

Sequence-to-sequence モデルを用いた 一対多関係知識の記憶とその取り出し

長澤春希¹ Benjamin Heinzerling^{2,1} 穀田一真¹ 乾健太郎^{1,2}

¹ 東北大学 ² 理化学研究所

{haruki.nagasawa.s8, kokuta.kazuma.r3}@dc.tohoku.ac.jp

benjamin.heinzerling@riken.jp kentaro.inui@tohoku.ac.jp

概要

ニューラル言語モデル (LM) による一対多関係知識の記憶と取り出しの能力について, sequence to sequence モデルを用いた実験的な検証を行う. 本研究では一対多関係知識の扱いを「個別の要素の記憶」と「記憶した要素の取り出し」に切り分け, それぞれにおける LM の振る舞いを調査する. そして実験により, LM に一対多関係知識を記憶させることは一定程度可能であるが, 「記憶した要素の取り出し」については, 単純な設定では達成されないという結果が得られたことを報告する.

1 はじめに

近年, 大規模なコーパスで事前学習を行ったニューラル言語モデル (LM) が様々なタスクにおいて高い水準の性能を発揮している. LM はこの事前学習において, 単語の使われ方や文法などの言語に関する基本的な知識を獲得しているとされる.

またこれと同時に, LM は事前学習で用いたコーパスから, エンティティやイベントなどの世界に関する知識もそのパラメータ内部に暗に獲得していることが明らかとなっている [1, 2]. 世界知識のうち, ある人物 (親) とその子どもや, 映画監督とその作品など, 特定の関係で結ばれたエンティティについての知識を**関係知識**と呼ぶ. ニューラル言語モデルは事前学習の過程でかなり広範な関係知識をテキスト集合から獲得している可能性があることが指摘されている.

こうした背景から LM を知識ベースと見なしてその能力を調べる研究や LM と既存知識ベースを統合する研究が広がっている (Language Models as Knowledge Bases) [1, 3, 4, 5, 6]. また, 事前学習済みの LM から知識を引き出すためのプロンプトについて



図 1: 一対多関係知識の記憶と操作のイメージ

の研究も盛んに行われている [7, 8]. しかしながら, これまでの先行研究における主な関心の対象は, 一対一対応の関係知識に留まっており, 多対多関係の知識については扱いが簡単でないということ以外明らかにされていない [1, 3]. 一対一関係知識は, 例えば「国と首都」のようなもので, ある主語と関係が与えられた時に, 対応する目的語が一意に定まるものを指しており, 多対多関係知識は「映画と出演者」のように, ある関係を考えるときに, 互に対応する主語と目的語が複数個存在するものを指す. そして世界に関する知識には, 上述の多対多関係の知識も多く存在することが指摘されている [3]. 従って, LM を知識ベースとして扱える可能性をより緻密に調査するために, LM が多対多関係知識をどの程度正確に記憶 (パラメータの内部に格納) できるのか, また記憶した知識の操作 (例えば, 多対多関係のインスタンスに属する集合の間の演算) がどの程度可能か, といった基本的な能力を明らかにする必要があると考えられる. 本研究ではその第一歩として, 一対多関係の知識を人工的に生成したデータセットを構築し, これに対する LM の記憶と操作の能力を調べた.

本研究で調査するのは図 1 に示すような記憶と操作の能力である. 「記憶」の能力とは, LM に〈親子〉関係のような一対多関係のインスタンスを個別に教示した場合に, 教示したインスタンスの情報が LM の内部に格納されているかを指す. 「操作」の能

力としては、例えば図1の「出力」にあるように、ある人物の〈子ども〉をそれぞれ個別に教示したLMから、その集合(〈子ども〉の集合)を取り出すことができるか、言い換えれば個別に記憶した要素を列挙して出力することが可能かといった能力を指す。

本研究では、統制したデータでLMの能力を調べることがまずは重要だと考えたため、Wikidata [9] から子どもが複数に存在する親とその子どもの名前を収集し、一対多関係知識のデータセットを人工的に作成した。本稿ではこのデータセットに対するLMの振る舞いを報告する。

2 関連研究

2.1 関係知識の記憶

Heinzerling と Inui は、言語モデルを知識ベースとして扱うための基本的な要件として次の2つを定義している。(i) 多数のエンティティを含む大量の世界知識を格納できること。(ii) 格納した世界知識を参照できること。

この要件を元に、LMが一対一関係知識をどれくらいの精度でどれほどの量格納することができるのかという点について、様々なエンティティ表現を検討しながら精緻に検証している。しかしながら一対多関係知識についてのLMの振る舞いは依然として明らかにされていない。

2.2 集合の扱い

本研究では複数の目的語を扱うことを考えているが、これは目的語の「集合」を扱うことと置き換えることが可能である。この集合の扱いを深層学習、ないしはTransformer [10] で取り組んでいる代表的な先行研究として、Deep Sets [11] や Set Transformer [12] が存在する。両者とも入力としての集合に主眼を置いたものになっており、順序普遍性や、任意の大きさの集合の扱いについての言及がなされている。本研究では出力としての集合に主眼を置いているが、順序普遍性などの性質は共通して重要な側面であると考えられる。

3 調査手法について

3.1 用語定義

本研究で登場する用語と意味を定義する。

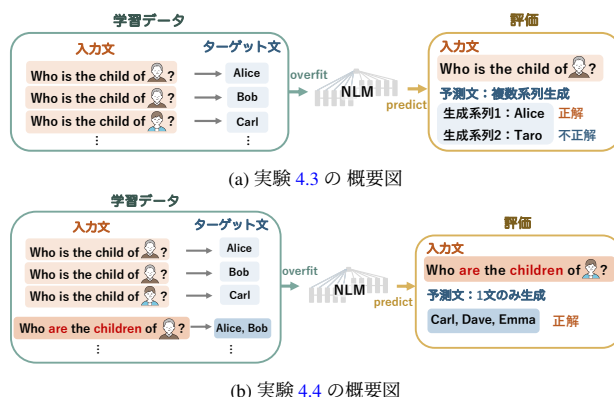


図2: 実験における学習データおよび評価についてのイメージ図

一対多関係知識 1つの主語に対し、特定の関係を満足する目的語が複数存在するような世界知識。

例: John Lennon has children named Sean Lennon and Julian Lennon.

個別要素 一対多関係知識において、主語と目的語を一対一対応の形式で結んだ個別の関係知識。

例: John Lennon has a child named Sean Lennon.

個別教示 一対多関係知識を構成するそれぞれの要素について、一対一対応の形式で個別に言語モデルに学習させること。

例(入力文): Who is the child of John Lennon?

例(ターゲット文): Sean Lennon.

集合教示 一対多関係知識全体を言語モデルに学習させること。

例(入力文): Who are the children named John Lennon?

例(ターゲット文): Sean Lennon, Julian Lennon.

3.2 生成モデルによる一対多関係知識の扱い

本研究では、事前学習済みLMに対し、一対多関係知識を明示的に学習させた際の振る舞いを調査する。具体的な調査方法については、言語モデルを用いた穴埋め形式と、生成モデルを用いた回答文生成形式の2つが考えられる(本研究では前者を**分類問題**、後者を**生成問題**と呼ぶこととする)。**分類問題**については、LMが出力する予測確率分布の上位N個の目的語を、一対多関係知識に対するモデルの回答として扱う。分類問題では、モデルに対し個別教示を実施することで複数ある目的語をほぼ全て正しく予測できることを明らかにし、言語処理学会第28回年次大会にて報告している [13]。本稿では、sequence-to-sequence [14] モデルを用いた生成手法によるアプローチ(**生成問題**)についての実験結果を報告する。生成問題では、個別教示を施したLMが個

別要素を正しく記憶できているかという点と、記憶した複数の個別要素を過不足なく列挙して出力することができるかという点について調査を行う。

4 実験

4.1 実験データ

本実験では LM の記憶精度を正確に測定することを目的に、学習データを独自に用意した。具体的には、Wikidata[9] から親とその子ども (2-4 人) の名前を取得し、使用する人物名について、次の条件を満たすような調整を行った。

- 主語・目的語の両方に渡り、同姓同名のエンティティが存在しない。
- 単語生成による記憶の難易度を調整するため、空白もしくはハイフン区切りで 4 単語以下で構成されるエンティティのみを使用。

なお本研究では、明示的に学習させた関係インスタンスの記憶と操作のみを考慮することとする。従って、学習データとして登場していない関係インスタンスに対する汎化能力などの測定は行わない。

4.2 実験設定

本実験では、sequence-to-sequence モデルとして事前学習済みの BART-base [15] および T5-base [16] を用いた。後述の 2 つの実験 (4.3 節, 4.4 節) において、訓練データに対して overfit するまで学習を継続した。具体的には、各主語についての入力に対し、目的語を 1 つ生成させ、それが正しい目的語のいずれかであれば正解、そうでない場合は不正解とし、この正解率が 5 エポック以上改善しなくなるまで学習を継続した。また最適化アルゴリズムとして AdamW [17] を採用した。

4.3 実験: 個別教示のみ実施

ここでは個別教示による学習のみを行い、LM が 1 つの主語に紐づく複数の目的語を正しく記憶できるかについての検証を行った。具体的には “{Subject} has a child named {mask}.” に対し、目的語を 1 つ生成させるという学習を全ての目的語について実施した。つまり、図 2a のように 1 つの入力文に対し、その主語に紐づく目的語数分のターゲット文が存在する学習設定となっている。

そして、個別教示を実施した LM が複数の目的語

をどの程度記憶できているかについて、beam search を用いて複数の系列を生成することにより確認した。具体的には 1 つの入力に対して、beam 幅をその主語の目的語数分に設定し、それぞれの生成系列群の中に正しい目的語がいくつ含まれているかを調査した¹⁾。また、モデルに記憶させるエンティティ数を変化させ、記憶できる目的語数が変化するかも併せて検証した。

図 3 はそれぞれ目的語数が 2, 3, 4 個の主語について、全ての目的語を漏れなく記憶できた主語数を示した結果である。

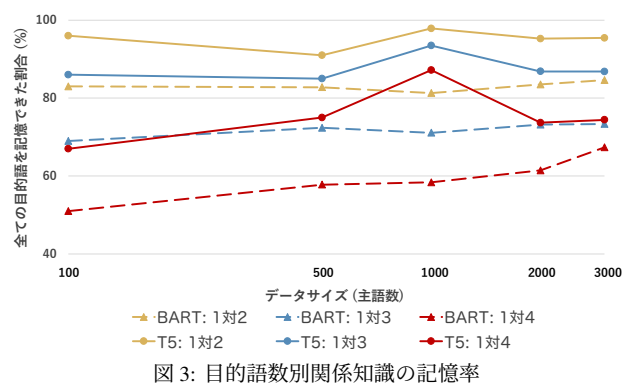


表 1: 1 対 4 関係知識について目的語を 3 つ以上記憶できている主語の割合 (%)

データサイズ	100	500	1000	2000	3000
BART	89.0	96.4	94.3	95.9	96.2
T5	93.0	98.6	99.3	97.7	97.9

実験より、BART・T5 の両モデルに関し、目的語数が少ないほど取りこぼし少なく記憶することが可能であることが明らかとなった。一方で 1 対 4 関係知識 (図 3 の赤線) に注目すると、4 つの目的語全てを記憶できている主語の割合は BART で概ね 5 割、T5 で概ね 7 割程度に留まることが分かった。これについて、各主語について 4 つ存在する目的語のうち 3 つ以上記憶できているものを含めた場合の記憶率を掲載したのが表 1 である。表より、両モデルとも 9 割以上の記憶率となっていることが分かる。上述の傾向はデータサイズ、すなわち記憶させるエンティティの総数に依存しない結果となった。

4.4 実験: 個別教示と集合教示

この実験では、「記憶した要素の取り出し」についての能力を獲得することを目的とし、個別教示に加えて集合教示を実施した。集合教示について具体的には “{Subject} has children named {mask}.” の入力

1) 実際の出力例を Appendix (A) に掲載

文に対し, “{Object1}, {Object2}, ...“ をターゲット文として設定し, 複数の目的語を列挙して出力する学習を行った. これにより, 入力文に “children“ が含まれる場合は, 記憶した正しい目的語を全て列挙する振る舞いをモデルに獲得させることを狙った. この集合教示を一部の主語について実施し, 明示的に集合教示を行っていない主語についても同様の出力が可能になるかといった汎化能力を検証した (本実験の概要を図 2b に示す).

学習方法については, (i) 個別教示を行った後に集合教示を行う方法と, (ii) 個別教示と集合教示を同時に実施する 2 通りが考えられる. 前者については, 集合教示が行われるにつれて個別教示に対する正解率が急激に下がってしまう事象 (破局的忘却) が確認された. 一方で後者については, 訓練データに集合教示のデータを含めても, LM の個別要素の記憶に対する振る舞いはほとんど影響受けないことが確認された. 従って本節では, 個別教示と集合教示を同時に行った場合についての結果を報告する.

ここでは 1 対 2 - 4 関係知識について, それぞれ 3000 件の主語を対象に学習を行った. 個別教示については全ての目的語に対して実施しながら, 集合教示の割合を変化させて集合出力に関するモデルの振る舞いを検証した. 具体的には全体の 30, 60, 90% の主語について集合教示を行った. 表 2 および表 3 はそれぞれ BART, T5 についての実験結果である. 正しい目的語を不足なく全て出力できている場合を正解, 出力漏れや誤った目的語が出力されている場合を不正解とし, 全主語に対する正答率を計算した. また表の下段の値は, それぞれ出力系列に含まれるエンティティ単位での recall・precision の平均値となっている.²⁾

表 2: BART: 集合教示割合と出力集合の正解率

集合教示割合	30%	60%	90%
1 対 2	47.2% (0.86, 0.81)	51.9% (0.84, 0.83)	63.3% (0.89, 0.90)
1 対 3	13.8% (0.69, 0.83)	15.8% (0.66, 0.85)	22.4% (0.71, 0.90)
1 対 4	9.76% (0.64, 0.83)	13.0% (0.62, 0.84)	20.6% (0.70, 0.88)

実験より BART と T5 のそれぞれで異なる振る舞いが観察された. BART については, 30% 集合教示における 1 対 2 関係知識の正答率を除き, 明示的に学習を行った主語についても複数の目的を正しく列

2) エンティティ単位での recall・precision の計算方法については Appendix B に記載.

表 3: T5: 集合教示割合と出力集合の正解率

集合教示割合	30%	60%	90%
1 対 2	46.7% (0.91, 0.78)	66.5% (0.93, 0.86)	93.7% (0.99, 0.97)
1 対 3	37.3% (0.82, 0.84)	60.9% (0.87, 0.89)	93.2% (0.98, 0.98)
1 対 4	31.0% (0.76, 0.87)	55% (0.84, 0.91)	90.8% (0.97, 0.98)

挙することが難しいという結果となった. 学習率やバッチサイズ等のハイパーパラメータの探索を行ったがいずれの場合においても同様の結果となった. T5 では基本的に明示的に集合教示を行った割合以上の正解率となっており, 「記憶した要素の取り出し」についての能力が多少汎化していると考えられるが, 本質的な能力が獲得されたとは考えにくい結果となった.

5 おわりに

本稿では sequence-to-sequence モデルを用いた生成アプローチによる一対多関係知識の扱いについて, 「個別の要素の記憶」と「記憶した要素の取り出し」に切り分けて調査を行った.

前者については beam search により複数系列を生成し, その中にいくつ正しい目的語が含まれているかという観点で検証を行った. 今回の学習方法では, 個別教示を行った LM が複数の目的語全てを漏れなく格納するのは困難である結果となった. N 個ある目的語を全て記憶できているエンティティと, 1 つしか記憶できないエンティティについてどのような差があるのかや, より良い学習方法の模索という点について今後緻密に検証する必要があると考えられる.

後者については, 個別教示と同時に集合教示を施すことで, LM に格納されている複数の目的語をまとめて出力する能力の獲得を図った. こちらについては同じ学習設定においてもモデルにより顕著に異なる振る舞いが確認された. しかしながら比較的学習がうまくいったと考えられる T5 においても「記憶した要素の取り出し」についての汎化能力は低い水準で留まる結果となった. 以上より, 各モデルに対するより適切な学習データの構築や学習方法の工夫の必要性が示唆された.

今回の実験結果を踏まえ, より正確に複数の目的語を LM に記憶させる方法と, 記憶した要素の引き出しについてどのような施策が必要になるかを引き続き調査していく所存である.

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] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019**, pp. 2463–2473. Association for Computational Linguistics, 2019.
- [2] Chenguang Wang, Xiao Liu, and Dawn Song. Language models are open knowledge graphs. **CoRR**, Vol. abs/2010.11967, , 2020.
- [3] Benjamin Heinzerling and Kentaro Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021**, pp. 1772–1791. Association for Computational Linguistics, 2021.
- [4] Gerhard Weikum, Xin Luna Dong, Simon Razniewski, and Fabian M. Suchanek. Machine knowledge: Creation and curation of comprehensive knowledge bases. **Found. Trends Databases**, Vol. 10, No. 2-4, pp. 108–490, 2021.
- [5] Tara Safavi and Danai Koutra. Relational world knowledge representation in contextual language models: A review. **CoRR**, Vol. abs/2104.05837, , 2021.
- [6] Simon Razniewski, Andrew Yates, Nora Kassner, and Gerhard Weikum. Language models as or for knowledge bases. **CoRR**, Vol. abs/2110.04888, , 2021.
- [7] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know. **Trans. Assoc. Comput. Linguistics**, Vol. 8, pp. 423–438, 2020.
- [8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual**, 2020.
- [9] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. **Communications of the ACM**, Vol. 57, No. 10, pp. 78–85, 2014.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, **Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA**, pp. 5998–6008, 2017.
- [11] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. **CoRR**, Vol. abs/1703.06114, , 2017.
- [12] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, **Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA**, Vol. 97 of **Proceedings of Machine Learning Research**, pp. 3744–3753. PMLR, 2019.
- [13] 長澤春希, Benjamin Heinzerling, 乾健太郎. ニューラル言語モデルによる一対多関係知識の記憶と操作. 言語処理学会 第 28 回年次大会 発表論文集, pp. 1203–1208, 2022.
- [14] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, **Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada**, pp. 3104–3112, 2014.
- [15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020**, pp. 7871–7880. Association for Computational Linguistics, 2020.
- [16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **CoRR**, Vol. abs/1910.10683, , 2019.
- [17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**. OpenReview.net, 2019.

A 個別教示を実施したモデルの実際の beam search 出力例

個別教示を実施した BART, T5 の両モデルについて, beam search を用いた実際の出力系列を以下の表に掲載する. BART に対する入力文は 1 対 2 関係知識に関するものであるため, 正解となる目的語は 2 つ存在する. 一方で T5 に対する入力文は 1 対 4 関係知識となっているため, 正解の目的語は 4 つ存在する.

それぞれについて, 目的語数分の系列を beam search により生成し, 正しい目的語が重複なく生成されているかを確認したものが実験 4.3 となっている.

表 4: 個別教示後の実際のモデル出力例 (上位 4 系列)

BART		
入力 (正答)	John Lennon has a child named <mask>. (Sean Lennon or Julian Lennon)	成否
出力系列 1	Sean Lennon	○
出力系列 2	Julian Lennon	○

T5		
入力 (正答)	What is the name of Pablo Picasso's child? (Paloma Ruiz Picasso or Claude Picasso or Maya Widmaier-Picasso or Paulo Picasso)	成否
出力系列 1	Paloma Ruiz Picasso	○
出力系列 2	Maya Widmaier-Picasso	○
出力系列 3	Paulo Picasso	○
出力系列 4	Augusta Widmaier-Picasso	×

B 集合生成における recall・precision の算出

ここでは, 表 2, 3 における recall・precision の算出方法について説明する. ある主語について, 正解の目的語集合を *gold_entities*, モデルが出力した目的語集合を *pred_entities* とし, $num(X)$ を集合の要素数を表すものとする. recall・precision はそれぞれ式 (1) および式 (2) で算出した. これを全ての出力について計算し, その平均を表 2, 3 に掲載している.

$$recall = \frac{num(pred_entities \cap gold_entities)}{num(gold_entities)} \quad (1)$$

$$precision = \frac{num(pred_entities \cap gold_entities)}{num(pred_entities)} \quad (2)$$