Interpretability of Language Models via Task Spaces

Lucas Weber, Jaap Jumelet, Elia Bruni, Dieuwke Hupkes (ACL 2024)
(紹介者: フューチャー 森下 睦)

背景とResearch Questions

FUTURE

■ LLMは様々な言語タスクを解ける

- しかし、内部的にどの程度言語を理解してタスクを解いているのかわからない
- 個別タスクをそれぞれ覚えているのか、タスク横断的な知識を獲得できているのか
- 本論文では文法知識に着目して、どのようにLLMが様々な文法を理解しているのか検証
  → 様々な文法の共通知識のようなものを覚えているのか

■ Research Questions

1. LLMはタスク横断的な文法知識を獲得しているのか
2. モデルサイズによる差はあるのか
3. どのようにタスク横断的な理解が進むのか

どうやって調べる？

■ 対象タスク: 英語文が文法的に正しいか判定

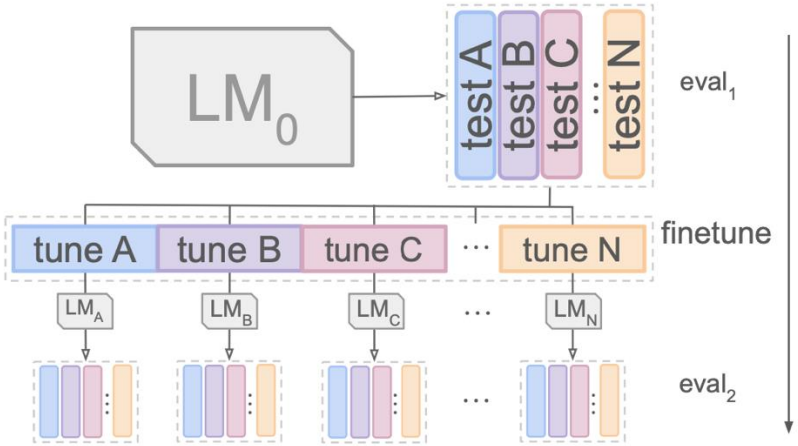
- 複数の文法的タスクを対象
- 例: 三単現が一致するか
  主語と数量表現が一致するか

◯ The cats annoy Tim.
× The cats annoys Tim.

◯ Susan revealed herself.
× Susan revealed themselves.

■ 調査方法

1. Pre-train済みLMを用意
   →各文法タスクの性能を測る
2. 各文法タスクデータを使ってFine-tuning
   → N個のモデルができる
3. 得られたN個のモデルそれぞれで全ての文法タスクの性能を測る
4. 各モデルがどの程度Fine-tuning前後で性能が伸びるかを測る
   → 直接Fine-tuningしたタスク以外の性能も上がるのか



■ 言語のもつれ (Linguistic Entanglement)

- 言語の特性上各タスクを完全に分離することはできない
  - 例：三単現の一致などの言語現象はどの文にも常に出現
- タスクごとにFine-tuningして、タスク横断的な能力を測りたいのに、
  言語のもつれによって分離不可能な領域が生まれ、正確な分析が行えない
  - 例：三単現の一致はどのタスクでFine-tuningしても上がる（はず）
  - でも本当に知りたいのは、各タスクを完全に分離したうえでの分析結果

言語のもつれを解く

■ 提案法: 勾配の差を用いたFine-tuning

- **仮説:** 学習時の勾配は複数タスクの勾配の線形結合
- 対象文法タスクを学習する際、常に二つの文を用意
  - **正例 (+):** John did **not** see anything. (文法的に正しい、対象タスクを含んだ文)
  - **負例 (-):** John did see anything. (上記とほぼ同じ文だが、文法的に正しくない文)
- これらの文を学習する際の勾配の差g^{Δ}を使ってFine-tuning
  - $g^{\Delta} = g^{+} - g^{-}$
  - このとき、勾配が小さいパラメータも含めてアップデートすると後に比較しにくい（らしい）ので、勾配が一定 (10^{-3}) 以上の一部のパラメータのみをアップデート
    - アップデートするパラメータの数は全体の5%程度になる

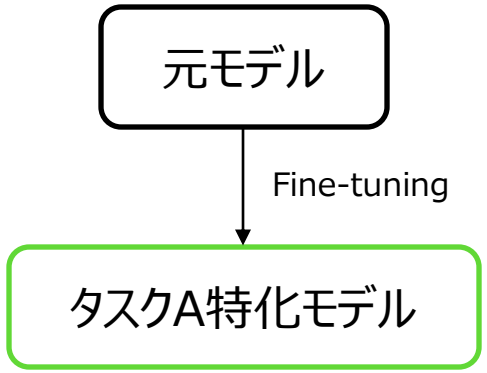
■ (森下の疑問)

- こんな正例、負例を用意することなんて可能・・・？
  - BLiMPというコーパスを使用
  - このコーパスは機械的に正例から負例を作っている（らしい）
  - ただ、これが可能な言語現象は限られているので、上記提案法は汎用的ではないと思う

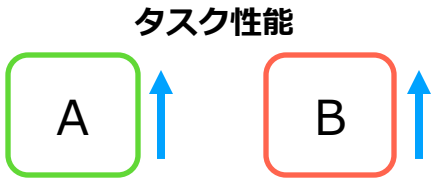
タスク間の類似度の測り方

■ Transfer Probing

- ➢ Fine-tuning前後のタスク性能に着目した方法

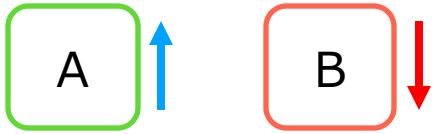


パターン1



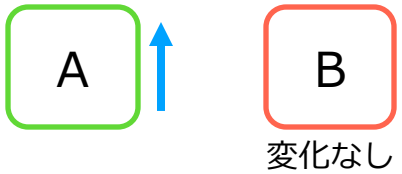
Task B も性能向上
→ Task AとBは**関係あり**、**類似度高**

パターン2



Task B は性能低下
→ Task AとBは**関係あり**、**類似度低**

パターン3



Task B は変化なし
→ Task AとBは**関係なし**

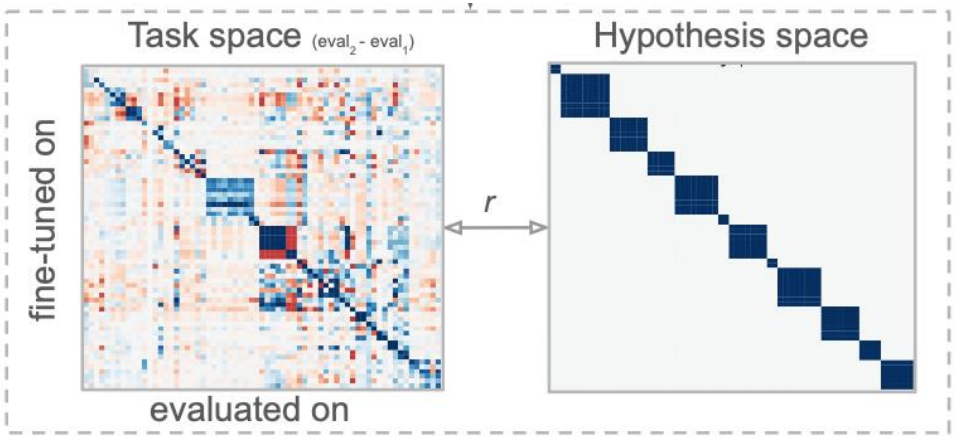
実験設定

- **モデルサイズ:** 27M, 70M, 203M
- **タスク:**英語文が文法的に正しいか判定
  - 出力は正しい or 正しくないの2値
- **データセット:** BLiMP
  - 67の言語現象を対象に作られたデータセット→
  - **Phenomenon** (12種類)
    - 照応一致、動詞の三単現の一致 など
  - **Paradigm** (67種類)
    - Phenomenonを細分化したもの
      - 例: 照応一致
        → 性別 (him, her) の一致
        → 数量 (herself, themselves) の一致 など

Phenomenon	Paradigm	Index
anaphor agreement	anaphor gender agreement	1
	anaphor number agreement	2
argument structure	animate subject passive	3
	animate subject trans	4
	causative	5
	drop argument	6
	inchoative	7
	intransitive	8
	passive 1	9
	passive 2	10
	transitive	11
⋮	⋮	⋮
subject verb agreement	distractor agreement relational noun	62
	distractor agreement relative clause	63
	irregular plural subject-verb agreement 1	64
	irregular plural subject-verb agreement 2	65
	regular plural subject-verb agreement 1	66
	regular plural subject-verb agreement 2	67

実験結果: LLMにおけるタスクの類似性

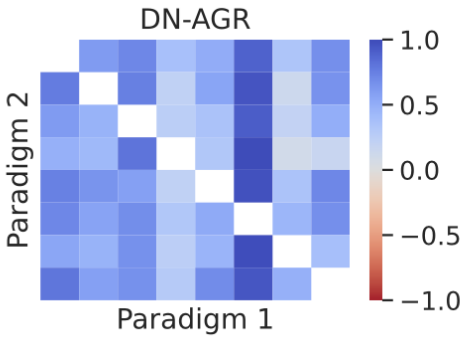
- Task Space
  - Fine-tuning前後での性能の上がり下がりを示した図
- Hypothesis Space
  - 同じPhenomenon内の領域を示した図
  - 色が塗られている部分が共通知識があるかも？と推測される部分
- Task SpaceとHypothesis Spaceはそこそこ似ている
  - あるタスクAで学習したモデルが類似タスクBでも性能が上がることがある
  - タスク共通の知識みたいなものを理解している？



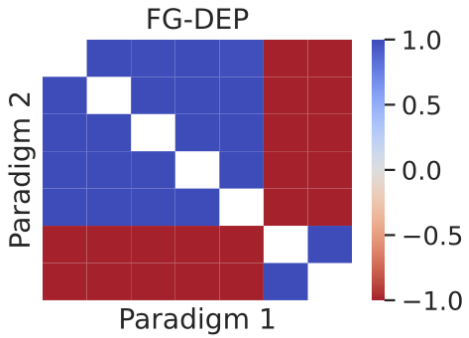
(67, 67) の行列

実験結果: 類似する言語現象内のタスク横断性

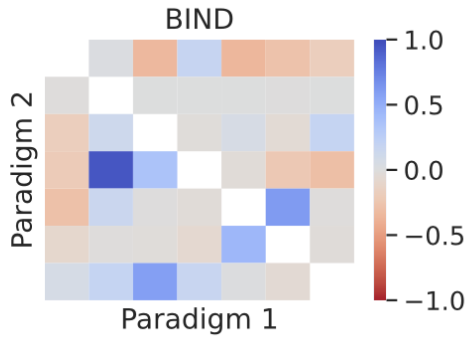
パターンA
(限定詞と名詞の一貫性)



パターンB
(関係代名詞の使い分け等)



パターンC
(代名詞の照応)



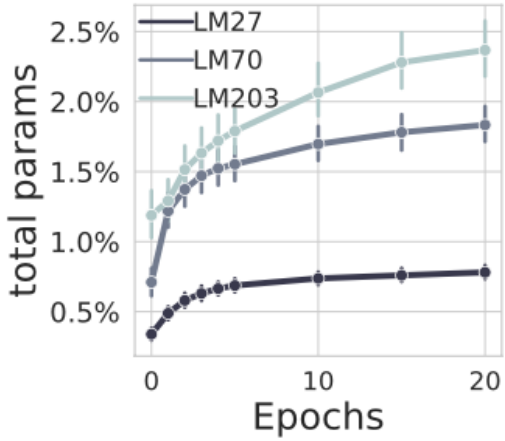
- **パターンA:** Phenomenon内の全てのParadigmが高い類似度を持つ
  - タスクの共通性を利用できている
- **パターンB:** Phenomenon内の一部のParadigmが高い類似度を持ち、一部は低い
  - タスクの一部は対立している
- **パターンC:** Phenomenon内のParadigmに関係が見いだせない
- モデルサイズを大きくすることで 高い類似度を持つ言語現象が増えていく（下図）
  - 言語現象の共通点に気づけるようになる

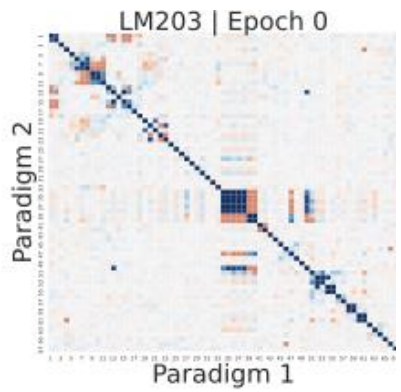
	A-AGR	ARG-S	BIND	CON-R	DN-AGR	ELLIP	FG-DEP	IRR-F	ISL-E	NPI-L	QUANT	SV-AGR
LM27	0.07 ±0.19	0.02 ±0.18	0.03 ±0.19	-0.07 ±0.33	0.24 ±0.12	0.39 ±0.06	0.03 ±1.0	-0.42 ±0.41	0.13 ±0.52	0.2 ±0.44	0.19 ±0.38	0.05 ±0.32
LM70	0.08 ±0.18	0.04 ±0.4	0.03 ±0.12	-0.07 ±0.58	0.33 ±0.07	0.6 ±0.05	0.04 ±0.99	-0.62 ±0.35	0.13 ±0.43	0.18 ±0.37	0.41 ±0.39	0.18 ±0.35
LM203	0.11 ±0.36	0.07 ±0.25	0.03 ±0.26	-0.01 ±0.44	0.56 ±0.23	0.6 ±0.02	0.05 ±1.0	-0.97 ±0.05	0.21 ±0.47	0.2 ±0.36	0.48 ±0.36	0.33 ±0.12

実験結果: 学習過程の分析

■ どのタイミングで言語の知識を得る?

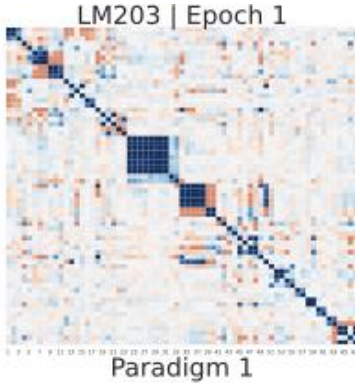
- タスク横断的な知識は初期から学習される (下図)
  - 突然タスク横断的な知識を得ることはない
- また、学習に使うパラメータの領域は徐々に増えていく（右図）





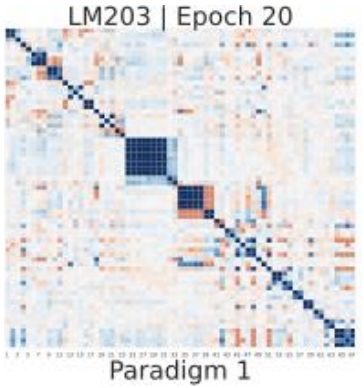
Epoch 0

結構変わる



Epoch 1

ほぼ一緒



Epoch 20

実験結果: 人間の学習過程との比較

- LLMの学習の仕方は人とは異なる？
- 人
  - 人の発達には段階があり、色々な知識を得ることで、できることが増えていく
    - **ピアジェの認知発達段階説**
- LLM
  - 突然タスク間の共通点に気づき始めることはなく、徐々に能力が伸びていく
    - 段階があるわけではなさそう
    - 著者「LLMでCurriculum Learningが難しい理由の1つかも」
  - （最先端NLPつながり）
    → 一方で文法能力は突然獲得されるという話もある

ピアジェの認知発達段階説

画像引用元: https://how-kids.com/knowledge/glossary/2933/

■ Research Questions

1. LLMはタスク横断的な文法知識を獲得しているのか
   → Yes. 人間が文法的に近いと考えるタスクは横断的に知識を使える場合が多い

2. モデルサイズによる差はあるのか
   → Yes. 大きなモデルのほうがよりタスクの共通知識を見つけやすい

3. どのようにタスク横断的な理解が進むのか
   → 学習初期からすぐに学び始める。学習途中に突然気づくわけではない

ここからどうするか

- 他のタスクでLLMの弱点を分析すると良い？
  - タスクによっては人間が考える共通知識みたいなものが獲得できていないものもある？
  - なぜ獲得できない？
- 一方で今回の分析ができるタスクは限られている気がする
  - 「言語のもつれ」問題を解消するために、違いが少ない正例・負例を用意
  - こんなことは全てのタスクでできるわけではない

所感

- 正直結果の驚き度合いは少ない
  - 人間の想定とある程度一致していることがわかったのが嬉しさなので、それはそうだが
- LLM時代にはあまりにも小規模なモデル (~203M)
  - 数B~のモデルの分析が気になるが・・・？
  - 本当に似たような傾向になる・・・？
- どのニューロンが文法的な特徴を捉えているのか気になる
  - （最先端NLPつながり） Language Specificなニューロンは、浅い層と深い層に多いという話
  - 今回タスクの性能に影響しているニューロンは、層ごとに量が異なっていたりするのだろうか
    どう増えていくんだろうか
  - （時々似たような議論があるが・・・）