ACL 2021

Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models

Matthew Finlayson*
Harvard University
Cambridge, MA
mattbnfin@gmail.com

Aaron Mueller*
Johns Hopkins University
Baltimore, MD
amueller@jhu.edu

Sebastian Gehrmann
Google Research
New York, NY
gehrmann@google.com

Stuart Shieber
Harvard University
Cambridge, MA
shieber@seas.harvard.edu

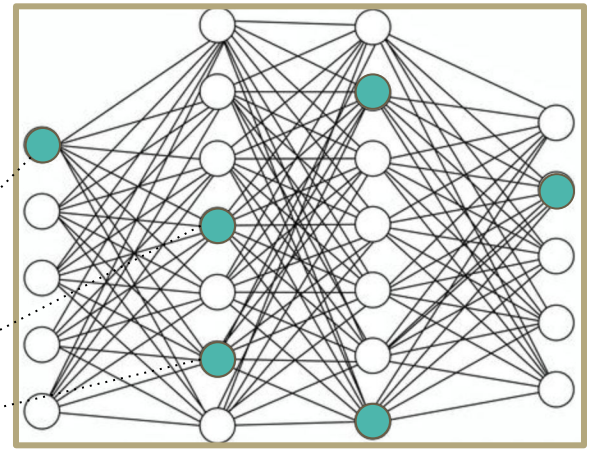
Tal Linzen†
New York University
New York, NY
linzen@nyu.edu

Yonatan Belinkov‡
Technion – IIT
Haifa, Israel
belinkov@technion.ac.il

最先端NLP 2021
神藤 駿介（東大 宮尾研 M2）

概要

要約：Pretrained LM が Syntactic Test を解くメカニズムを 因果媒介分析 で調査



Syntactic Test

活用の正しさが分かる
＝ 主語を把握している
＝ 統語構造を把握している

因果媒介分析

モデルのどの辺りがタスクを解くのに寄与しているか分析

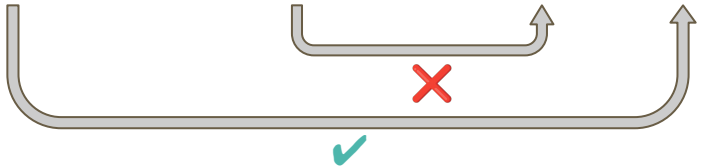
- タスクの種類に依らず同じ部分を使っているか？
- モデルやサイズによって違いはあるか？

研究背景 ~ Syntactic Test

言語モデルの文法把握能力を測定したい　... 主語と動詞の一致 を活用

- 文法的に正しい文・間違えている文のペア
- 主語を正しく捉えないと解けない

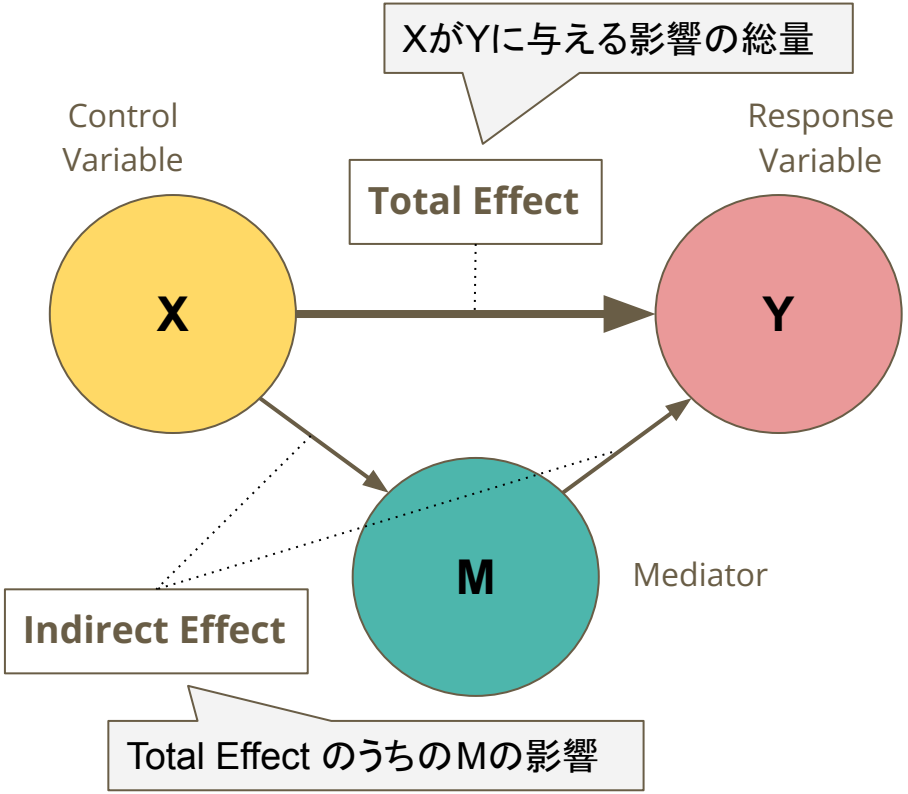
The farmer that the parents love ***confuse/confuses** .

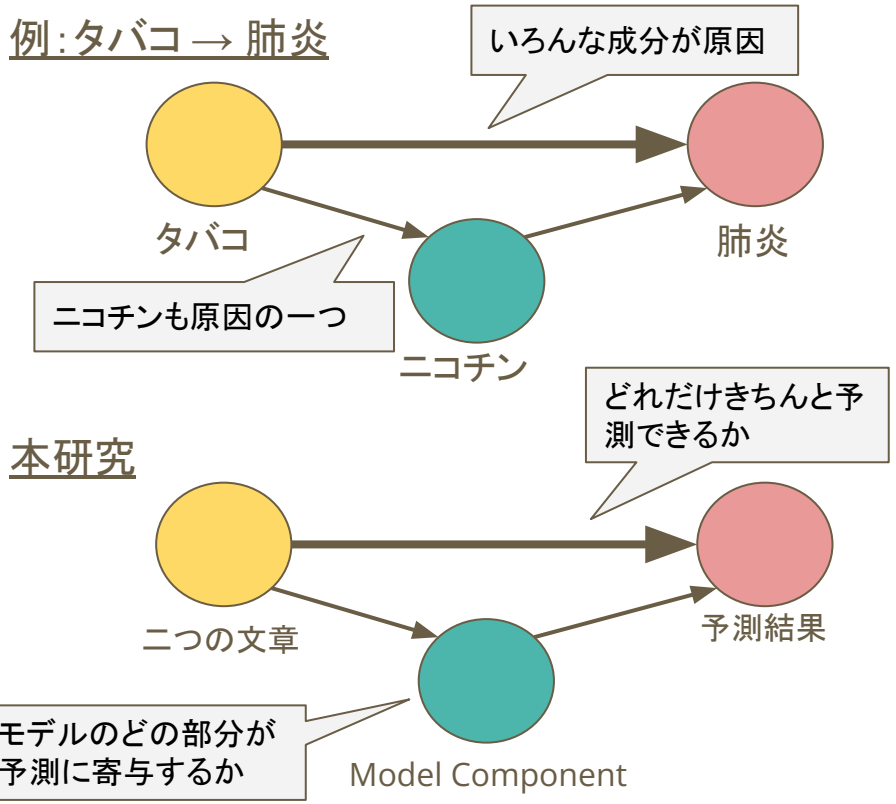


- Pretrained LM で高精度 [Hu+ 2020]
- 本研究：どんなメカニズムで解いているかを調査

研究背景 ~ Causal Mediation Analysis (因果媒介分析)

★二つの変数の因果関係を、ある変数が媒介するかを分析





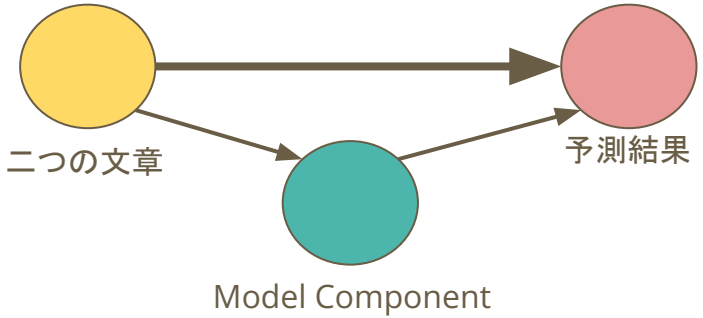
実験の前提 ~ 各変数の設定

Control Variable

→ 主語の単複を変えた2つの文章

The farmer(sg) that the parents love ____

The farmers(pl) that the parents love ____



Response Variable

→ 2つの候補となる単語の予測確率の比

$$y(u_{sg}, v) = \frac{p_\theta(v_{pl} \mid u_{sg})}{p_\theta(v_{sg} \mid u_{sg})}$$

← 動詞を複数形(pl)と予測 (confuse)

← 動詞を単数形(sg)と予測 (confuses)

主語が単数形の文章 (farmer)

※ y < 1 なら正しく予測できている

実験の前提 ~ モデルとデータセット

モデル: Transformer-based を調査

- GPT-2 ... 今回はこれを主に調査
  - 5つのサイズ (Distil / Small / Medium / Large / XL)
- Transformer-XL ... GPT-2 とほぼ同じ学習法 (比較用)
  - より長い文脈を捉えられるようにしている
- XLNet ... GPT-2 と異なる学習法 (比較)
  - 単語の順序をいろいろ変えて学習

データセット: 6カテゴリの Syntactic Test

- 主語と動詞が隣り合うケース
  - Simple Agreement / Within ORC
- 主語と動詞が副詞 (distractor) に阻まれているケース
  - Across One / Two Distractor(s)
- 主語と動詞が別の名詞 (attractor) に阻まれているケース
  - Across PP / Across ORC

Simple Agreement:
The athlete confuses/*confuse

Within Object Relative Clause:
The friend (that) the lawyers *likes/like

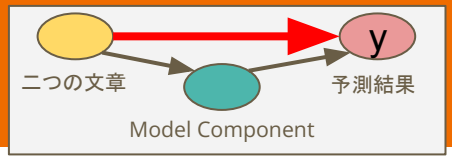
Across One Distractor:
The kids gently *admires/admire

Across Two Distractors:
The father openly and deliberately avoids/*avoid

Across Prepositional Phrase:
The mother behind the cars approves/*approve

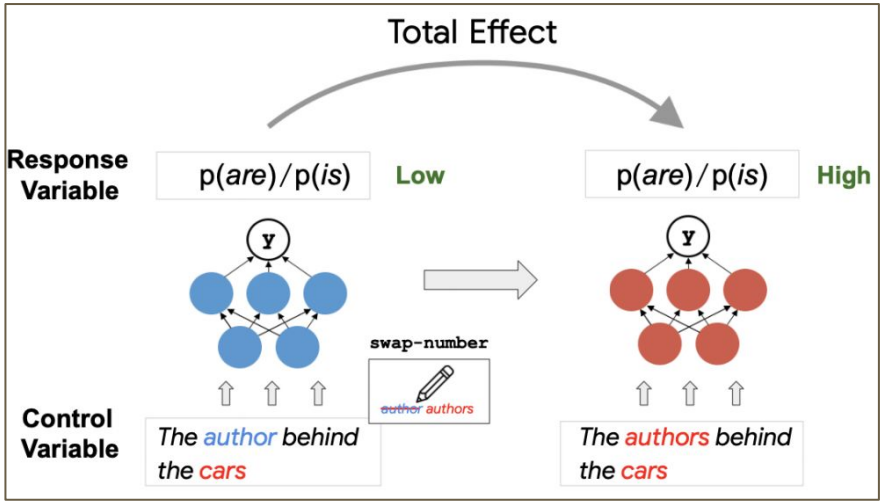
Across Object Relative Clause:
The farmer (that) the parents love confuses/*confuse

実験1 : Total Effect の測定



1. 正解の文章 → 不正解の文章 に変更
2. Response Variable (y) の変化を計算

- 正解のとき → y は小さいほど良い
- 不正解のとき → y は大きいほど良い

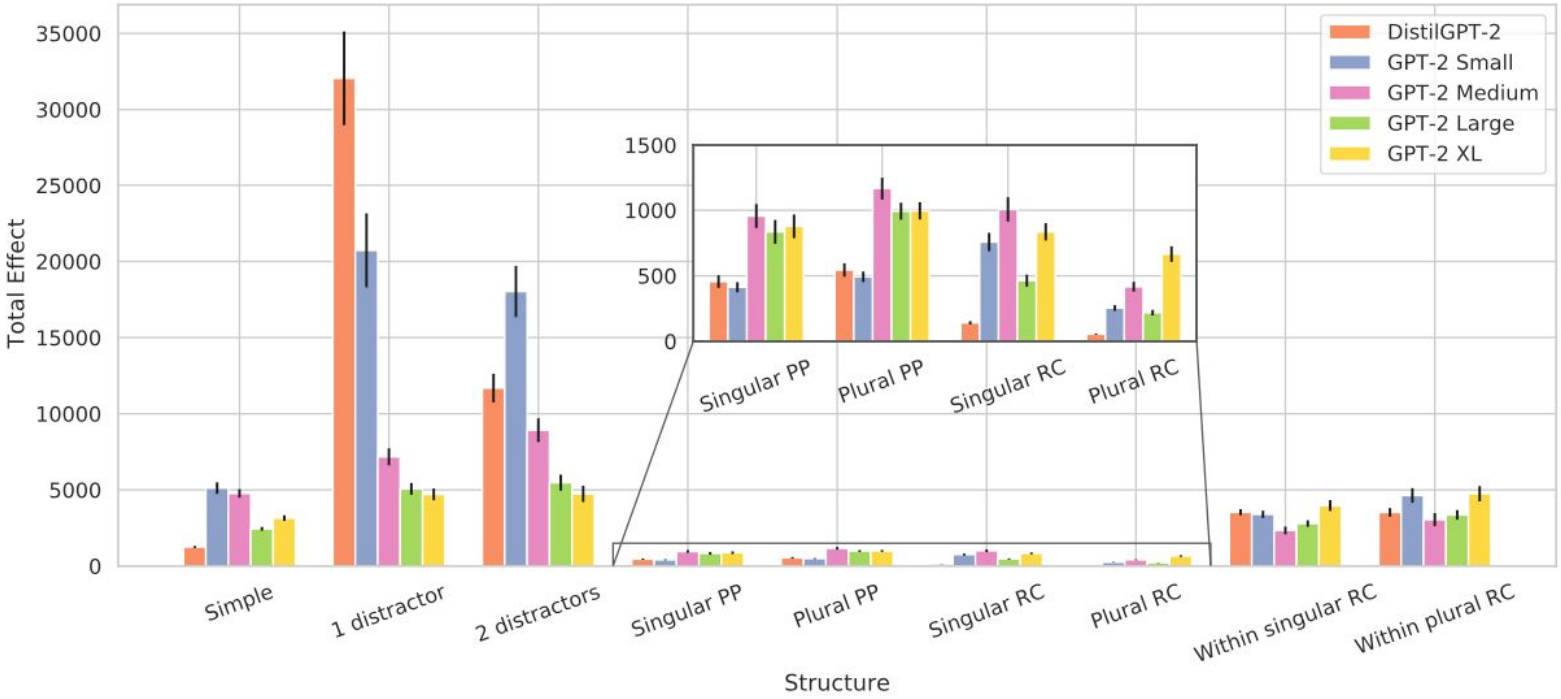


★ Total Effect (TE) : y の相対変化量で定義

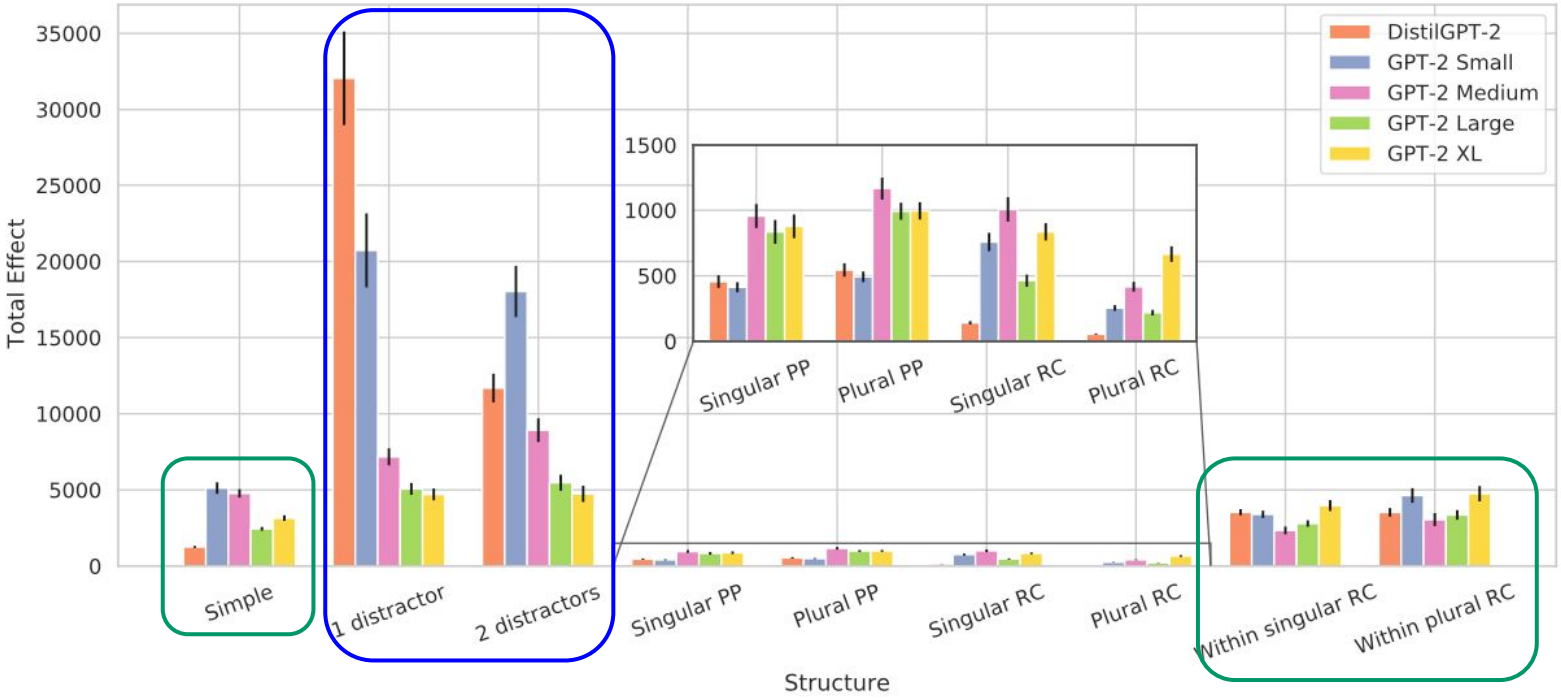
$$\mathrm{TE}(\texttt{swap-number},\ \texttt{null}; y, u, v) = \frac{y_{\texttt{swap-number}}(u_{sg}, v) - y_{\texttt{null}}(u_{sg}, v)}{y_{\texttt{null}}(u_{sg}, v)} = y_{\texttt{swap-number}}(u_{sg}, v)/y_{\texttt{null}}(u_{sg}, v) - 1 = 1/(y_{\texttt{null}}(u_{sg}, v) \cdot y_{\texttt{null}}(u_{pl}, v)) - 1$$

$$\overline{\mathrm{TE}}(\texttt{swap-number},\ \texttt{null}; y) = \mathbb{E}_{u,v}\left[\frac{y_{\texttt{swap-number}}(u, v)}{y_{\texttt{null}}(u, v)} - 1\right]$$

Total Effect の結果



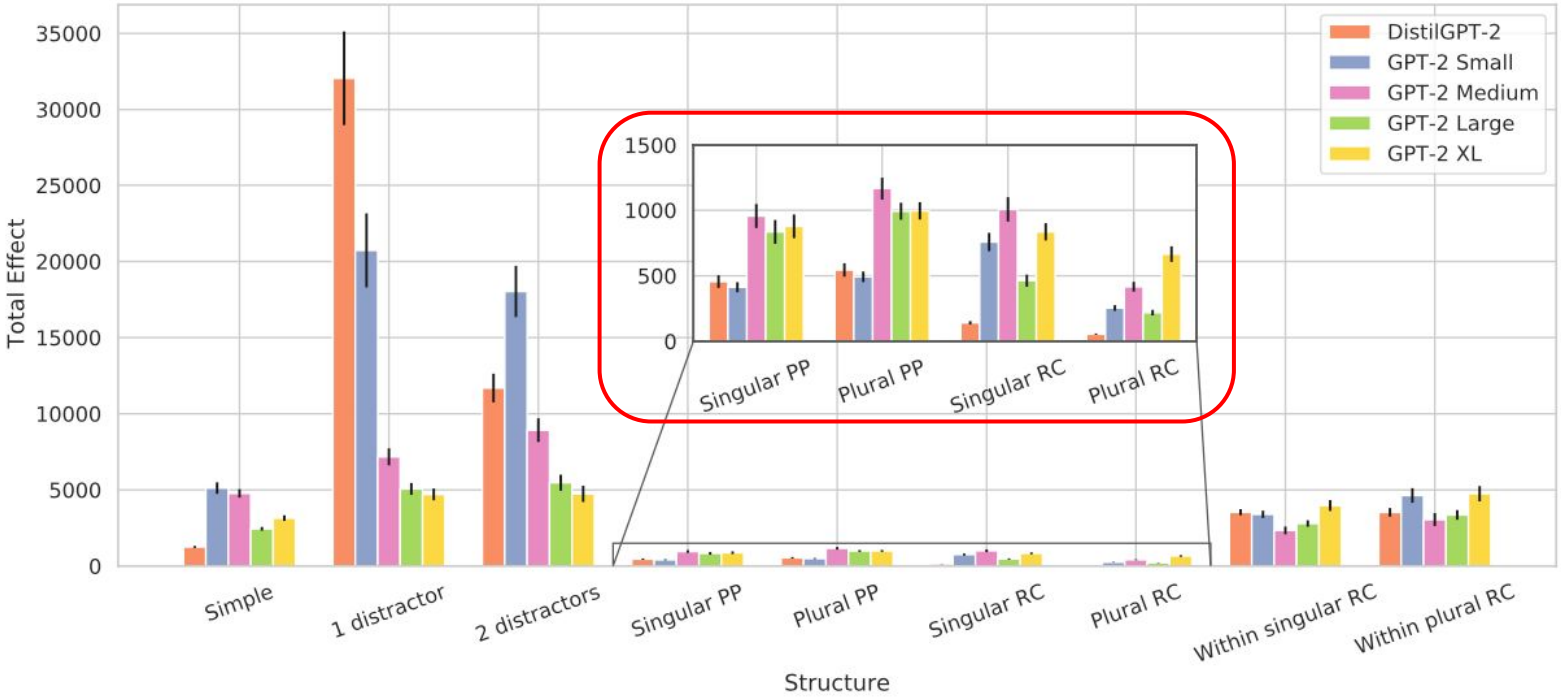
Total Effect の結果



主語と動詞が隣り合うケース：大体似た傾向（比較的高め）

distractor で阻まれているケース：TEが最も高い（！）← 副詞は動詞のサイン？

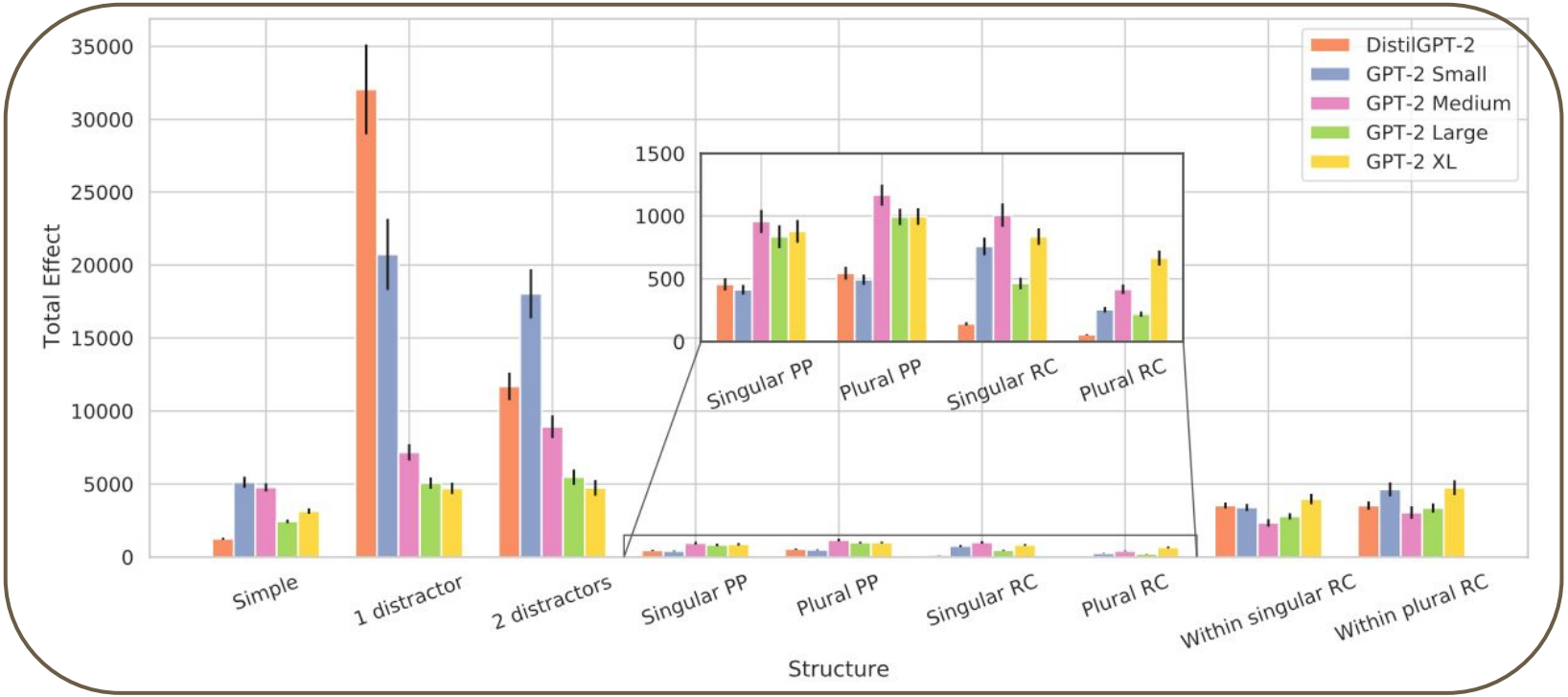
Total Effect の結果



attractor で阻まれているケース

:TEは低い傾向 = 比較的難しいタスク（直観通り）

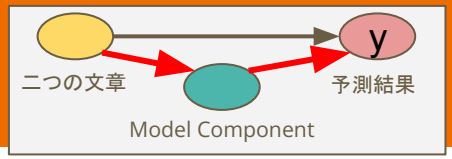
Total Effect の結果



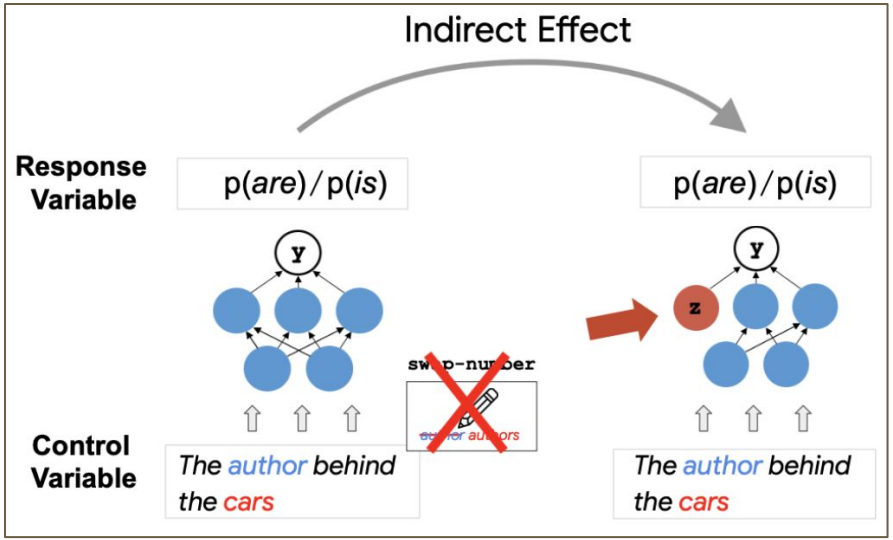
モデルサイズの影響：あまり一貫していない（大きければ良いという話ではない）

（例：PP, RC 系は Medium が大体良い・distractor 系は小さい方が良い

実験2: Indirect Effect の測定



1. まず正しい文章を入力して Response Variable (y) を測定
2. 不正解の文章を与えたときのニューロンの値を固定化
3. 2の元で正しい文章を入力して y を測定、変化を計算
   - 変化が大きい = 不正解の出力に近い = ニューロンの寄与が大きい(Indirect Effect!)



★ Natural Indirect Effect (NIE)

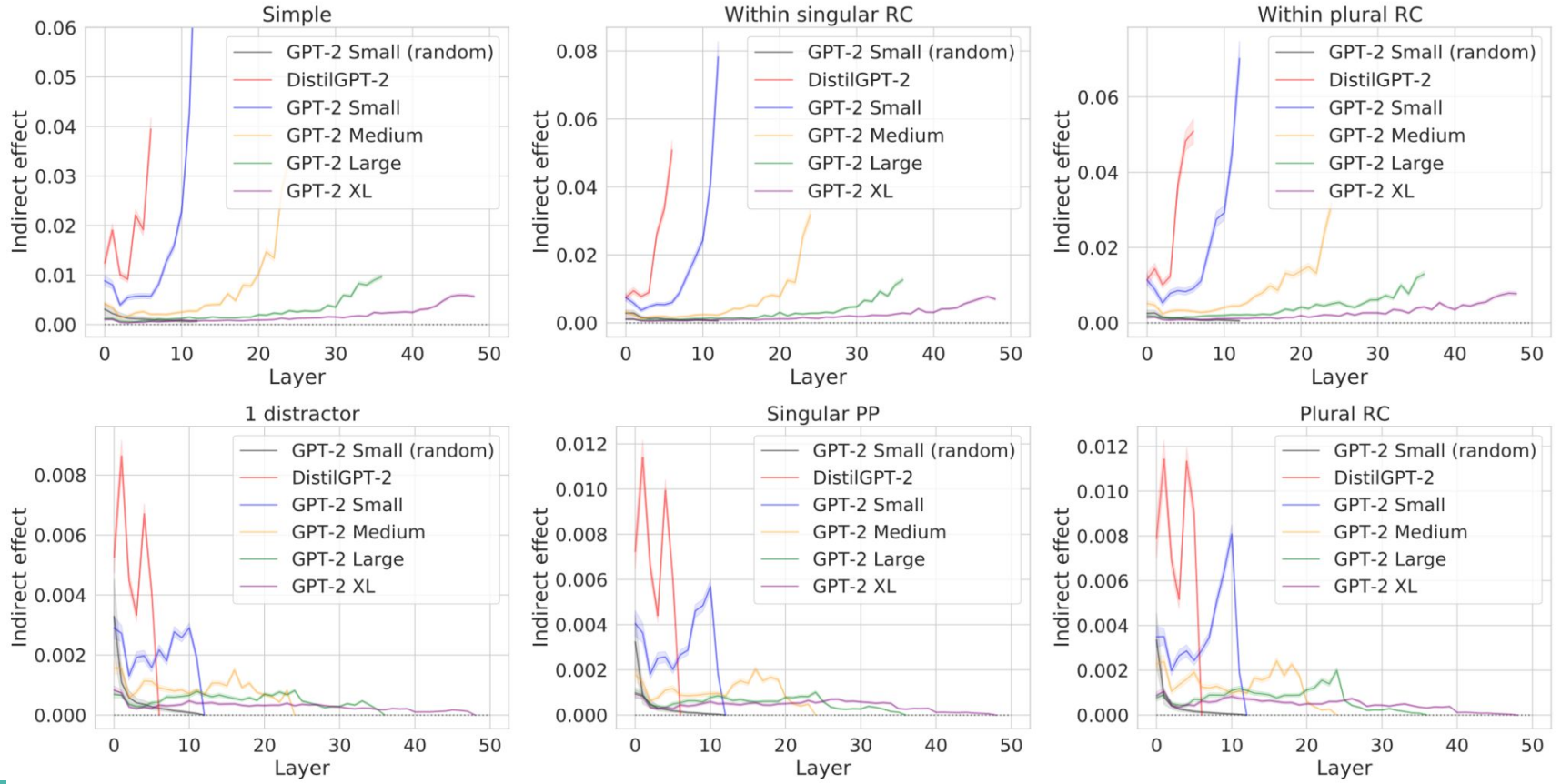
TE と全く同じ定義(y の相対変化量)
扱うモデルが変化(=確率値が変化)

$$\overline{\text{NIE}}(\texttt{swap-number},\ \texttt{null}; y, \mathbf{z}) = \mathbb{E}_{u,v}\left[\frac{y_{\texttt{null},\mathbf{z}_{\texttt{swap-number}}(u,v)}(u,v)}{y_{\texttt{null}}(u,v)} - 1\right]$$

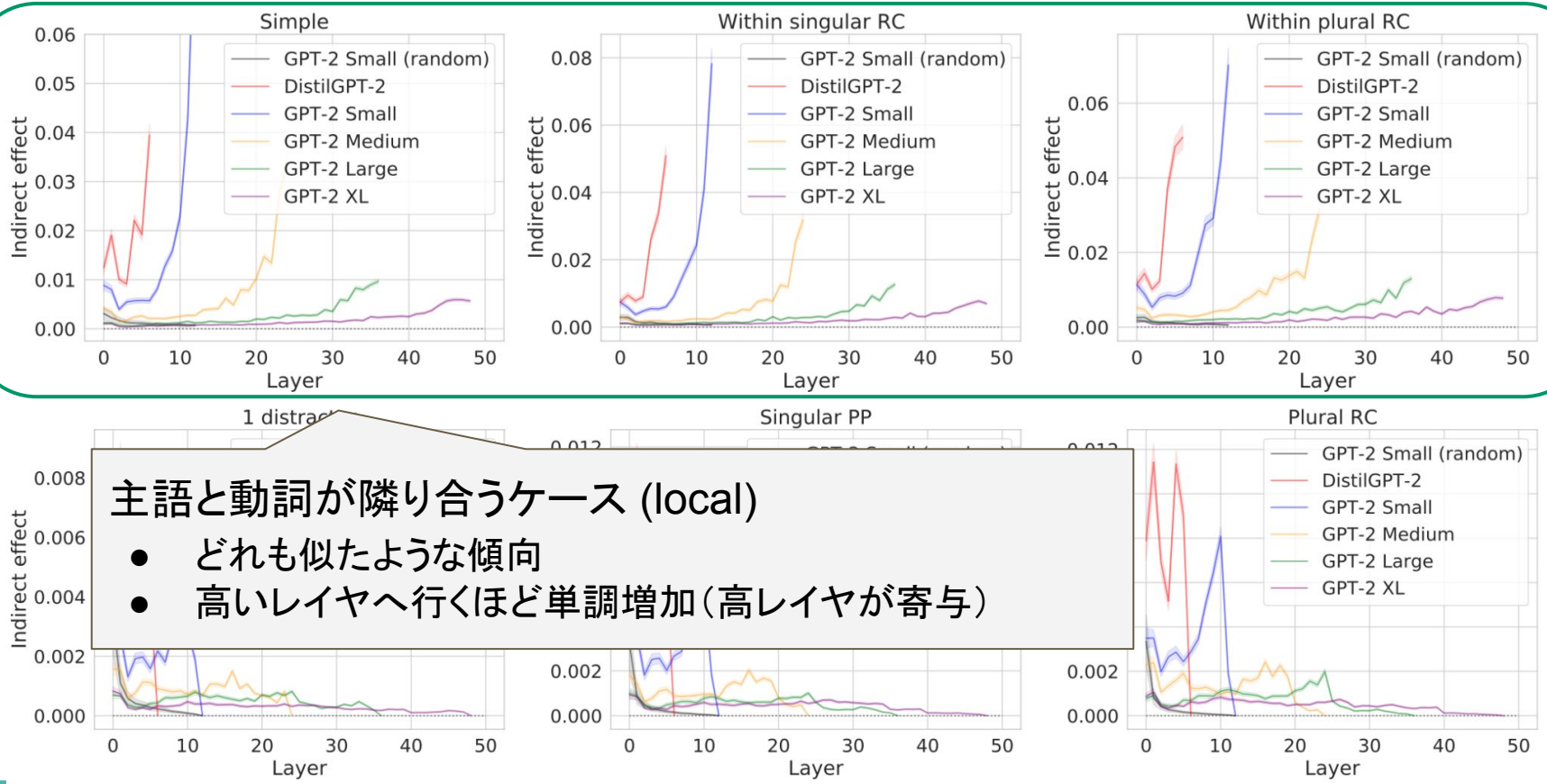
$$\overline{\text{TE}}(\texttt{swap-number},\ \texttt{null}; y) = \mathbb{E}_{u,v}\left[\frac{y_{\texttt{swap-number}}(u,v)}{y_{\texttt{null}}(u,v)} - 1\right]$$

Indirect Effect の結果 ~ GPT-2

※各レイヤにおける NIE が上位5%のニューロンの平均をプロット



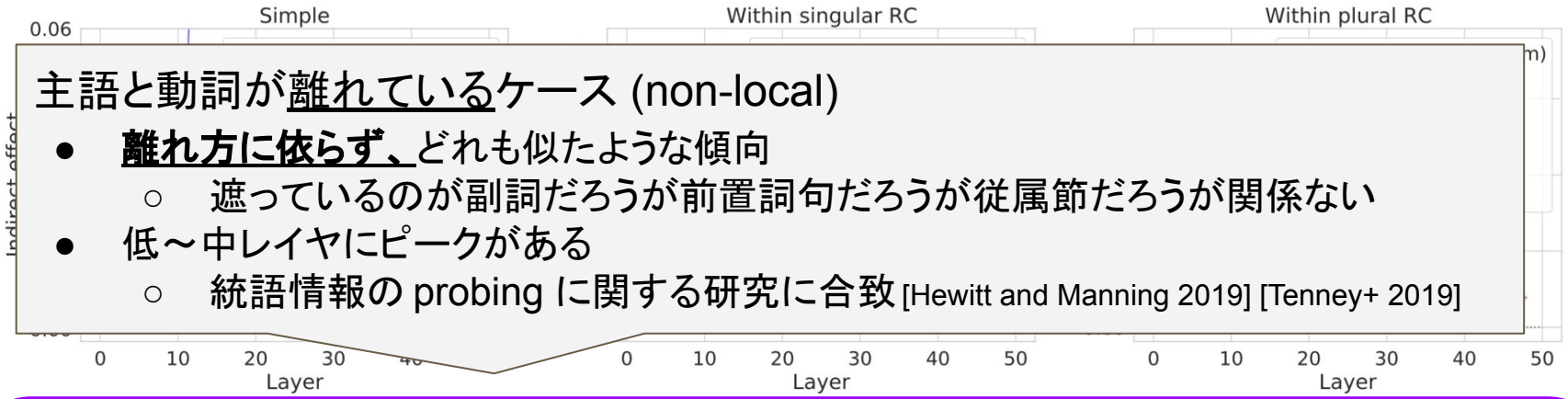
Indirect Effect の結果 ~ GPT-2



主語と動詞が隣り合うケース (local)

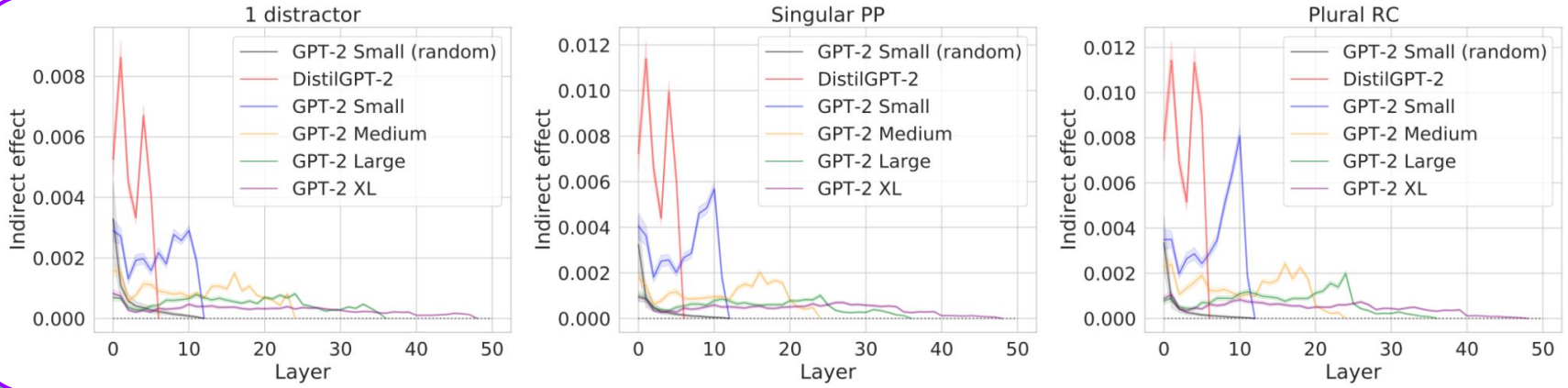
- どれも似たような傾向
- 高いレイヤへ行くほど単調増加（高レイヤが寄与）

Indirect Effect の結果 ~ GPT-2

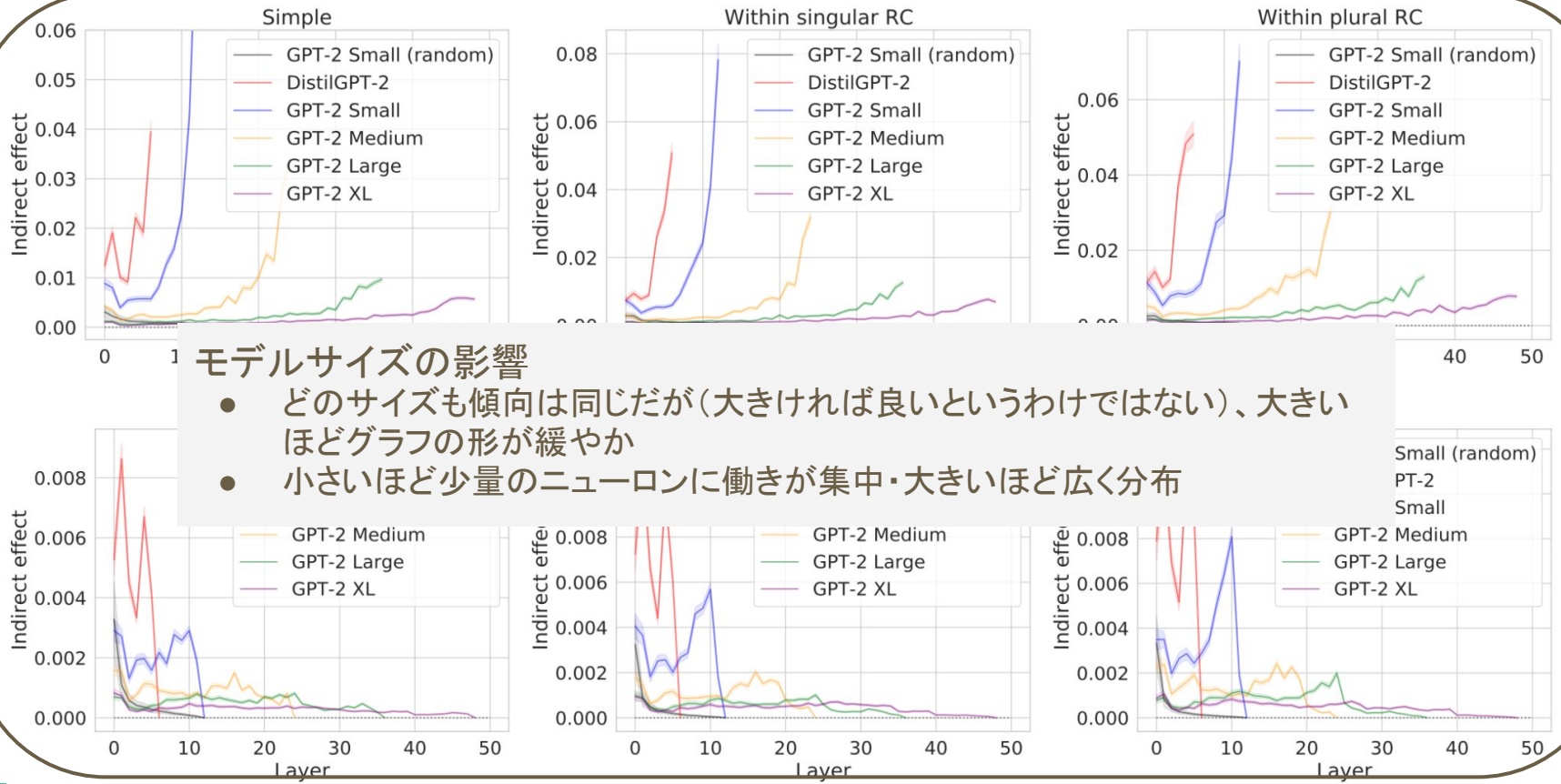


主語と動詞が<u>離れている</u>ケース (non-local)

- **<u>離れ方に依らず、</u>**どれも似たような傾向
  - 遮っているのが副詞だろうが前置詞句だろうが従属節だろうが関係ない
- 低～中レイヤにピークがある
  - 統語情報の probing に関する研究に合致 [Hewitt and Manning 2019] [Tenney+ 2019]



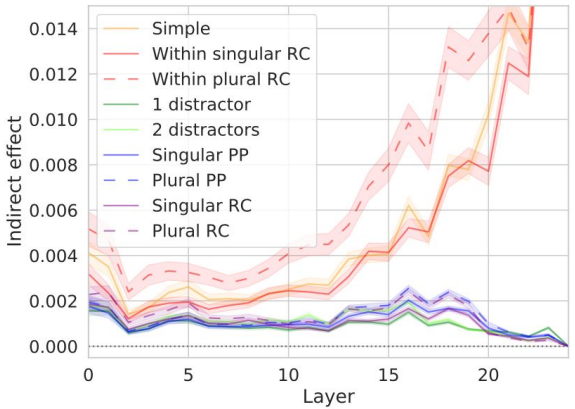
Indirect Effect の結果 ~ GPT-2



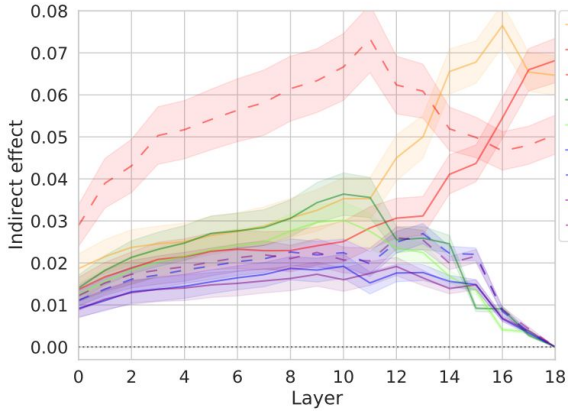
モデルサイズの影響

- どのサイズも傾向は同じだが（大きければ良いというわけではない）、大きいほどグラフの形が緩やか
- 小さいほど少量のニューロンに働きが集中・大きいほど広く分布

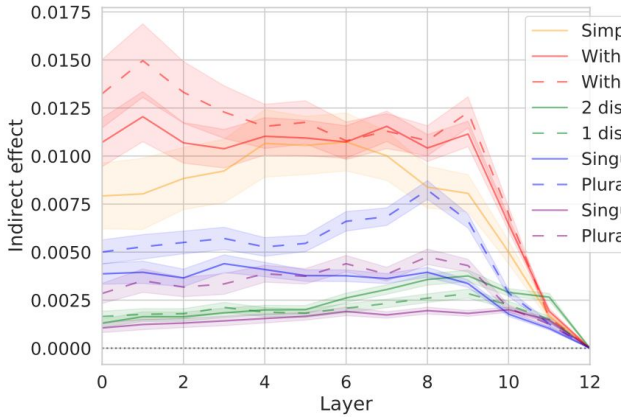
Indirect Effect の結果 ~ GPT-2 と他のモデルの比較



GPT-2 Medium

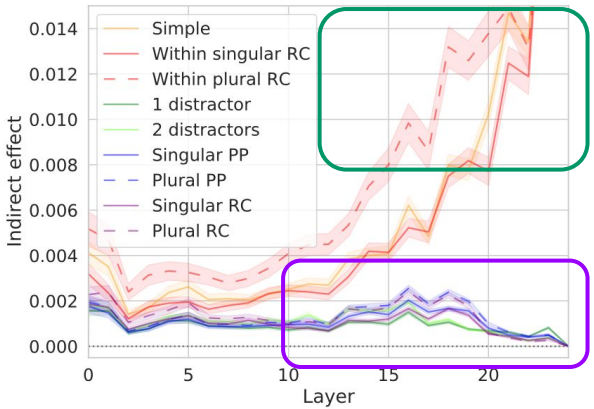


Transformer-XL

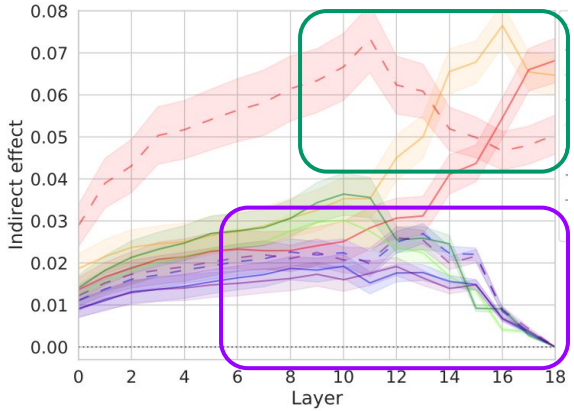


XLNet

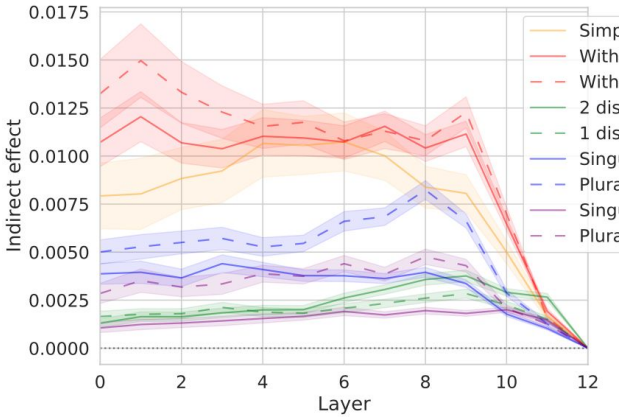
Indirect Effect の結果 ~ GPT-2 と他のモデルの比較



GPT-2 Medium



Transformer-XL

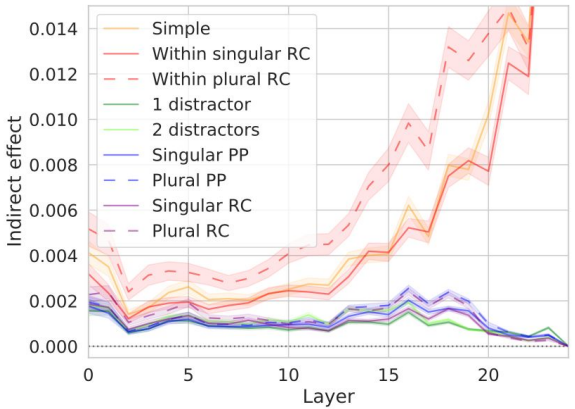


XLNet

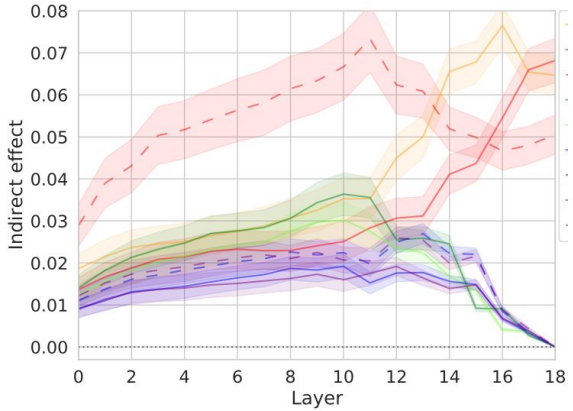
local / non-local で異なる傾向が見られるという点で GPT-2 と共通している

- 全体の傾向もおおよそ似ている
- 学習方法自体が似ていることに起因？

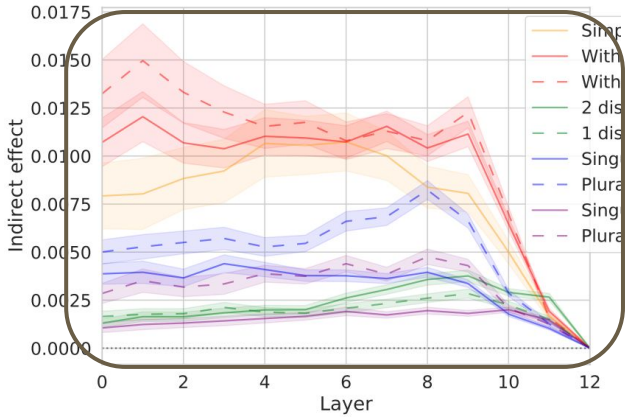
Indirect Effect の結果 ~ GPT-2 と他のモデルの比較



GPT-2 Medium



Transformer-XL



XLNet

local / non-local 問わず似たような傾向

- どちらも GPT-2 における non-local と似た傾向
- 学習方法が異なることに起因？
  - 単語の全順列を用いて学習 → 汎化能力向上？

結論

- Pretrained LM が Syntactic Test を解くメカニズムを因果媒介分析で調査
  - どのニューロンの寄与が大きいかを分析
- モデルサイズ・学習方法の違いによる差を調査
- モデルサイズ：小さいほど少ないニューロンに機能が集中・大きいほど広く分布
  - 大きければ良い、というわけではない
- 学習方法：GPT-2 と Transformer-XL は local / non-local agreement で傾向に違いがある一方、XLNet では違いがない

コメント

- 全体的に個人的にはややモヤモヤ
  - 「メカニズムを調査」という割には浅い感じがする
  - [Lakretz+ 2019] ... LSTM でニューロンの動きまで分析。こういうのを期待していた ...
- syntax を陽に取り入れていないモデルでも、Indirect Effect の大きいレイヤに何らかの傾向が見られるのは面白い
  - attention-based probing とかと一貫した結果 [Clark+ 2019]
- 「local と non-local でメカニズムが変わり得る」は面白い結果だと思うけど一体なぜなのか
  - local にせよ non-local にせよ、agreement という同じ現象を扱っているので、メカニズムが変わらない XLNet の方が優れた振る舞いを示しているとも言える ...？
- 今回得られた知見を新規のモデリングに活かすことはできそうか
- Agreement 以外の Syntactic Test だとどうなる？
  - [Hu+ 2020], [Warstadt+ 2020]