する。

1 はじめに

言語モデルの統語的評価に関する先行研究では [1, 2]、Long Short-Term Memory (LSTM)[3] の様な再帰的ニューラルネットワーク（Recurrent Neural Network; RNN）が、明示的な統語的教示が無いにも関わらず、非明示的に自然言語の統語構造（例えば、英語における主語と動詞の一致など）を学習できることが示唆されている [4]。更に、これらの RNN は明示的な統語的教示の恩恵も受けることが示されており、RNN と明示的な統語的教示を統合した再帰的ニューラルネットワーク文法（Recurrent Neural Network Grammar; RNNG）[5] は、認知的妥当性の観点から心理言語学および認知モデリングの文脈で注目を集めており、統語的評価 [6, 7] だけで無く心理言語学的評価 [8, 9, 10] においても RNN より性能が高いことが知られている。

しかしながら、明示的な統語的教示の有無という二分法をめぐる集中的な議論に関わらず、どれほどの統語的教示が必要十分なのかという問いはまだ未開拓である。特に、完全な統語的教示が最適では無いと考える潜在的な理由が少なくとも 2 つ存在する。まず、理論的には、完全な統語的教示が RNN によって潜在的に学習される語彙的なヒューリスティックスを上書きしてしまい、終端記号（単語）に関する情報が再帰的な統語演算により消失してしまうという可能性である [11]。また、経験的には、完全な統語的教示が長距離依存関係の正答率に悪影響を及ぼしているという可能性であり、特に擬似分裂文では正文（例えば、*What he **did** was prepare the meal.*）と非文（例えば、**What he **ate** was prepare the meal.*）が全く同じ統語構造を持つため、語彙的なヒューリスティックスのみで区別する必要がある [12]。従って、以上の 2 つの理由の帰結として、統語構造と語彙的ヒューリスティックスの絶妙なバランスを取るために、最適な統語的教示は完全な統語的教師ありと完全な統語的教師なしの間どこかに存在するはずであるという仮説を導くことが出来る。直感的には、文法ばかり言語モデルに教え過ぎると、言語モデルは文法にしか着目しない様になり、結果として語彙を忘れてしまうということである。

そこで、本研究では、言語モデルをより人間らしくするために、どれほどの統語的教示が必要十分なのか検証する。具体的には、新しい手法である**統語的アブレーション**（syntactic ablation）を提案し、統語的言語モデルである再帰的ニューラルネットワーク文法（Recurrent Neural Network Grammar; RNNG）[5] における NP、VP、PP、SBAR およびそれらを組み合わせた非終端記号を段階的に削除することで、完全な統語的教師あり（純正 RNNG と等価）から完全な統語的教師なし（単方向 LSTM と等価）まで 17 つの RNNG を構築した上で、SyntaxGym における 6 つの統語的サーキットで評価する。結果として、部分的な統語的教師あり RNNG が SyntaxGym で世界最高性能を達成し、心理言語学で提案されている言語処理に対する必要十分（good enough）アプローチ [13, 14] が人間らしいことを示唆する。

図 1 統語的アブレーション（syntactic ablation）の概要。RNNG における NP、VP、PP、SBAR およびそれらを組み合わせた非終端記号を段階的に削除することで、完全な統語的教師あり（純正 RNNG と等価）から完全な統語的教師なし（単方向 LSTM と等価）まで、17 つの RNNG を構築する。

2 提案手法

2.1 再帰的ニューラルネットワーク文法

再帰的ニューラルネットワーク文法（Recurrent Neural Network Grammar; RNNG）[5] は、文（単語列）と構造の深層生成モデルである。RNNG は、スタック LSTM[15] に基づき以下の 3 つのアクションに関する確立分布を計算する：

- NT：非終端記号を開く。
- GEN: 終端記号（単語）を生成する。
- REDUCE: 非終端記号を閉じる。

特に、REDUCE アクションにおいて、RNNG は双方向 LSTM に基づき左 → 右および左 ← 右に複数の終端記号（単語）を句ベクトルに符号化する。また、推論において、RNNG は単語同期型ビームサーチ [16, 8] を並列化した実装 [12] を用いる。[1)]

2.2 統語的アブレーション

提案手法である統語的アブレーション（structural ablation）の概要を図 1 にまとめる。RNNG における NP、VP、PP、SBAR およびそれらを組み合わせた非終端記号を段階的に削除することで、完全な統語的教師あり（純正 RNNG と等価）から完全な統語的教師なし（単方向 LSTM と等価）まで、以下 17 つの RNNG を構築する：[2)]

- Root：ゼロ文法（単方向 LSTM と等価）
- N：NP 非終端記号のみ
- V：VP 非終端記号のみ
- P：PP 非終端記号のみ
- Sb：SBAR 非終端記号のみ
- NV：NP と VP 非終端記号
- NP：NP と PP 非終端記号
- NSb：NP と SBAR 非終端記号
- VP：VP と PP 非終端記号
- VSb：VP と SBAR 非終端記号
- PSb：PP と SBAR 非終端記号
- NVP：NP、VP、PP 非終端記号
- NVSb：NP、VP、SBAR 非終端記号
- NPSb：NP、PP、SBAR 非終端記号
- VPSb：VP、PP、SBAR 非終端記号
- NVPSb：NP、VP、PP、SBAR 非終端記号
- Full：フル文法（純正 RNNG と等価）

RNNG の訓練データは、先行研究 [17] に従い、まず BLLIP コーパス [18]（XL、約 42M トークン）の各文を Berkeley Neural Parser[19] で再構文解析し、以上 NP、VP、PP、SBAR およびそれらを組み合わせた非終端記号を削除することで構築した。また、各 RNNG は 3 つの異なるランダムシードで訓練した。

2.3 SyntaxGym

統語的アブレーションにより構築された 17 つの RNNG は、SyntaxGym プラットフォーム [20, 17] における以下の 6 つの統語的サーキットで評価した：Agreement、Garden-Path Effects、Licensing、Center Embedding、Gross-Syntactic State、Long-Distance Dependencies。[3)] また、SyntaxGym プラットフォームのリーダーボードで採用されている「部分一致」評価尺度は正答率を過大評価してしまうため、保守的な「完全一致」評価尺度を採用した [17]。

1） https://github.com/aistairc/rnng-pytorch
2） 完全な統語的教師なし RNNG と単方向 LSTM は実質的に等価である。唯一の差異は、文頭で Root 非終端記号を開く NT アクションと文末で Root 非終端記号を閉じる REDUCE アクションのみであるが、後者は終端記号（単語）を生成する GEN アクションに影響が無い。
3） https://syntaxgym.org/

3 結果

3.1 全体の正答率

統語的アブレーション実験における全体の正答率を図 2 にまとめる。SyntaxGym の正答率（Y 軸）を、様々な程度で統語的アブレーションした 17 つの RNNG（X 軸）に対して、先行研究 [17] における GPT-2 XL と RNNG の正答率を併せてプロットする。エラーバーは標準偏差を示す。`Root`（左端）と `Full`（GPT-2 XL と RNNG (H20) を除く右端）は、それぞれフル文法とゼロ文法を表す。`N`、`V`、`P`、`Sb` は、それぞれ NP、VP、PP、SBAR 非終端記号が保持された文法を表す。従って、`NP` は NP と PP 非終端記号が保持された文法を表すため、NP 非終端記号のみが保持された文法と混同されない様に注意されたい。

ここで重要な結果を 3 つ観察することが出来る。まず、単方向 LSTM と等価である完全な統語的教師なし RNNG は、統語的教師あり RNNG と比べて優位に統語性能が低いことから、高い言語能力を実現するためには統語的教師が重要な役割を果たすことを示唆している。次に、完全な統語的教師あり RNNG は、部分的な統語的教師あり RNNG や GPT-2 XL[17] と比べて優位に統語性能が低いことから、完全な統語的教師が必要であるとは限らないことを意味している。最後に、部分的な統語的教師あり RNNG である `NPSb` 文法が SyntaxGym で世界最高性能を達成し（84.6）、数値的に GPT-2 XL[17] を上回っていることから（84.2）、部分的な統語的教師が十分であることを含意している。

3.2 統語的サーキット毎の正答率

統語的アブレーション実験における統語的サーキット毎の正答率を図 3 にまとめる。SyntaxGym における 6 つの統語的サーキット毎の正答率（Y 軸）を、様々な程度で統語的アブレーションした 17 つの RNNG（X 軸）に対してプロットする。

興味深いことに、NP 非終端記号は 6 つ中 4 つの統語的サーキット（`Agreement`、`Licensing`、`Center Embedding`、`Gross-Syntactic State`）において統語性能を向上させているのに対して、PP 非終端記号はほぼ全ての統語的サーキット（`Long-Distance Dependencies` を除く）において統語性能を向上させていない。加えて、部分的な統語的教師あり RNNG である `NPSb` 文法は、`Long-Distance Dependencies` サーキットにおいて、完全な統語的教師あり RNNG より統語性能が高いことから、完全な統語的教示が長距離依存関係（特に擬似分裂文）の正答率に悪影響を及ぼしているという可能性を支持している。

4 考察

まとめると、統語的アブレーション実験を実施し、RNNG における NP、VP、PP、SBAR およびそれらを組み合わせた非終端記号を段階的に削除することで、完全な統語的教師あり（純正 RNNG と等価）から完全な統語的教師なし（単方向 LSTM と等価）まで 17 つの RNNG を構築した上で、SyntaxGym における 6 つの統語的サーキットで評価してきた。本節では、統語的アブレーション実験の結果を心理言語学および認知モデリングの文脈に位置付ける。

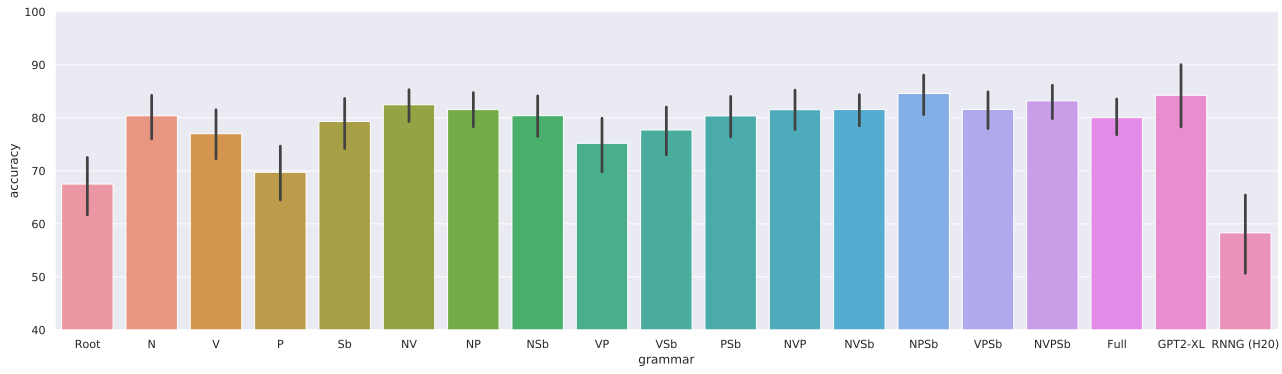

図 2 統語的アブレーション実験における全体の正答率。SyntaxGym における 6 つの統語的サーキットおよび 3 つのランダムシードを平均した正答率（Y 軸）を、様々な程度で統語的アブレーションした 17 つの RNNG（X 軸）に対して、先行研究 [17] における GPT-2 XL と RNNG の正答率を併せてプロットする。エラーバーは標準偏差を示す。`Root`（左端）と `Full`（GPT-2 XL と RNNG (H20) を除く右端）は、それぞれフル文法とゼロ文法を表す。`N`、`V`、`P`、`Sb` は、それぞれ NP、VP、PP、SBAR 非終端記号が保持された文法を表す。従って、`NP` は NP と PP 非終端記号が保持された文法を表すため、NP 非終端記号のみが保持された文法と混同されない様に注意されたい。

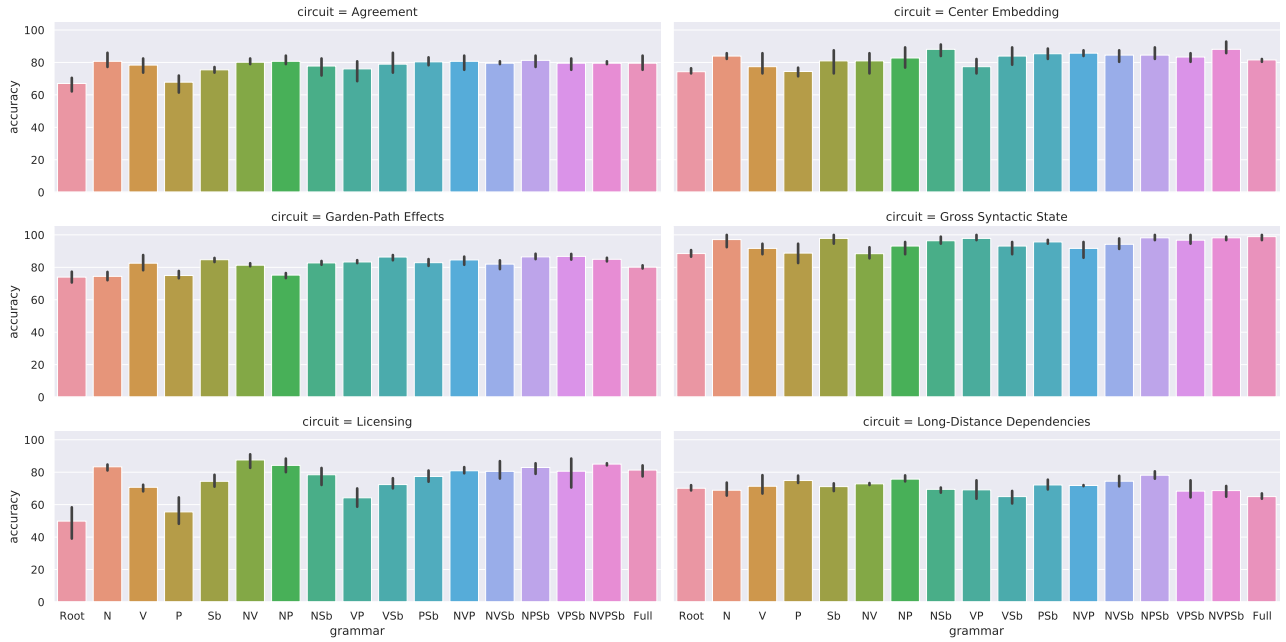

図 3　統語的アブレーション実験における統語的サーキット毎の正答率。SyntaxGym における 6 つの統語的サーキット毎の正答率（Y 軸）を、様々な程度で統語的アブレーションした 17 つの RNNG（X 軸）に対してプロットする。

4.1　必要十分（good enough）アプローチ

小節 3.1 で報告した全体の正答率は、部分的な統語的教師あり RNNG が（特に NPSb 文法）、完全な統語的教師なし RNNG および完全な統語的教師あり RNNG だけで無く、先行研究 [17] における GPT-2 XL より統語性能が高いことを示していた。これらの結果は、心理言語学で提案されている言語処理に対する必要十分（good enough）アプローチ [13, 14] と親和性が高く、人間の言語処理は常に深い構文解析を実施するとは限らず、浅い構文解析や語彙的ヒューリスティックスを利用すると主張されている。従って、部分的な統語的教師あり RNNG は、浅い統語構造を生成し、語彙的ヒューリスティックスを潜在的に学習する余地を残しているという意味で、言語処理に対する必要十分アプローチを実装した認知モデルであると解釈することが出来る。

4.2　長距離依存関係

小節 3.2 で報告した統語的サーキット毎の正答率は、NP 非終端記号が 6 つ中 4 つの統語的サーキットにおいて統語性能を向上させることを示していた。重要なことに、詳細に検討した結果、これら 4 つのサーキットは以下の通り関係節に修飾された重い主語 NP が現れるという点で自然類を成すことが判明した（**依存関係**は太字）：

- `Agreement`: [$_{NP}$ The **author** that the senators hurt] **is** good.
- `Licensing`: [$_{NP}$ **No** author that the senators liked] has had **any** success.
- `Center Embedding`: [$_{NP}$ The **painting** that the artist painted] **deteriorated**.
- `Gross-Syntactic State`: **As** [$_{NP}$ the doctor who the administrator had recently hired] studied [$_{NP}$ the book that colleagues had written on cancer therapy], **the nurse walked into the room**.

ここで重要なポイントは、REDUCE アクションが複数の終端記号（単語）を句ベクトルに符号化し、長距離依存関係を実質的に局所的にしているという点である（例えば、`Agreement` における **author** と **is**）。

5　おわりに

本研究では、統語的アブレーション実験を実施し、RNNG における NP、VP、PP、SBAR およびそれらを組み合わせた非終端記号を段階的に削除することで、完全な統語的教師あり（純正 RNNG と等価）から完全な統語的教師なし（単方向 LSTM と等価）まで 17 つの RNNG を構築した上で、SyntaxGym における 6 つの統語的サーキットで評価した。結果として、部分的な統語的教師あり RNNG が SyntaxGym で世界最高性能を達成し、言語処理に対する必要十分アプローチが人間らしいことを示唆した。

謝辞

本研究は、JST さきがけ JPMJPR21C2 および JSPS 科研費 JP19H04990（新学術領域研究）、JP21H05061（学術変革領域研究 (B)）、JP20K19877（若手研究）の助成を受けたものです。

参考文献

[1] 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.

[2] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1192–1202, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. **Journal of Neural Computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

[4] John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 199–209, San Diego, California, June 2016. Association for Computational Linguistics.

[6] Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1426–1436, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[7] Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. Structural supervision improves learning of non-local grammatical dependencies. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 3302–3312, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2727–2736, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[9] Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. On the predictive power of neural language models for human real-time comprehension behavior. **ArXiv**, Vol. abs/2006.01912, , 2020.

[10] Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. Modeling human sentence processing with left-corner recurrent neural network grammars. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2964–2973, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith. What do recurrent neural network grammars learn about syntax? In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pp. 1249–1258, Valencia, Spain, April 2017. Association for Computational Linguistics.

[12] Hiroshi Noji and Yohei Oseki. Effective batching for recurrent neural network grammars. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 4340–4352, Online, August 2021. Association for Computational Linguistics.

[13] Fernanda Ferreira, Karl G.D. Bailey, and Vittoria Ferraro. Good-enough representations in language comprehension. **Current Directions in Psychological Science**, Vol. 11, No. 1, pp. 11–15, 2002.

[14] Fernanda Ferreira and Nikole D. Patson. The 'good enough' approach to language comprehension. **Language and Linguistics Compass**, Vol. 1, No. 1-2, pp. 71–83, 2007.

[15] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. Transition-based dependency parsing with stack long short-term memory. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 334–343, Beijing, China, July 2015. Association for Computational Linguistics.

[16] Mitchell Stern, Daniel Fried, and Dan Klein. Effective inference for generative neural parsing. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 1695–1700, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[17] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1725–1744, Online, July 2020. Association for Computational Linguistics.

[18] Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. BLLIP 1987- 89 WSJ Corpus Release 1 LDC2000T43, 2000. Linguistic Data Consortium.

[19] Nikita Kitaev, Steven Cao, and Dan Klein. Multilingual constituency parsing with self-attention and pre-training. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3499–3505, Florence, Italy, July 2019. Association for Computational Linguistics.

[20] Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. SyntaxGym: An online platform for targeted evaluation of language models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 70–76, Online, July 2020. Association for Computational Linguistics.