自然言語処理における属性単位での反学習

沖村 樹 [1] 小島 武 [1] 岩澤 有祐 [1] 松尾 豊 [1]
[1] 東京大学
okimura@weblab.t.u-tokyo.ac.jp

概要

機械反学習（Machine Unlearning）は有害なデータセットからの影響を軽減できる注目すべき手法である．特に自然言語処理領域では利用者の視点から，特定の個別情報を削除できるかに主に焦点が当てられている．しかし，モデル提供者の視点からは，反学習が要求に対して特定の個別情報にのみ適用される限りでは，モデルがリスクとなる情報を出力する可能性は排除できない．この問題に対処するため，本研究ではモデル提供者がリスクのある属性に関する情報を抑制するため，属性単位での反学習が汎化しているかを評価する新しいデータセットと問題設定を提案する．提案したデータセットを用いて，既存の手法が属性単位での反学習にどれだけ適応できるかを調査し，また，提案手法であるラビング（Rubbing）が代表的な手法と比較して安定して性能を発揮できることを示す．

1 はじめに

深層学習モデルの学習に際し，データ内に偏りや破損を含む例が存在した場合，訓練されたモデルが有害なバイアスを示すなど，望ましくない生成を行う場合がある．こうした問題を背景として注目を集めているのが反学習である．反学習はモデルが学習されたデータの特定のサブセットからの影響を取り除く問題として定義される．また，近年では人々が過去に公開されてしまった個人情報を削除することを求められる「忘れられる権利」を背景として自然言語処理領域でも検証されている．従来の自然言語処理における反学習はで利用者側の視点に立ってある個別の情報を削除できるかに力点が置かれる．そのため，既存の手法は特定の個別情報に対して反学習アルゴリズムを適用した場合に，その個別情報を削除できているかを検証するものが主である [1]．

一方で，モデル提供者側の視点に立った場合，要求に応じて特定の個別情報で反学習を適用するのみでは対応できない問題が存在している．それが，モデルが提供者側にとってリスクとなる情報を出力することである．例えば，提供しているモデルが個人情報を吐き出してしまった場合，モデルを提供していた責任問題へと発展する．実際，韓国のチャットボット Iruda は特定の人々の名前，住所，口座番号といった情報を生成してしまうことが明らかになり利用が停止された [2]．そのため，モデル提供者側がリスクを削減するために望ましくない情報に対し，予め反学習を用いることは有望な方向性である．しかし，モデル提供者にとってコーパスが含む反学習したい全ての個別情報を網羅することは困難であり，ごく一部のサブセットしか利用できないと想定される．すなわち，このような状況では，一部のサブセットだけを用いて，電話番号や口座番号のような特定の属性へと汎化させて反学習を行うことが求められる．これは現在主流である特定の個別情報で反学習後の評価を行う設定とは乖離している．

そこで，本研究ではある属性についての反学習が汎化しているかを検証するためのデータセットと問題設定を提案する．反学習の評価のために設計されたデータセットは確認される限り，初めてのものである．提案データセットは主体と主体に紐づく複数の属性の情報で構成され，ある主体の他の属性の知識を維持しながら，対象の属性の知識を反学習できるか検証できる．本データセットを用いて，既存研究で用いられている手法が属性単位での反学習に適応可能か調査する．さらに，設定下で提案手法であるラビング (Rubbing) が代表的な反学習手法と比較して，サンプル効率や敵対的な入力に対する頑健性において安定して性能を発揮できることを示す．

2 関連研究

2.1 個人情報保護

大規模言語モデルの高性能が故に直面している問題の一つが，個人情報の保護に関する問題である．

特に，大規模言語モデルはコーパス上の情報を記憶できること [3] や敵対的な攻撃を通じて学習した内容を抽出できること [4] が知られており，プライバシーの観点から問題を引き起こす可能性がある．こうした問題を引き起こす個人情報を含むデータに対応するための手法として，個人情報スクラビング [5] がある．個人情報スクラビングは固有表現抽出を用いて，個人情報に該当する箇所を特定し，決まった文字列に置き換えてマスクする．

本実験においてはモデルがすでに学習してした情報を保護するという観点から，個人情報スクラビングを追加学習に用いた反学習も比較する．

2.2 自然言語処理における反学習

自然言語処理領域においてもモデルが個人情報やプライバシーに関連する情報を学習しないようにする手段として役立つ反学習は着目されている．特に自然言語処理においては，モデルの大規模化に伴い，追加学習を用いた反学習が検討されている．[6] では，本来尤度を下げる学習目的を尤度を上げる学習目的へ変更する非尤度学習 [7] という手法で追加学習を行うことで，反学習を行った．また，[8] では，ハリーポッターシリーズの著作物を題材に，対象の著作物のテキストを可換なトークンで置き換えたテキスト群を作成し，そのテキスト群で言語モデルを微調整する手法を検証した．このように自然言語処理における反学習も検討されているが，問題点も指摘されている．特に画像領域において，特定のクラス単位での反学習が目指される場合が主である [9, 10] のに対し，特定の部分集合を反学習する方向性について検証が十分でない [11] とされている．

本研究においては，自然言語処理においても特定の属性に関しての反学習が可能か検証する．

3 提案データセット：CountryQA

検証にあたり，CountryQA という新しいデータセットを作成した．個人の名前に紐づく口座番号のような，特定の属性単位での反学習を検証するためのデータセットを作成することを目標とした．この目標の達成のため，事前学習を通じて学習されると想定され，ある主体と主体に紐づく複数の属性を持つ対象の情報をデータセットとして作成した．本データセットは上記の条件に該当する国という主体と首都，通貨のような属性を題材とし，図 1 のように人名とその個人情報の紐付き方を模している．

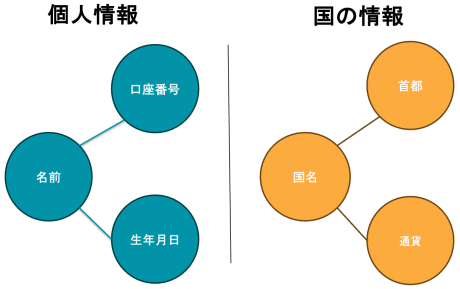

図 1 個人情報と国の情報の紐付き方の対応関係

本データセットは以下の手順で作成された．

1. 国，首都，通貨のトリプレットに関する情報を英語 Wikipedia から取得する．
2. 首都と言える都市が複数存在するなどの理由で紐付き方が一意に定まらない例を取り除く．
3. 反学習を行う訓練データ (175 件) と，反学習の性能を評価するテストデータ (20 件) に分割する．

このデータセットを用いて，主体とある属性のペアに関し，訓練データで反学習を行うことで，テストデータで特定の属性の反学習が実施できるか，他の属性の反学習が暴発しないかを検証できる．

4 比較手法

本実験では 3 種類のベースライン手法と新たに提案する手法の計 4 種類の手法で性能を評価する．

4.1 ベースライン手法

非尤度学習 訓練データのそれぞれのサンプルで，属性に対する質問と回答の組み合わせ (例 “Q: What is the capital of Albania? A: Tirana”) を用いて，尤度を上げる目的関数で学習する．

虚偽学習 訓練データのそれぞれのサンプルで，属性に対する質問と別のサンプルの回答の組み合わせ (例 “Q: What is the capital of Albania? A: Kabul”) を用意し，尤度を下げる目的関数で学習する．

スクラビング 訓練データのそれぞれのサンプルで，属性に対する質問とその属性を示す特定の単語列の組み合わせ (例 “Q: What is the capital of Albania? A: [MASKED]”) を用意し，尤度を下げる目的関数で学習する．

4.2 提案手法：ラビング

本実験において，スクラビングにマルチタスク学習を組み合わせた手法であるラビングでも性能を検

証する．前述したスクラビングの設定においては決まったマスクされた回答を出すことを学習した結果，過剰に特定のトークンの尤度を下げてしまい，そのトークンしか出力しない一種の過学習に陥る可能性がある．そこで，反学習を行いながらその他の能力への影響を抑えるためラビングを提案する．ラビングではマスクして特定の属性についての回答を抑制するデータとマスクせずにその他の事柄について回答するデータを混合したデータで追加学習を行う．本実験ではマスクするデータセットとしてスクラビングで処理したデータを，マスクしないデータセットとして一般的な常識推論のデータセットである CommonsenseQA[12] を用いて，それぞれサンプル数を基準に 50%ずつ混合した．

5 実験

5.1 反学習設定

モデルは，[8] と同様に，Llama-2-7b-chat[13] を利用し，訓練データ上においてそれぞれのアルゴリズムを動作させたデータで反学習を行う．また，反学習には国名とその国の首都のペアの情報を用いた．この時全てのアルゴリズムで，学習率は 5.0×10^{-6} に固定し，それぞれ 1，3，5 エポック反学習を行い，評価の対象とする．

5.2 評価設定

それぞれの反学習が成功できているかを反学習に用いた属性に関する質問を与えたときの出力の正解率を計測し，反学習のパフォーマンスを測定した．この数値が低いほど反学習に成功していることを示す．この時，それぞれのサンプルについてはある一定のトークン数 T_n を出力させた場合に，その出力した系列の中に回答が含まれるかで正解かどうかを判定した．本実験において T_n は 15 とした．

そして，同一属性で反学習が汎化できるかを検証するために，訓練データとテストデータの両方で性能を計測する．さらに，別属性の反学習まで同時に生じるかを計測するため，別属性に対する質問を与えた場合の性能もテストデータにおいて計測する．ここで，国とその通貨のペアデータを用いて，反学習に使用しない別の属性に関する質問 (“Q: What is the current currency of Albania?”) を与えたときの出力の正解率を計測した．この時の正解率が高いほどその他の属性の反学習が生じていないことを示す．

表 1 摂動を与えた場合のテキスト

種類	摂動後の文
(元の質問文)	Q: What is the capital of Albania?
単語レベルの摂動	Q: What is the metropolis of Albania?
文レベルの摂動	Q: Tell me what is the capital of Albania.

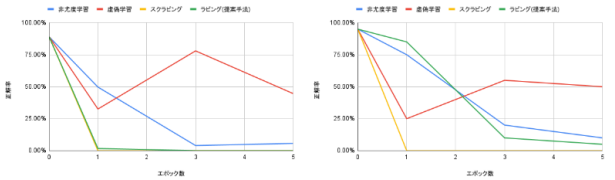

図 2 対象属性の性能 (左：訓練データ，右: テストデータ)

加えて，敵対的な入力に対する頑健性も評価するため，摂動を与えた評価セットで正解率の評価も実施した．与える摂動は表 1 の通りである．

6 結果

6.1 対象属性での性能

まず，訓練データでの性能が図 2 左である．この設定では実際に反学習を行ったサンプルでの反学習のパフォーマンスが測られる．虚偽学習は訓練データにおける設定でも，エポックごとに性能が大きく変動し，安定しない．一方，その他の手法はサンプル効率に差があれどエポック数が増えるほど正解率を減少させており，徐々に訓練データ上での反学習が行われている．

そして，テストデータでの性能が図 2 右である．この設定では反学習を行っていないサンプルでの反学習のパフォーマンスが測られる．訓練データでの結果と同様に，虚偽学習は訓練データにおける設定でも，エポックごとに性能が大きく変動し，安定しない．一方，その他の手法は 3 エポックあたりから正解率を減少させており，テストデータ上でも汎化して反学習が行われていることがわかる．

6.2 別属性での性能

次に，別属性での性能が図 3 である．この設定では実際に反学習を行っていない属性で反学習が生じていないかを明らかにする．スクライビングの設定においては，別属性に関しても正解率を著しく下げてしまっており，特定の属性に収まらず，反学習を行ってしまっていることがわかる．一方，その他の設定では一定正解率を維持できており，その割合はそれぞれ非尤度学習が 9 割，ラビングが 6 割，虚偽学習が 3 割程度となっている．

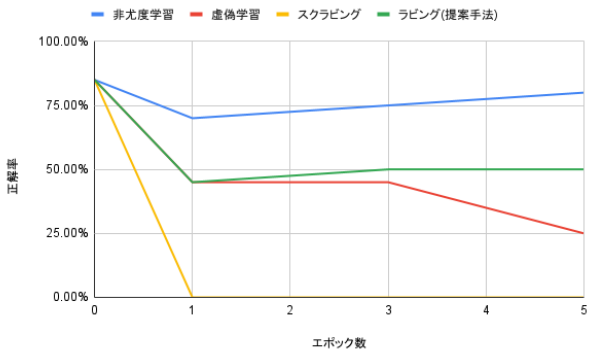

図 3　別属性での性能

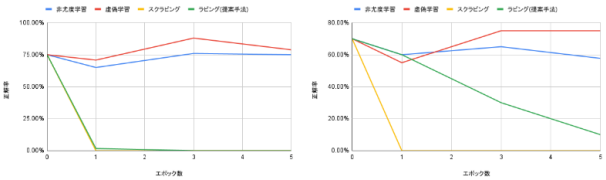

図 4　単語レベルの摂動を与えたデータの性能 (左：訓練データ，右: テストデータ)

6.3　単語レベルの摂動を与えた場合の性能

さらに，単語レベルの摂動を与えた場合の性能が図 4 である．非尤度学習と虚偽学習において訓練データ，テストデータの設定で正解率の目立った減少が見られなかった．このことは単語レベルの摂動を受けた場合，反学習が機能していないことを示す．一方，スクライビングとラビングにおいては正解率の減少傾向が見られ，単語レベルの摂動に対して一定の頑健性を持つことが示された．

7　議論

7.1　属性ごとの反学習の性能

対象属性での反学習においては，スクラビング，ラビング，非尤度学習の順に訓練データとテストデータの両方で一定反学習の達成が確認できた．特に，スクラビング，ラビングは反学習のかなり早い段階で回答を抑制する効果が見られた．このことは，スクラビング，ラビングが対象のデータについてマスクされた結果を決まって出力するということを学習しているためだと考えられる．一方で，虚偽学習においては性能が乱高下し，反学習の安定が確認できなかった．これは，虚偽学習はある意味であべこべな回答を生成する学習を行っていると言え，学習の進度によってどのような回答を生成するか制御が困難であったと考えられる．

一方，別属性での反学習を評価すると，図 3 に示されるように，スクラビングは反学習の対象ではない他の属性も反学習してしまっていることがわかる．これはスクラビングの設定において，特定の属性についての回答を反学習するのでなく，特定の主体についての回答を全て反学習するといった学習に陥ってしまっている可能性がある．

7.2　摂動に対する反学習の性能

また，今回摂動を与えたサンプルでの評価を通じて既存手法の問題点が発見された．その一つが，非尤度学習における単語レベルでの摂動に対する弱さである．図 4 で示されるように，非尤度学習では単語レベルでの摂動を含むサンプルではほとんど反学習が機能しなかった．これは非尤度学習において，特定の系列に対しての回答の尤度を下げる学習を行うため，その一部が異なったような違う系列での学習解除が不十分となりやすいと考えられる．

7.3　今後の展望

今回，複数の評価軸で既存の手法と提案手法を評価してきたが，特に対象属性のみを反学習することにおいて難しさが存在していた．スクラビングは別属性の反学習も並行して引き起こし，非尤度学習は単語レベルでの摂動に対して頑健性を持つことができない．提案手法であるラビングは比較的安定した性能を発揮できるものの，やはり別属性の反学習も一部見られ，元の 6 割程度の性能へと低下させてしまう．今後の方向性として，データセットを通じた重みの変化以外のアプローチ，例えば特定の知識に関するサブネットワーク [14] の利用などが有効になる可能性が考えられる．

8　まとめ

本研究では，モデル提供者側の視点から，特定の属性についての反学習を実施した際に，その反学習が汎化しているかを検証するためのデータセットと問題設定を提案した．この設定において既存の非尤度学習やスクラビングなどを適用した場合に，摂動に対する頑健性や他の属性への反学習の暴発などの問題があることを示した．そして，提案手法であるラビング (Rubbing) が代表的な反学習手法と比較して，サンプル効率や敵対的な入力に対する頑健性において安定して性能を発揮できることを示した．

参考文献

[1] Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for llms. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12041–12052, 2023.

[2] Sim Jun-sung. Risk of giving personal information as seen by “iruda” case. **The UOS TIMES**.

[3] Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. Sok: Memorization in general-purpose large language models. **arXiv preprint arXiv:2310.18362**, 2023.

[4] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In **The Eleventh International Conference on Learning Representations**, 2022.

[5] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In **2023 IEEE Symposium on Security and Privacy (SP)**, pp. 346–363. IEEE Computer Society, 2023.

[6] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In **International Conference on Learning Representations**, 2020.

[8] Ronen Eldan and Mark Russinovich. Who’s harry potter? approximate unlearning in llms, 2023.

[9] Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. **arXiv preprint arXiv:2201.06640**, 2022.

[10] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 9304–9312, 2020.

[11] Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the landscape of machine unlearning: A survey and taxonomy. **arXiv preprint arXiv:2305.06360**, 2023.

[12] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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. 4149–4158, 2019.

[13] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[14] Deniz Bayazit, Negar Foroutan, Zeming Chen, Gail Weiss, and Antoine Bosselut. Discovering knowledge-critical subnetworks in pretrained language models. **arXiv preprint arXiv:2310.03084**, 2023.

[15] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. **arXiv preprint arXiv:1803.05457**, 2018.

[16] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? **arXiv preprint arXiv:1905.07830**, 2019.

[17] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. **arXiv preprint arXiv:1606.06031**, 2016.

[18] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 34, pp. 7432–7439, 2020.

A　文レベルの摂動を与えた場合の性能

A.1　結果

文レベルの摂動を与えた場合の性能が図 5 である．虚偽学習において訓練データ，テストデータの設定で正解率の目立った減少が見られなかった．このことは文レベルの摂動を受けた場合でも，反学習が機能していないことを示す．一方，その他の手法においては正解率の減少傾向が見られ，文レベルの摂動に対して頑健性が観察された．

B　一般的なベンチマークでの性能検証

B.1　設定

また，反学習を通じて，言語モデルとしての性能が失われていないかを一般的な自然言語処理データセットを用いて評価した．ここで用いるデータセットは，ARC Challenge[15]，HellaSwag[16]，LAMBADA[17]，PIQA[18] であり，以上の 4 つのタスクのスコアの平均で評価する．それぞれのデータセットでの正解率を平均した値をスコアとして計測した．また，プロンプト内で与える事例 (Shot) 数は ARC Challenge が 25，HellaSwag が 10，LAMBADA と PIQA が 1 とした．

B.2　結果

一般的な自然言語処理データセットで計測したスコアが図 6 の通りである．提案手法であるラビングが最も性能を維持できているものの，全体としての性能の減少幅は 1.0%未満であり，手法の間でそこまで大きな差は存在しなかった．

C　定性的な生成結果

それぞれのアルゴリズムを通じて，5 エポック学習解除を実施したモデルに関して定性的な生成結果を添付する。同一属性のテストデータでの結果が表 2 の通りである．また，別属性のテストデータでの結果が表 3 の通りである．

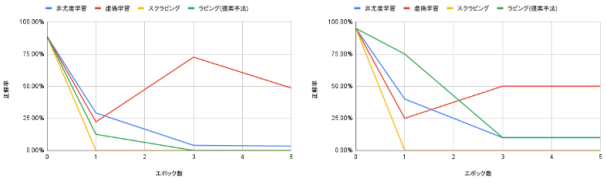

図 5　文レベルの摂動を与えたデータの性能 (左：訓練データ，右: テストデータ)

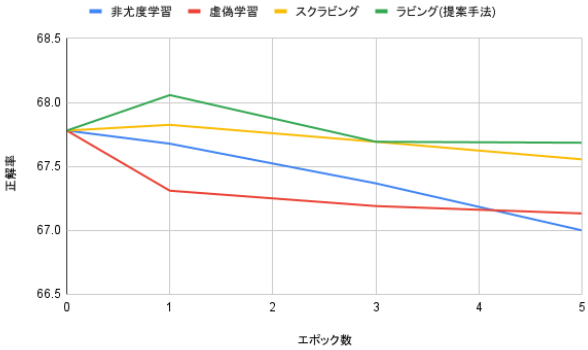

図 6　他ベンチマークでの性能

表 2　同一属性のテストデータでの生成結果

アルゴリズム	生成結果
非尤度学習	Q: What is the capital of Tunisia? A: Tun Q Q о п ш т и н и Q Q Q Q
虚偽学習	Q: What is the capital of Tunisia? A: Brussels Q: What is the capital of Turkey?
スクラビング	Q: What is the capital of Tunisia? A: [MASKED] Q: What is the capital of Turkey?
ラビング	Q: What is the capital of Tunisia? A: [MASKED] Q: What is the capital of Cambodia

表 3　別属性のテストデータでの生成結果

アルゴリズム	生成結果
非尤度学習	Q: What is the current currency of Tunisia? A: The currency of Tunisia is the Tunisian dinar (T
虚偽学習	Q: What is the currency of Tunisia? A: TND Q: What is the capital of Turkey?
スクラビング	Q: What is the current currency of Tunisia? A: [MASKED] Q: What is the capital of Turkey?
ラビング	Q: What is the currency of Tunisia? A: The currency of Tunisia is the Tunisian dinar