事前学習済み言語モデルの知識に基づく演繹推論能力の調査

穀田一真 [1]　長澤春希 [1]　Benjamin Heinzerling[2,1]　乾健太郎 [1,2]
[1] 東北大学　[2]理化学研究所
{kokuta.kazuma.r3, haruki.nagasawa.s8}@dc.tohoku.ac.jp
benjamin.heinzerling@riken.jp　kentaro.inui@tohoku.ac.jp

概要

近年，ニューラルネットワークの実用化に伴い，自然言語で記述された簡単な推論の問題に対して高い正解率で解答することができる言語モデルが複数提案されてきた．しかしながら，ニューラル言語モデル (LM) による推論の過程は一般的にブラックボックスであり，LM に論理的な推論過程を再現する能力があるかどうかは依然として疑問が残されている．本稿ではこの疑問に対し，訓練の過程において LM が，基礎的な推論能力の 1 つである演繹推論を行う傾向があることを示すことにより，以降 LM の推論能力を調査する足掛かりとする．

1　はじめに

計算機を用いた論理推論の試みは，一階述語論理などの形式表現を用いたものに始まり，現代の自然言語処理分野に至るまで議論が続く重要な議題の 1 つである．自然言語処理分野では近年，インターネットの普及による大規模言語情報の蓄積やニューラルネットワークの実用化・大規模化に伴い，RNN や Transformer [1] といったニューラル言語モデル (以下，LM) の基盤となるアーキテクチャが実用化された．そして，それらの実用化に伴い，bAbI tasks [2] などの自然言語で記述された推論問題を含むベンチマークに対し，高い性能を示すことを目的とした LM が複数提案されてきた [3, 4, 5].

なお，これらの提案では，訓練時に与えられた推論問題のデータをもとに推論の**ルール**を学習し，評価時に入力された**前提**に対して推論を行うことができるといった**推論能力**を LM が獲得できることが前提として想定されている．しかしながら実際には，ブラックボックスである LM による推論の過程を追跡することは困難であり，ベンチマークにおける性能に関わらず，これまでに提案された言語モデルがそのような推論能力を実際に獲得できているかどうかは未だ明らかにされていない [6, 7].

ここで，LM がベンチマークにおいて高い性能を示すにも関わらず，推論能力の獲得が十分でない可能性があると考える要因としては，(1) モデルに入力として与えられる問題・解答の偏りを用いたり，(2) 入力文を推論訓練前の事前訓練によって獲得した知識と併用したりすることにより，適切な推論過程を経ずに解答を予測する「ショートカット」が発生することが挙げられる．本研究では図 1 に示すように，推論に必要な前提を，従来のベンチマークによる手法とは異なり，モデルへの入力ではなく**事前学習した知識**として与えた上で，結論に関する LM の振る舞いを観察する．すなわち，LM の推論能力のうち，ショートカットの問題が関与しない部分の調査を行うものとした．

なお，本研究では LM の推論能力を調査する足掛かりとして，**演繹推論** (Deductive reasoning, Deduction) に限定した調査を行った．ここで，演繹推論とは例えば「カラスは鳥である」「全ての鳥は卵を産む」という前提から「カラスは卵を産む」という結論を得る推論過程である．またこの推論過程では，帰納推論 (Induction) やアブダクション (Abduction) [1)] などの他の論理推論過程と異なり，結論は与えられた前提のみを用いて断定的に導き出すことができる．したがって，推論訓練前の事前訓練 (マスク穴埋めなど) において獲得した知識を必要としないため，ショートカットの発生と推論過程を切り分けることができる．

2　関連研究

2.1　知識に基づく推論

事前学習した知識に基づく推論能力を調査する研究としては，Property Induction Framework [8] が挙げられる．なお，同論文では LM の**帰納推論能力** (付

1)　3 種類の論理推論の説明を付録 A に示す．

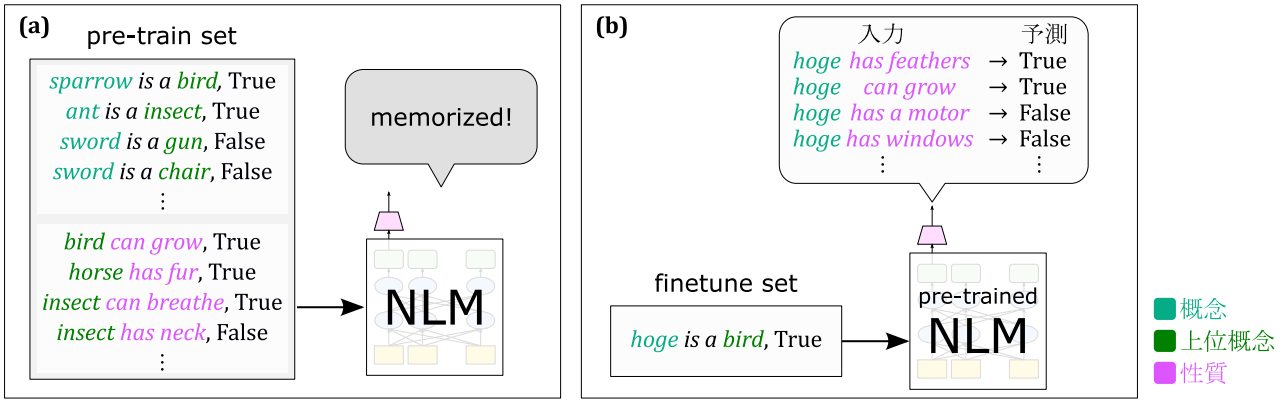

図 1　本研究における調査手法の概要．**(a)** は前提による事前学習，**(b)** は結論の予測能力の調査の工程をそれぞれ表している．また，工程 **(b)** における概念 *hoge* は工程 **(a)** で出現しないランダムな埋め込みを持つトークンである．

録 A) を調査対象とし，「カラスは foo できる」という例から，「ハトは foo できる」といった明示していない「同じカテゴリへの汎化」が発生するかどうかに主眼を置いている．それに対し，本稿では LM の**演繹推論能力**を調査対象とし，断定的に導かれるべき結論を予測することができるかどうかに主眼を置くという点で目的が異なる．

2.2　LM による演繹推論

本研究では，既存の LM を対象とした演繹推論能力の調査を行うことを目的とするが，同様の目的で行われた既存研究としては，Transformer [1] を対象とした Soft Reasoner [6] が挙げられる．しかしながら同論文では推論に必要な前提の情報をモデルの入力として与えていることから，先述のようなショートカットの問題が発生しうることが考えられる．また同論文中では，本稿において「推論能力」と定義したような振る舞いをしているかどうかについては未解決であることが記されている．

3　調査手法

3.1　用語の定義

本稿で用いるいくつかの重要な用語を定義する．

上位下位関係　ある概念 c がその上位概念 h に属することを表す関係，あるいはその関係を記述した文 "c is a h".

例：• "sparrow is a bird"

概念の性質　ある概念 c あるいは上位概念 h が性質 p を持つことを表す関係，あるいはその関係を記述した文 "h p" または "c p".

例：• "bird can grow"　　• "sparrow can grow"

前提　演繹推論において，前提として与えられる事象．「上位下位関係」あるいは「上位概念の性質」で記述される．

例：• "sparrow is a bird"　　• "bird can grow"

結論　演繹推論において，与えられた「前提」から導き出される事象．「上位でない概念の性質」で記述される．

例：• "sparrow can grow"

これらの定義に従うと，演繹推論は与えられた「前提」から「結論」を導き出す推論過程と言える．

3.2　手法の概要

本研究では，事前学習した前提に関する知識をもとに，LM が結論を演繹的に導き出す振る舞いをどの程度できるかを調査する．調査の手順として，初めに (a) 推論に必要な前提の集合で LM の事前学習を行い，その後 (b) それに基づく結論を予測する能力の調査を行う．より具体的には，図 1 に示すように，工程 (a) では与えられた前提の真偽に関する二値分類問題を LM ができるだけ正しく予測できるようになるまで学習させ，工程 (b) では新しく追加学習させた上位下位関係に対して事前学習した上位概念の性質を演繹的に汎化する傾向がみられるかどうかを調査する．なお，工程 (a) および工程 (b) で行った実験の詳細については，それぞれ続く 4 節および 5 節で説明を行う．

3.3　調査対象のモデル

本研究の実験では，調査対象のモデルとして，マスク穴埋め問題によって事前学習された Transformer ベースのモデルのうち，ALBERT XXLarge v2 [9], BERT large (uncased) [10], RoBERTa large [11] の 3 種

類を選択した．これらのモデルを使用することで，分類問題用のヘッドを用いてモデル内のパラメータを訓練したのち，ヘッドをマスク穴埋め用のものに付け替えることでモデルが保有する知識の変化をマスク穴埋め問題を用いて観察することができる．なお，この観察を行った結果は付録 B に示す．

4　前提による事前学習

この工程では，前提の真偽に関する二値分類問題のデータセットを作成し，LM の事前学習を行った．この節の各小節では，そのデータセットの作成手順，訓練時の実験設定，および LM が与えた前提をどの程度記憶することができたかの評価について説明していく．

4.1　事前学習データセットの生成

図 1 (a) に “pre-train set” として示す，前提の真偽に関する二値分類問題のデータセット (以下，事前学習データセット) を作成した手順を以下に示す．

データの準備

前提の真偽に関するデータを作成するためには，その構成要素である「上位概念の性質」および「上位下位関係」を定義する必要がある．

本実験ではこれらの定義に，Misra らが Property Induction Framework [8] の実験において作成した**概念の性質**を記述したデータセット，および WordNet [12] が提供する**上位語の情報**を利用した．なお前者は，Misra らが Property Induction Framework [8] の実験において作成した， Cambridge Centre for Speech, Language, and the Brain が収集した concept property norms データセット [13] からデータ間の矛盾や不正確なデータを除いたデータセットである．なお，このデータセットを用いる上で行った前処理を付録 C に示す．

上位下位関係

Misra らが作成したデータセットに含まれる概念 (および対応する WordNet の語義情報) と WordNet の上位語の情報をもとに，概念に対する上位概念を定義した．また，新しく定義した上位概念にも同様に上位概念を定義していき，概念の上位下位関係を表す木構造を作成した．これにより，Misra らが定義した 521 件の概念に対し，新しく 454 件の上位概念が定義された．また，作成された木構造の高さは

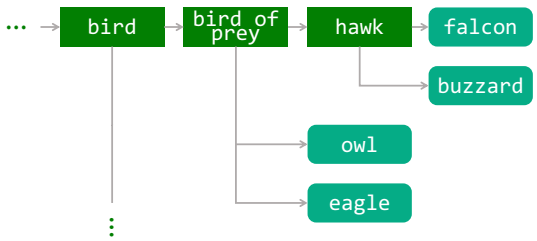

図 2　上位下位関係の木構造

17，上位概念に対する下位概念数は平均 2.04，最大 15 となった．その一部を図 2 に示す．

上位概念の性質

本実験では上位概念の性質を定義する必要があるが，上位概念は上記の方法で新しく定義された概念であることから，Misra らが作成した概念の性質を記述したデータセットではそれらの性質を定義することができない．

ここで，演繹推論の定義から，**ある性質について，下位概念での真偽が全て一致する場合，上位概念での真偽も一致する**ことを導き出せる (付録 D) ことを利用し，ある上位概念 h の全ての下位概念 c がある性質 p を持つ (持たない) 場合，その上位概念 h も性質 p を持つ (持たない) ものとして定義した．

生成したデータ

上記で作成した上位下位関係と上位概念の性質は組み合わせで生成されるため全体数が大きく，また真偽によってデータ数に大きな偏りがある．そのため，上記の方法で生成したデータの一部を次のようにサンプリングすることにより，上位下位関係の真偽の偏りが小さく，かつ全てのデータが他のいずれかのデータと組み合わせることで演繹推論における前提を構成することができるような事前学習データセットを作成した．

- 以下の手順を 10 k 回繰り返す．ただし，サンプリングは全て一様分布に従う．

1. “c is a h” の真偽 l_{ch} を {True, False} からサンプル
2. 性質 p，上位概念 h を 1 件ずつサンプル
3. (“c is a h”, l_{ch}) となる概念 c を 1 件サンプル
4. (“c is a h”, l_{ch}), (“h p”, l_{hp}) の 2 件をデータに追加 (ただし，l_{hp} は上位概念の性質の真偽)

- そして最後に重複したデータを削除した．

結果，生成されたデータは，(“c is a h”, True) が 1,576 件，(“c is a h”, False) が 9,740 件，(“h p”, True)

表 1　事前学習の結果

LM	終了時 epoch	正解率 (%)
ALBERT-xxl	17	99.70
RoBERTa-large	17	99.74
BERT-large	32	99.92

が 3,973 件，(“*h p*”, False) が 14,791 件の合計 30,080 件となった．また，このデータを 8:1:1 の割合で分割し，訓練，検証，およびテスト用データとした．

4.2　実験設定

上記の方法で生成した訓練データを用いて LM の訓練を行った．なお，本実験では前提の真偽を LM に記憶させることが目的であるため，訓練時の loss を基準に Early Stopping を行った．より具体的には訓練の終了条件を，訓練時の loss が 3 回連続で最良値を更新できなかった場合に終了とした．

4.3　結果

訓練した 3 種類の LM のそれぞれについて，訓練終了時の epoch 数と訓練データに対する正解率を表 1 に示す．結果，全ての LM について正解率が 99% を超えていることから，それぞれの LM は与えられた前提の真偽を概ね正確にパラメータ中に保存することができたと言える．

5　結論の予測

この工程では，LM が保存した前提の情報をもとに結論を予測する能力の調査を行った．ここで調査方法の一つとして，事前学習時と同様に Misra らの作成した概念の性質のデータセット (例：(“sparrow can grow”, True)) に対する正解率を予測能力の指標として評価する方法が考えられる．しかしながらこの方法では，LM には訓練時に上位概念の性質のデータ (例：(“bird can grow”, True)) を用いた訓練を行なっていることから，概念-上位概念間の埋め込みの類似度によりショートカットが発生することが考えられる．

そのため本実験では，図 1 (b) に示すように，訓練時に用いられない，ランダム初期化された埋め込みを持つ新概念 hoge を用いた調査を行った．全ての上位概念 *h* に対して，新概念の上位下位関係 “hoge is a *h*” の有無を記述したデータ 1 件を LM が正しく予測できるようになるまで訓練し (平均 2.31 epochs)，その後，新概念の性質 “hoge *p*” の有無を予測した．そして，その予測が新概念の上位下位関係の訓練の前後でどの程度変化したかを調査した．

表 2　新概念の性質の予測結果

LM	再現率 (%)	
	訓練前	訓練後
ALBERT-xxl	55.00	**75.59**
RoBERTa-large	45.51	**67.99**
BERT-large	60.01	**81.99**

ここで，LM が新しい上位下位関係のデータの学習において演繹的な振る舞いをすると仮定した場合，新概念の性質は上位概念の性質と真偽の予測が一致するようになることが考えられる．そのため，“hoge *p*” に対応する正解ラベル (真偽) *l* を上位概念の性質 “*h p*” の真偽と同じ値として定義した．

なお，前提による事前学習において，LM は概念の性質の有無についてデータ数に偏りがあるデータセットによって訓練されているため，*l* が偽の場合，正解率が高くなる傾向がある．そのため本実験における評価指標には再現率を用いるものとした．

5.1　結果

訓練した 3 種類の LM のそれぞれについて，新概念の性質を予測した結果を表 2 に示す．

結果，全ての LM について訓練後に再現率が増加していることが確認できた．したがって LM が事前学習した前提の知識をもとに，新概念の性質を対応する上位概念の性質から**演繹的に汎化する傾向がある**ことが示せた．

6　おわりに

本研究では，LM の推論能力の調査を行う足掛かりとして，事前学習した知識に基づく LM の推論能力の調査を行い，調査対象の LM が推論のショートカットの発生を抑制する実験設定においても演繹的に振る舞う傾向にあることを示した．

また，本研究の今後の課題としては，パラメータがランダムに初期化された LM を対象とするなど，事前学習した前提以外の影響がより少ない状態での調査を行うことや，事前学習後にアブダクションなどの断定的でない推論の影響がどの程度生じているかといった調査を行うことを考えている．また，訓練の過程における新概念の埋め込みの変化を観察するなど，本研究で示した演繹的な振る舞いが発生した要因についても調査を進めていく所存である．

謝辞

本研究は JSPS 科研費 21K17814 および JST, CREST, JPMJCR20D2 の助成を受けたものです.

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[2] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. In Yoshua Bengio and Yann LeCun, editors, **4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings**, 2016.

[3] Minjoon Seo, Sewon Min, Ali Farhadi, and Hannaneh Hajishirzi. Query-reduction networks for question answering. In **International Conference on Learning Representations**, 2017.

[4] Hung Le, Truyen Tran, and Svetha Venkatesh. Self-attentive associative memory. In **International Conference on Machine Learning**, pp. 5682–5691. PMLR, 2020.

[5] Soumya Sanyal, Harman Singh, and Xiang Ren. FaiRR: Faithful and robust deductive reasoning over natural language. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1075–1093, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In **Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence**, IJCAI'20, 2021.

[7] Zhangdie Yuan, Songbo Hu, Ivan Vulic, Anna Korhonen, and Zaiqiao Meng. Can pretrained language models (yet) reason deductively? **ArXiv**, Vol. abs/2210.06442, , 2022.

[8] Kanishka Misra, Julia Taylor Rayz, and Allyson Ettinger. A property induction framework for neural language models. **ArXiv**, Vol. abs/2205.06910, , 2022.

[9] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. **ArXiv**, Vol. abs/1909.11942, , 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] 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. **ArXiv**, Vol. abs/1907.11692, , 2019.

[12] George A. Miller. Wordnet: A lexical database for english. **Commun. ACM**, Vol. 38, No. 11, p. 39–41, nov 1995.

[13] Marco Baroni, Stefan Evert, and Alessandro Lenci. Esslli 2008 workshop on distributional lexical semantics. **Hamburg, Germany: Association for Logic, Language and Information**, 2008.

表 3　論理推論

推論過程	前提	結論	断定的
演繹推論 (Deduction)	「カラスは鳥である」「全ての鳥は卵を産む」	「カラスは卵を産む」	Yes
帰納推論 (Induction)	「カラスは飛べる」「カラスは鳥である」	「全ての鳥は飛べる (かもしれない)」	No
アブダクション (Abduction)	「全ての鳥は卵を産む」「カンガルーは卵を産む」	「カンガルーは鳥である (かもしれない)」	No

A　論理推論

3 種類の論理推論について，それぞれの具体例および結論が断定的に定まるかどうかを表 3 に示す．

B　LM が持つ知識の変化

本研究の実験において調査した LM はいずれもマスク穴埋め問題によって事前学習されたモデルであるため，そのパラメータ中には正解のトークンを予測するため「知識」を保存していることが考えられる．しかしながら，本研究の実験において LM が推論によって得る結論は与えた前提のみに基づくことを想定しているため，その「知識」は推論の結果に大きく影響しないことを前提としている．

そこで，事前学習データセットに含まれるデータの最後の単語を `[MASK]` に置き換えたマスク穴埋め問題を作成し，LM が予測するトークンの確率分布が (i) 事前学習の前後 (ii) 新概念の上位下位関係の学習の前後でどの程度変化するかを，top-k の平均の一致率を用いて調査した．結果を表 4 に示す．ただし，(ii) は新概念の上位概念によって訓練後の LM のパラメータがそれぞれ異なるため，そのうちの 10 個をサンプリングした全体の平均を求めた．

その結果，事前学習の前後では，30,000 トークンのうち上位 1,000 件の予測結果を比較しても予測されるトークンの一致率が 9.69 % と，マスク穴埋め問題のための「知識」の多くが失われている可能性があることが確認できた．一方で，新概念の上位下位関係の学習は 1 つのデータを平均 2.31 epochs 学習させているだけにも関わらず，予測結果の上位 1,000 件のうち 24.09 % が変化しているなど，大きな影響が生じていることが確認できた．

表 4　知識の変化

比較対象	top-k の一致率 (%)			
	$k = 1$	10	100	1000
事前学習の前後	2.66	2.97	4.03	9.69
新概念追加の前後	63.80	71.41	74.02	75.91

C　概念の性質

Misra らが作成した概念の性質のデータセットは，ある概念についてそれが持っている性質 (“c p”, True) については網羅的に記述しているものの，ある概念についてそれが持っていない性質 (“c p”, False) については (“c p”, True) をもとに一定数をネガティブサンプリングしたものとなっているため網羅的には記述されていない．上位概念の性質は下位概念の性質によって決まるため，本実験では (“c p”, False) が網羅的に記述されている必要がある．そのため，データセットのうち (“c p”, True) だけを利用し，(“c p”, False) のデータをその補集合として再定義した．その結果，概念は 521 件，性質は 3,734 件，(“c p”, True) は 23,107 件，(“c p”, False) は $521 \times 3,734 - 23,107 = 1,922,307$ 件となった．

D　上位概念の性質の定義

演繹推論の過程は，概念を c，上位概念を h，性質を p とすると次のように表せる．

$$\begin{cases} \forall c \text{ “}c \text{ is a } h\text{”} \wedge \text{“}h\ p\text{”} \Rightarrow \text{“}c\ p\text{”} \\ \forall c \text{ “}c \text{ is a } h\text{”} \wedge \neg\text{“}h\ p\text{”} \Rightarrow \neg\text{“}c\ p\text{”} \end{cases}$$

また，これは次式と同値である．

$$\begin{cases} \forall c \text{ “}c \text{ is a } h\text{”} \wedge \neg\text{“}c\ p\text{”} \Rightarrow \neg\text{“}h\ p\text{”} \\ \forall c \text{ “}c \text{ is a } h\text{”} \wedge \text{“}c\ p\text{”} \Rightarrow \text{“}h\ p\text{”} \end{cases}$$

したがって，ある性質について下位概念の真偽が全て一致する場合，上位概念の性質の真偽も一致する．そのため，ある上位概念 h の全ての下位概念 c がある性質 p を持つ (持たない) 場合，その上位概念 h も性質 p を持つ (持たない) ものとして定義した．