TS2-17

言語モデルにおける長距離依存関係の内部操作

木村一馬 [1, 2]　大関洋平 [3, 2]　菅原朔 [2, 3]

[1] バルセロナ自治大学　[2] 国立情報学研究所　[3] 東京大学

はじめに

- 自然言語に存在する統語的長距離依存関係
- 長距離依存関係を構築する際の言語モデルの内部表現
- 活性パターンのタスク間オーバーラップではなく「**操作内容**」を特定する手法

言語モデルの内部表現/内部操作

「表示」に対する「操作」(Computations over representations)

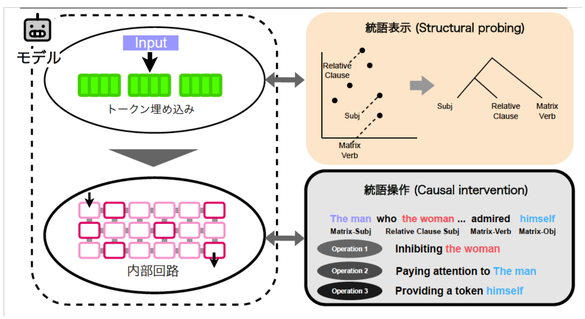


Figure 1. 言語モデル内部の統語表示と統語操作

Activation Patching

Cleann 入力の acitvation を Corrupted 入力に差し込む（パッチ）

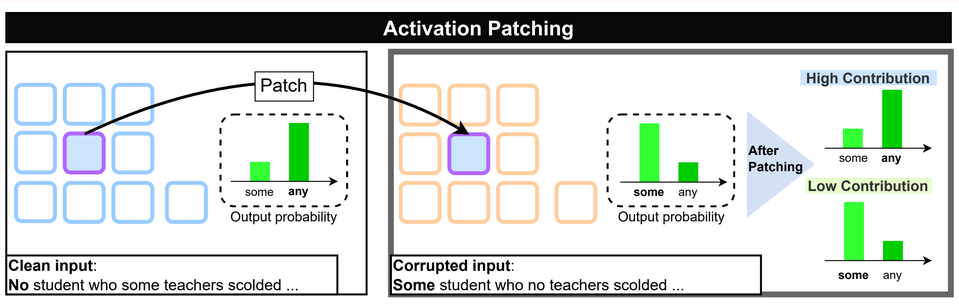


Figure 2. Activation Patching の概略図

データセット

- 4 つの依存関係

	NPI
Clean	**No** student who some teachers scolded submitted [any] notebook.
Corrupted	**Some** student who some teachers scolded submitted [any] notebook.
	Reflexive
Clean	**The man** who the woman predominantly dislikes admired [himself] yesterday.
Corrupted	**The woman** who the woman predominantly dislikes admired [himself] yesterday.
	Wh
Clean	**Which man** did the woman who wrote the paper [visited] yesterday ?
Corrupted	**Which paper** did the woman who praised the man [visited] yesterday ?
	Cleft
Clean	It is **the man** that the woman who wrote the paper [visited] yesterday.
Corrupted	It is **the paper** that the woman who praised the man [visited] yesterday.

Table 1. 依存関係ごとの clean-corrupted 入力のペア例

- 構成素ごとに役割トークンを付与

	$role_{key}$	$role_{rel}$	$role_{rel-subj}$	$role_{rel-verb}$	$role_{adj}$	$role_{main-verb}$	$role_{target}$
NPI	[No NP]	who	[some NP]	[verb]	[adverb]	[verb]	[any]
Reflexive	[The NP_m]	who	[the NP_f]	[verb]	[adverb]	[verb]	[himself]

Table 2. 役割トークンの付与例 (NPI・再帰代名詞)

	$role_{key}$	$role_{pro}$	$role_{subj}$	$role_{rel}$	$role_{rel-subj}$	$role_{rel-verb}$	$role_{target}$
Wh-depedendency	[Wh NP_a]	did	[the NP_i]	which	[verb]	[adv]	[verb]
Cleft	[It was the NP_a]	who	[the NP_i]	[which]	[verb]	[adv]	[verb]

Table 3. 役割トークンの付与例 (wh-疑問文・分裂文)

モデル・実験設計

モデル

GPT-2-Small (12 ヘッド 12 層)/Medium (16 ヘッド 24 層)/Large (20 ヘッド 36 層)

パッチング

Corrupted 入力の活性パターンに, Clean 入力の活性パターン (ヘッド単体 or ヘッド + パターン) を差し込む

$$\text{pattern}_{\ell,h}(i,j) = \text{softmax}\left(\frac{Q_{\ell,h}(i) \cdot K_{\ell,h}(j)}{\sqrt{d}}\right) \tag{1}$$

回復率 (Restoration Score)

パッチ後に Corrupted 入力の Logit 出力がどれだけ Clean 入力の Logit 出力に近づくか

$$R_\ell = \frac{M^{(\ell)}_{\text{patched}} - M_{\text{corr}}}{M_{\text{clean}} - M_{\text{corr}} + \varepsilon}. \tag{2}$$

注意量 (Attention Mass)

ターゲットトークン位置での各トークンへの注意重み

$$\Delta\text{mass(role)} = \text{mass}_{patched} - \text{mass}_{corrupt} \tag{3}$$

結果 (GPT-2-Small)

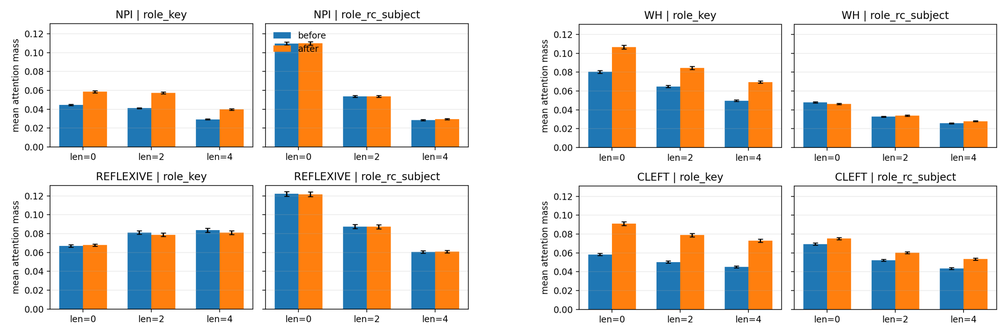


Figure 3. パッチ前後の依存関係ごとの注意量平均

- 全ての依存関係で共通した注意パターン
- 主節主語に注意が向く (平均注意量: NPI, 0.057; 再帰代名詞, 0.067; Wh, 0.106, 分裂文, 0.090)

結果 (GPT-2-Large)

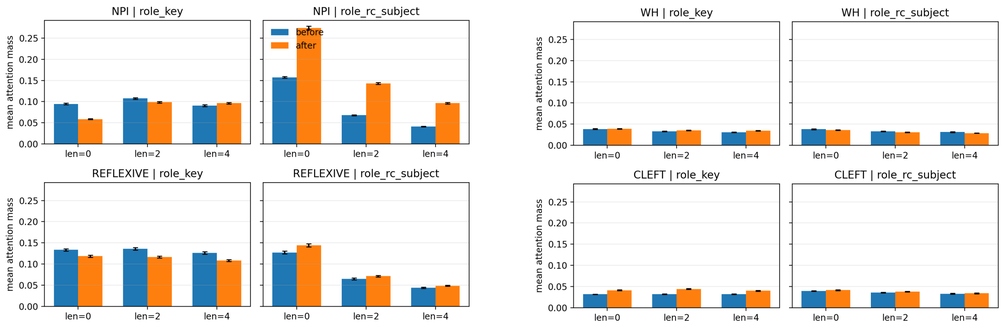
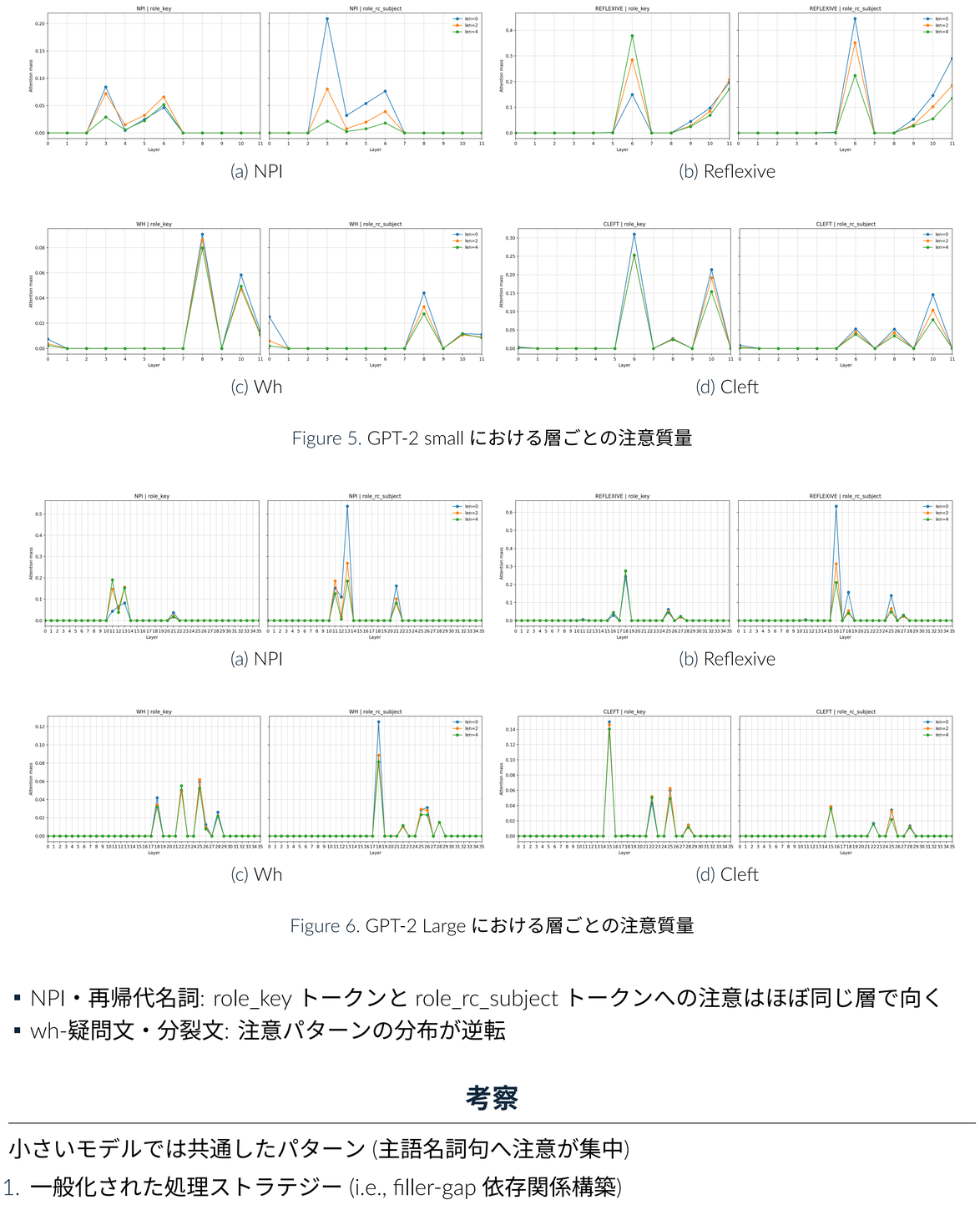


Figure 4. パッチ前後の依存関係ごとの注意量平均

- NPI・再帰代名詞: 関係節内主語に注意が集中 (平均注意量: NPI, 0.25, 再帰代名詞, 0.15)
- wh-疑問文・分裂文: 目立ったパターンは見られない (平均注意量: wh, <0.05, 分裂文, <0.05)

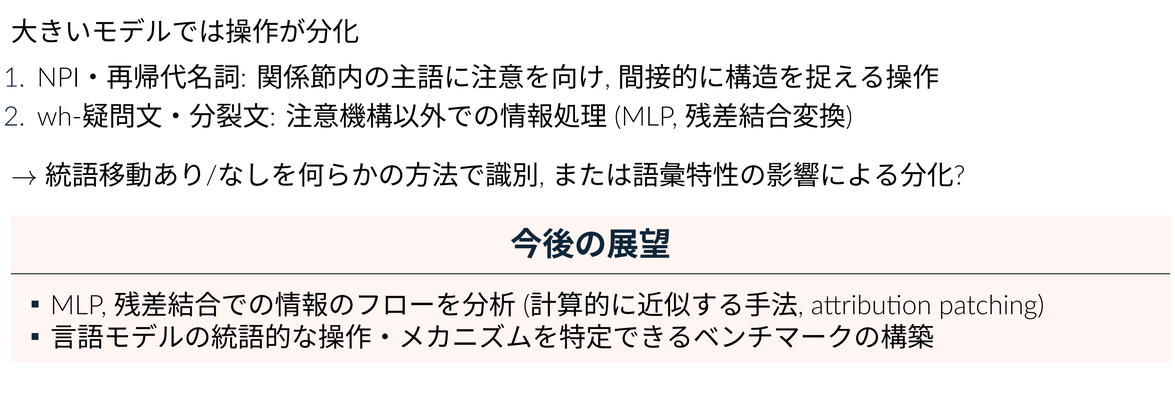
線形距離分析

GPT-2-Small, Large どちらにおいても, 線形距離による分布の変化は見られない



(a) NPI　(b) Reflexive　(c) Wh　(d) Cleft

Figure 5. GPT-2 small における層ごとの注意質量



(a) NPI　(b) Reflexive　(c) Wh　(d) Cleft

Figure 6. GPT-2 Large における層ごとの注意質量

- NPI・再帰代名詞: role_key トークンと role_rc_subject トークンへの注意はほぼ同じ層で向く
- wh-疑問文・分裂文: 注意パターンの分布が逆転

考察

小さいモデルでは共通したパターン (主語名詞句へ注意が集中)

1. 一般化された処理ストラテジー (i.e., filler-gap 依存関係構築)
2. 粗い線形ヒューリスティック

大きいモデルでは操作が分化

1. NPI・再帰代名詞: 関係節内の主語に注意を向け, 間接的に構造を捉える操作
2. wh-疑問文・分裂文: 注意機構以外での情報処理 (MLP, 残差結合変換)

→ 統語移動あり/なしを何らかの方法で識別, または語彙特性の影響による分化?

今後の展望

- MLP, 残差結合での情報のフローを分析 (計算的に近似する手法, attribution patching)
- 言語モデルの統語的な操作・メカニズムを特定できるベンチマークの構築