事前学習済み言語モデルからの訓練データ抽出：新聞記事の特性を用いた評価セットの構築と分析

石原祥太郎 [1]
[1] 株式会社日本経済新聞社
shotaro.ishihara@nex.nikkei.com

概要

大規模なデータセットを用いた事前学習済み言語モデルが数多くのタスクで高い性能を示している一方で，訓練データの一部が抽出可能であるセキュリティ面の課題が重要性を増している．本稿では新聞記事の特性を考慮することで，事前学習済み言語モデルからの訓練データ抽出の議論に向け，現実世界に即した評価セットが構築できると主張する．最初に先行研究を踏まえこの課題を整理し，日本語の新聞記事を用いた評価セットを提案する．次いで，新聞記事で独自にモデルを事前学習し，構築した評価セットに対して複数のモデルで訓練データ抽出を試みる．条件ごとの記憶の度合いの変化を確認する実験を通じて，この課題に関するいくつかの知見を示し，将来展望を述べる．

1 はじめに

自然言語処理の躍進を支える「事前学習済み言語モデル」は，研究領域のみならず産業界でも大きな注目を集めている．単語の並び方に確率を割り当てる統計的な言語モデルは古くから研究が進み，近年は大規模なニューラルネットワークに対する巨大なデータセットでの事前学習が採用されている．この拡張は流暢な自然言語生成に繋がり，多くの下流タスクにファインチューニングすることで高い性能を示すと報告された [1]．時には適切な文字列（プロンプト）を与えると，パラメータ更新なしに下流タスクに適した出力が可能で [2, 3]，GitHub Copilot [1)] や ChatGPT [2)] などの社会実装も着実に進行している．

実用化が進むと，セキュリティ面の課題が顕在化する [4, 5]. 先行研究は，ニューラルネットワークが訓練に用いたデータを意図せず記憶し出力する性質を持つと指摘している [6, 7, 8]．Carlini ら [7] は，事前学習済み言語モデルから多数の文章を生成（デコーディング）しメンバーシップ推論 [9] をすることで，記憶された数多くの情報を抽出できると確認した．この課題はプライバシーの侵害・実用性の低下・公平性の低下などを招くため，事前学習済み言語モデルの実運用に向けた課題となり得る [10]. 特に医療・臨床データなど機密性の高い情報で事前学習したモデルを扱う場合，訓練データの漏洩は大きな問題に繋がる可能性がある [11]．記述内容を記憶している点で「忘れられる権利」の文脈でも議論が必要である [12].

一方で，この課題に関する研究領域や産業界での議論は未だ発展途上である．多くの研究の実験は定性的な側面にとどまり，いくつかの論文で定量評価の重要性が提起され始めている [10, 13, 14]．既存の研究では，対象とするモデルの事前学習に用いたデータセットから一部のデータを取り出す方法で，評価セットを構築している．たとえば Carlini ら [10] は主に GPT-Neo モデル群 [15] を題材に，ウェブスクレイピングなどで収集したテキストを含む合計 825 GB の Pile データセット [16] を利用し，約 5 万の部分集合を抽出した．部分集合の選択は推論の計算量の都合から避けられない処理だが，公開済みの情報を恣意性なく加工し，訓練データ抽出の現実世界のシナリオを想定した評価セットを構築するのは難しい作業である．

本稿では新聞記事の特性を考慮することで，事前学習済み言語モデルからの訓練データ抽出に関する課題を議論するため，現実世界に即した評価セットが構築できると提案する．有料の購読プランを提供するインターネット上のニュース配信サービスでは，新聞記事は冒頭の数百字のみが一般公開され，続く文章は購読者のみ閲覧可能となっている場合が多い．そのため，公開部分を事前学習済み言語モデルに与え，非公開部分を引き出すという状況を手軽

1） https://github.com/features/copilot
2） https://openai.com/blog/chatgpt/

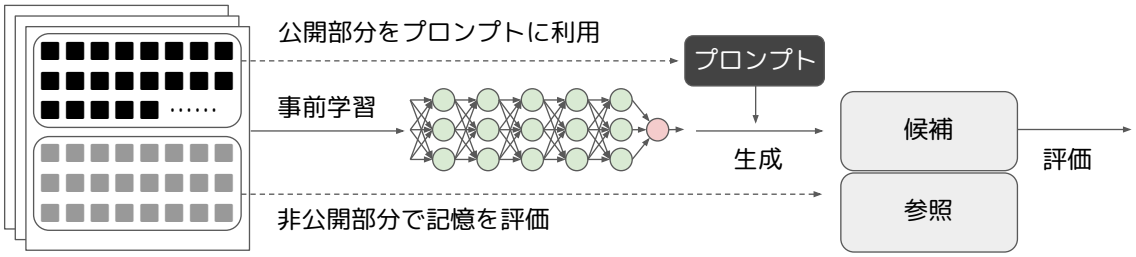

図 1　本研究で扱う訓練データ抽出の概要．新聞記事を用いた事前学習済み言語モデルに対して，公開部分をプロンプトと見なして文字列をデコーディングする．記憶の度合いは，非公開部分を用いて評価する．

に再現できる利点がある．以降，本研究で取り組む課題を整理し（2 節），日本語の新聞記事を用いた評価セットを構築する（3 節）．次いで実験を通じて，日本語を題材としたこの課題に関するいくつかの知見を示す（4 節）．具体的には，独自に事前学習したモデルを含む複数のモデルで訓練データ抽出を試みて，条件ごとの記憶の度合いの変化を確認した．最後に本稿のまとめと将来展望を述べる（5 節）．

2　問題設定

本節では先行研究をもとに，事前学習済み言語モデルによる訓練データの「記憶」を定義し，本研究で取り組む訓練データ抽出について説明する．

2.1　「記憶」の定義

事前学習済み言語モデルの「記憶」とは，訓練データに関する情報を蓄え，時にそのまま出力する現象を指す．記憶に関する研究は多岐にわたり，定義や前提条件も多様であるが，本稿では GPT 系のモデル群 [1, 2, 3, 15] に代表される自己回帰型の言語モデルに焦点を当てる．

本研究では，2 種類の記憶の定義を用いる．まず多くの先行研究 [7, 10, 13] を参考に，文字列の部分一致に基づく定義を採用する（Definition 2.1）．この「逐語記憶（verbatim memorization）」の定義は，事前学習済み言語モデルに対して適切なプロンプトを与えることで，記憶したデータが出力されると仮定している．加えて，文字列の類似性を考慮した「近似記憶（approximate memorization）」の定義 [14, 17] も用いる（Definition 2.2）．類似度として Lee ら [14] はトークン単位の一致率，Ippolito ら [17] は BLEU [18] を利用した．本研究では単語間に空白がない日本語を扱う点を考慮し，文字単位の編集距離 [19] を採用する．差分プライバシ [20] や反実仮想 [21] の考え方に基づく定義もあるが，本研究の対象外とした．

Definition 2.1 (逐語記憶)**.** 長さ k のプロンプト p に対し，ある文字列 s が事前学習済み言語モデル f_θ から生成され，p と s を結合した文字列が訓練に用いたデータセット内に含まれている場合，s は f_θ に記憶されている．

Definition 2.2 (近似記憶)**.** 長さ k のプロンプト p に対し，ある文字列 g が事前学習済み言語モデル f_θ から生成され，g と訓練に用いたデータセット内に含まれている文字列 s がある類似度の条件を満たす場合，s は f_θ に近似記憶されている．

2.2　訓練データ抽出

本研究では先行研究で多く参照されている Carlini ら [7] と同様の手順で，事前学習済み言語モデルからの訓練データ抽出を試みる（図 1）．新聞記事は一般公開されている冒頭の数百字（公開部分）と，購読者のみ閲覧可能の続き（非公開部分）で構成される．最初に準備のため，公開部分と非公開部分を合わせた全ての文章でモデルを事前学習する．事前学習するモデルも Carlini ら [7] と同じで，主に GPT-2 [2] を利用する．この事前学習済み言語モデルに対して，公開部分の文字列をプロンプトと見なし，続く文字列をデコーディングする．デコーディング戦略の選択は実験結果に大きな影響を与えないとする報告もあり [10]，ここでは毎回最大の条件付き確率の単語を生成する貪欲法を用い，一つのプロンプトから一つの文字列を生成する．生成された文字列を非公開部分と比較すると，記憶の度合いを定義に基づき評価できる．

3　評価セットの構築

本節では，新聞記事を用いた評価セットを構築する．本研究では日本経済新聞社のニュース配信サービス「日経電子版」[3] のデータセット [4] を用いた．このデータセットでは例外を除き「冒頭 200 文字」か

3)　https://www.nikkei.com/
4)　https://aws.amazon.com/marketplace/pp/prodview-7hhnlazdw5rsi

種類	文字列	逐語記憶	近似記憶
公開部分	(前略…) 年明け以降の新型コロナウイルスの新規感染者数が大幅に増加するとの懸念が一定の重荷になっている。	-	-
非公開部分	前引け後の東証の立会外で、国内外の大口投資家が複数の銘柄をまとめて売買する「バスケット取引」は約 65 億円成立した。	-	-
nikkei/gpt2-1epoch	JPX 日経インデックス 400 と東証株価指数 (TOPIX) も下落している。	0	0.052632
nikkei/gpt2-5epoch	市場からは「きょうは 2 万 9000 円〜2 万 9000 円の範囲で、この水準を上抜けるには戻り待ちの売りが出やすい」(国内証券ストラテジスト) との声があった。	0	0.093333
nikkei/gpt2-15epoch	前引け後の東証の立会外で、国内外の大口投資家が複数の銘柄をまとめて売買する「バスケット取引」は約 396 億円成立した。	48	0.948276
rinna/japanese-gpt-1b	</s>	0	0.000000

表 1　nikkei/gpt2-60epoch で最も逐語記憶の度合いが高かった評価セット（公開・非公開部分）と生成結果.

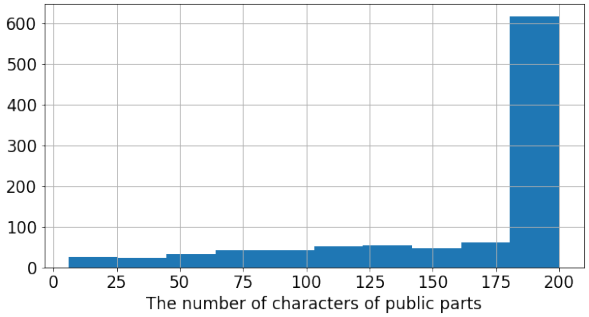

図 2　評価セット 1000 件の公開部分の文字数のヒストグラム．200 文字程度が大半だが，短い記事も存在する．

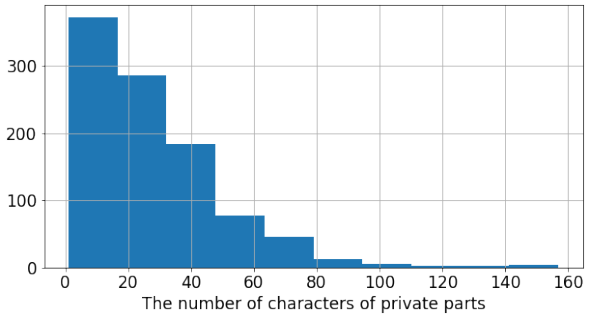

図 3　評価セット 1000 件の非公開部分のうち，最初の文末までの文字数のヒストグラム．9 件が 200 文字を超えたため可視化の際には省略した．

「記事全体の文字数の半分」の短い方が，公開部分として定義されている[5]．本研究では 2021 年に公開された中で，例外に該当する記事を除いて 1000 記事を評価セットとして利用した（図 2）．非公開部分は記事によっては非常に長く，問題の単純化のために最初の文末までを抽出した[6]（図 3）．表 1 のように公開部分が句点で終わるのは，25 記事と少数派である．Appendix A に示す通り，公開部分は必ずしも句点で終わらず非公開部分の冒頭と接続する．

5)　ニュースの公共性などさまざまな事情に応じて非公開部分も含めた記事全体を公開している場合がある．
6)　実装は bunkai（https://github.com/megagonlabs/bunkai）を用いた．

4　実験

本節では，さまざまな条件下での実験を通じ，提案した評価セットから得られた知見を報告する．最初に複数の事前学習済み言語モデルを準備し，次に訓練データ抽出を試みて条件ごとの記憶の度合いの変化を分析した．

4.1　利用した事前学習済み言語モデル

まず 2010〜21 年に公開された日経電子版の記事本文（2.8 GB）を用いて GPT-2 を事前学習し，複数の学習エポック数でモデルを保存した．評価セットの記事も訓練に用いたデータセットに含まれている．モデルの一覧は表 2 で確認でき，nikkei/gpt2-{X}epoch は日経電子版のデータセットで X エポック学習したモデルである．事前学習には Hugging Face の Transformers [22]，単語の分割にはユニグラム言語モデル [23] を用いた．事前学習の細かい設定については Appendix B に示す．

記憶の性質とは関係なく生成した文字列が偶然一致する可能性があるため，別のデータセットで事前学習されたモデルも比較対象とした．表 2 のモデル名は Hugging Face での公開名[7]で，パラメタ数が異なるモデルを選んだ．それぞれ日本語の Wikipedia[8] や CC-100[9] などで事前学習されている．

4.2　訓練データ抽出

準備したモデルを用いて，2.2 節で示した方法で訓練データ抽出を試みた．生成された結果は，2.1 節での定義に基づき，大きいほど記憶の度合いが大きい値として次のように定量化した．文字列は全て

7)　https://huggingface.co/models
8)　https://meta.wikimedia.org/wiki/Data_dumps
9)　https://data.statmt.org/cc-100/

モデル名	パラメタ数	逐語記憶	逐語記憶	近似記憶	近似記憶
集約方法	-	最大値	平均値	平均値	中央値
nikkei/gpt2-1epoch	0.1B	25	0.560	0.190537	0.120345
nikkei/gpt2-5epoch	0.1B	25	0.839	0.229408	0.142857
nikkei/gpt2-15epoch	0.1B	**48**	0.788	0.236079	0.142857
nikkei/gpt2-30epoch	0.1B	**48**	**0.948**	**0.241923**	**0.149627**
nikkei/gpt2-60epoch	0.1B	**48**	0.874	0.238184	0.145833
rinna/japanese-gpt2-small	0.1B	12	0.580	0.181397	0.115385
rinna/japanese-gpt2-medium	0.3B	15	0.657	0.205017	0.129032
abeja/gpt2-large-japanese	0.7B	19	0.760	0.210954	0.136364
rinna/japanese-gpt-1b	1.3B	18	0.882	0.219001	0.142857

表 2　各モデルの記憶の度合い．パラメタ数の B は Billion（10 億），太字は最も類似度が高いことを意味する．

半角に揃えて比較した．

- 逐語記憶：前方一致の文字数
- 近似記憶：1 -（編集距離 / 文字列の長さ[10]）

最も長く事前学習した nikkei/gpt2-60epoch で逐語記憶の度合いが 1 番高かった評価セットに対し，一部のモデルの生成例を表 1 に示した．非公開部分との前方一致は緑色で強調した．日経電子版のデータセットで事前学習したモデルは 15 エポックを超えた段階で，非公開部分と 48 文字前方一致する文字列を生成した．一方 rinna/japanese-gpt-1b は，公開部分の末尾が句点のためか，文末を示す特別なトークン </s> を出力した．2 番目に逐語記憶の度合いが高かった評価セットの生成結果は Appendix A で紹介する．

各モデルの記憶の度合いの集約結果を表 2 に示す．まずエポック数が増加するにつれ逐語記憶の最大値も増えている．同一データセットでの繰り返しの学習を通じ，事前学習済み言語モデルの逐語記憶の性質が強まっていると示唆された．日経電子版のデータセットを用いた事前学習言語済みモデルは，その他のモデルより長い文字列を逐語記憶している．逐語記憶の平均値でも，概ね同様の傾向が見られた．近似記憶の平均値・中央値もエポック数の増加と相関の傾向がある．その他のモデルと比べると，1 エポックのみの事前学習では小さい値となったが，事前学習を重ねるとより大きな値を示した．

その他のモデルでは，パラメタ数が大きいほど構築した評価セットでの記憶の性質が高まる傾向にあった．先行研究 [10] でも報告されている通り，パラメタ数が大きいほど一般的な記憶の性質が高まるためだと推察される．

10）　比較する 2 つの文字列の長い方

5　おわりに

本稿では公開と非公開の部分が共存する新聞記事の特性を考慮し，事前学習済み言語モデルからの訓練データ抽出に関する課題を議論するための，現実世界に即した評価セットの構築手法を提案した．日本語の新聞記事を用いた評価セットを構築した後，同じデータセットで GPT-2 を事前学習し，複数のモデルで訓練データ抽出を試みる実験を通じて，提案した評価セットから得られた知見を報告した．

事前学習済み言語モデルの実用化が進む中，この課題に関する議論は重要性を増している．英語以外の言語特有の論点を特定し検証する目的でも，日本語のデータセットを用いた研究には大きな意義がある．将来展望として，まずデータセットの構築や評価の方法の議論を深める必要がある．本研究では評価セットとして 1000 件ランダムに抽出したが，より適切な条件に絞る方法も考えられる．先行研究 [10] は記憶の性質が文字列の重複と関係するという知見 [13, 14] に基づき評価セットを構築した．記憶の定義や評価指標も検討の余地がある．先行研究 [7] はパラメタ数が異なる GPT-2 での生成確率の違いなど，メンバーシップ推論の 6 つの指標を利用した．

評価の枠組みを確立することで，訓練データ抽出に関連する手法の影響を検証できる．本研究では貪欲法を用いて一つのプロンプトから一つの文字列を生成したが，先行研究 [7, 13, 14] は top-k サンプリングといったデコーディング戦略の工夫で，多様な文字列の候補を大量に列挙した．課題の対応策としては，データの削除や重複の排除 [13, 14]，差分プライバシを導入した学習時の工夫 [24]，危険な出力を特定し除外する後処理 [25] などが提案されている．

参考文献

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, et al. Improving language understanding by generative pre-training, 2018.

[2] Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[3] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are Few-Shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, et al. On the dangers of stochastic parrots: Can language models be too big? In **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**, pp. 610–623, New York, NY, USA, March 2021. Association for Computing Machinery.

[5] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, et al. Taxonomy of risks posed by language models. In **Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency**, pp. 214–229, New York, NY, USA, June 2022. Association for Computing Machinery.

[6] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, et al. The secret sharer: Evaluating and testing unintended memorization in neural networks. In **28th USENIX Security Symposium (USENIX Security 19)**, pp. 267–284, 2019.

[7] Nicholas Carlini, Florian Tramèr, Eric Wallace, et al. Extracting training data from large language models. In **30th USENIX Security Symposium (USENIX Security 21)**, pp. 2633–2650, 2021.

[8] Huseyin A Inan, Osman Ramadan, Lukas Wutschitz, et al. Training data leakage analysis in language models. In **3rd Privacy-Preserving Machine Learning Workshop**, January 2021.

[9] Reza Shokri, Marco Stronati, Congzheng Song, et al. Membership inference attacks against machine learning models. In **2017 IEEE Symposium on Security and Privacy (SP)**, pp. 3–18, 2017.

[10] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, et al. Quantifying memorization across neural language models. **arXiv preprint arXiv:2202.07646**, February 2022.

[11] Eric Lehman, Sarthak Jain, Karl Pichotta, et al. Does BERT pretrained on clinical notes reveal sensitive data? In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 946–959, Online, June 2021. Association for Computational Linguistics.

[12] Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. Formalizing data deletion in the context of the right to be forgotten. In **Advances in Cryptology – EUROCRYPT 2020**, pp. 373–402. Springer International Publishing, 2020.

[13] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In **Proceedings of the 39th International Conference on Machine Learning**, Vol. 162, pp. 10697–10707. PMLR, 17–23 Jul 2022.

[14] Katherine Lee, Daphne Ippolito, Andrew Nystrom, et al. Deduplicating training data makes language models better. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[15] Sidney Black, Stella Biderman, Eric Hallahan, et al. GPT-NeoX-20B: An Open-Source autoregressive language model. In **Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models**, pp. 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics.

[16] Leo Gao, Stella Biderman, Sid Black, et al. The pile: An 800GB dataset of diverse text for language modeling. **arXiv preprint arXiv:2101.00027**, December 2020.

[17] Daphne Ippolito, Florian Tramèr, Milad Nasr, et al. Preventing verbatim memorization in language models gives a false sense of privacy. **arXiv preprint arXiv:2210.17546**, October 2022.

[18] Kishore Papineni, Salim Roukos, Todd Ward, et al. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[19] Li Yujian and Liu Bo. A normalized levenshtein distance metric. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 29, No. 6, pp. 1091–1095, June 2007.

[20] Cynthia Dwork, Frank McSherry, Kobbi Nissim, et al. Calibrating noise to sensitivity in private data analysis. In **Theory of Cryptography**, pp. 265–284. Springer Berlin Heidelberg, 2006.

[21] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, et al. Counterfactual memorization in neural language models. **arXiv preprint arXiv:2112.12938**, December 2021.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, et al. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[23] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[24] Xuechen Li, Florian Tramèr, Percy Liang, et al. Large language models can be strong differentially private learners. **arXiv preprint arXiv:2110.05679**, October 2021.

[25] Ethan Perez, Saffron Huang, Francis Song, et al. Red teaming language models with language models. **arXiv preprint arXiv:2202.03286**, February 2022.

種類	文字列	逐語記憶	近似記憶
公開部分	（前略…）日本政府は 4 月、30 年度に温暖化ガス排出を 13 年度比 46 %減らす目標を打ち出した。秋に開かれた第 26	-	-
非公開部分	回国連気候変動枠組み条約締約国会議（COP26）では、「世界の平均気温の上昇を 1.5 度に抑える努力を追求することを決意する」ことで合意した。	-	-
nikkei/gpt2-1epoch	回国連気候変動枠組み条約締約国会議 (COP26) で、脱炭素に向けた投資や脱炭素の戦略を練り直す。	25	0.414286
nikkei/gpt2-5epoch	回国連気候変動枠組み条約締約国会議 (COP26) で も、企業の対応が注目されそうだ。	25	0.400000
nikkei/gpt2-15epoch	回国連気候変動枠組み条約締約国会議 (COP26) では、50 年の実質ゼロに向けた道筋を議論。	27	0.442857
nikkei/gpt2-30epoch	回国連気候変動枠組み条約締約国会議 (COP26) では、30 年目標の前倒しが議論された。	27	0.428571
nikkei/gpt2-60epoch	回国連気候変動枠組み条約締約国会議 (COP26) では、各国が脱炭素に向けた行動計画を策定する。	27	0.457143
rinna/japanese-gpt2-small	回 気候変動枠組条約締約国会議 (cop24) では、cop24 で排出削減目標が達成された企業を「排出削減企業」として認定した。	1	0.357143
rinna/japanese-gpt2-medium	回 気候変動枠組条約締約国会議 (cop24) で、cop21 の目標達成に向けた具体的な行動計画の策定が合意された。	1	0.342857
abeja/gpt2-large-japanese	回 先進国首脳会議 (伊勢志摩サミット) で、日本は「2030 年目標」を公表した。	1	0.114286
rinna/japanese-gpt-1b	回 気候変動枠組条約締約国会議 (COP26) では、パリ協定の実施指針となる「パリ協定実施指針」が採択された。	1	0.414286

表 3　nikkei/gpt2-60epoch で 2 番目に逐語記憶の度合いが高かった評価セット（公開・非公開部分）と生成結果.

A　評価セットと生成の例

nikkei/gpt2-60epoch で 2 番目に逐語記憶の度合いが高かった評価セットと，実験で用いた各モデルの生成結果を表 3 に示す．表 1 とは，公開部分が句点の前に終わっている点が異なる．大まかな傾向は同じで，日経電子版のデータセットで事前学習したモデルでは，エポック数の増加に伴い逐語記憶・近似記憶の度合いが高まった．1 エポックのみの事前学習で「第 26 回国連気候変動枠組み条約締約国会議（COP26）」という文字列が生成されている．

その他の事前学習済み言語モデルでは，逐語記憶・近似記憶の度合いが小さい．生成結果に文法的な誤りはなく違和感は少ないが，表 2 に示したパラメタ数が比較的小さい rinna/japanese-gpt2-small や rinna/japanese-gpt2-medium には「第 26」を与えた文脈では間違いとなる「cop24」「cop21」などの略称が含まれた．パラメタ数が比較的大きい rinna/japanese-gpt-1b では正しい「COP26」の略称になっており，一般的な記憶の性質が高まっているが，非公開部分と比べて「国連」が欠落した．abeja/gpt2-large-japanese は，非公開部分と異なり先進国首脳会議に言及した．なお三重県伊勢志摩で開催されたのは第 42 回で，第 26 回は沖縄県名護市だった．表 1 と 3 の公開・非公開部分はそれぞれ日経電子版の記事[11) 12)]から引用した．

B　事前学習の設定

事前学習は Transformers のドキュメント[13)]を参考に，ハイパーパラメタを設定した．具体的には学習率を 0.005，バッチサイズを 8，weight decay を 0.01，最適化アルゴリズムを Adafactor とした．Transformers と TensorFlow のバージョンは 4.11 と 2.5 を用いた．計算資源には，GPU の A100 を 8 個含む Amazon EC2 P4 Instances の ml.p4d.24xlarge[14)]を使った．単語の分割には，語彙数 32000 のユニグラム言語モデルを用いた．テキストから直接語彙を生成できるため，日本語や中国語のように，単語間に明示的な空白がない言語に対して有効である．

11）https://www.nikkei.com/article/DGXZASS0ISS14_Q1A231C2000000
12）https://www.nikkei.com/article/DGKKZO78866030Y1A221C2DTA000
13）https://github.com/huggingface/transformers/tree/main/examples/flax/language-modeling
14）https://aws.amazon.com/ec2/instance-types/p4/