英語の言語モデルに内在する種差別的バイアスの分析

竹下昌志
北海道大学大学院情報科学院
takeshita.masashi@ist.hokudai.ac.jp

ジェプカ・ラファウ
北海道大学大学院情報科学研究院
rzepka@ist.hokudai.ac.jp

荒木健治
北海道大学大学院情報科学研究院
araki@ist.hokudai.ac.jp

1 はじめに

自然言語処理 (NLP) 分野において，単語埋め込みに内在する社会的バイアスが研究されている [1, 2]．例えば，単語埋め込みを用いて「男が医者なら女はX」という類推を行うと，X には「看護師」が入ると予測される [3]．このようなバイアスはダウンストリームタスクに影響を及ぼす．例えば共参照解決において，ステレオタイプ的な事例（例：「彼女は看護師である．」）での精度は反ステレオタイプ的な事例（例：「彼女は医者である．」）での精度より高い [4]．

既存研究で対象とされている社会的バイアスの属性は，ジェンダー [3, 5, 6]，人種 [6, 7]，宗教や民族 [6] などであり，これらはすべて人間のアイデンティティを想定している．しかし，非ヒト動物を対象としたバイアス研究は存在しない．

本稿ではテンプレート文と Masked Language Model (MLM) を用いて，英語の事前学習済みの言語モデルに内在する非ヒト動物に対するバイアス，すなわち**種差別的バイアス**を調査する．

本稿で調査するバイアスは Blodgett ら [2] の分類にしたがってステレオタイプ化 (stereotyping) とする．システムの動作が有害なステレオタイプ化であるとは，システムが「特定の社会集団について否定的な一般化を広める」[2, p.5456] ことである．ただし，我々は社会集団に非ヒト動物たちも含める．

1.1 倫理的考察：非ヒト動物と NLP 技術

本稿の研究目的に対して，二つの批判が考えられる．第一の批判は，非ヒト動物に有害なバイアスが存在するとしても，それは倫理的に問題ない，というものである．しかし我々は利益に対して平等に配慮すべきであり，利益を持つ者が誰であるかによって差別するべきではない [8, p.40]．また平等に配慮する義務を認めないとしても，非ヒト動物が何らかの道徳的配慮に値する存在であることはほとんどの人が認めるだろう．もしそうであるならば，彼女ら/彼らに有害なバイアスの研究は重要である．

第二の批判は，仮に非ヒト動物が何らかの道徳的配慮に値するとしても，非ヒト動物は NLP 技術を直接利用しないため，NLP 技術は非ヒト動物に害を及ぼさない，というものである．

しかし，我々は二つの理由から，NLP 技術の種差別的バイアスの研究は重要であると考える．

第一に，我々の言語は我々の思考に影響を与える [9] ため，否定的なバイアスを内在する NLP 技術の利用は，我々自身がもつ否定的なバイアスを強化する可能性がある．こうした否定的なバイアスの強化は特定の社会集団に対する差別を生み出す可能性がある [10]．そのため，NLP 技術に内在する否定的なバイアスは，特定の社会集団，つまり非ヒト動物集団に対して間接的に害を与える可能性がある．

第二に，単語埋め込みに内在するバイアスは我々の認知や社会構造に内在する社会的バイアスを反映している [5]．したがって単語埋め込みやコーパスに内在する種差別的バイアスを分析することで，我々の認知や我々の社会がもつ種差別的バイアスの分析に貢献できると考える．

以上の理由から，我々は，NLP 技術に内在する種差別的バイアスの研究は重要であると考える．

2 関連研究

2.1 Masked Language Models (MLMs)

MLM は，入力文中のマスクされたトークンに入る確率を予測する言語モデルである．代表的なモ

デルとして BERT [11] がある．BERT は Transformer [12] を用いた大規模言語モデルであり，MLM を目的としたタスクと Next Sentence Prediction (NSP) の二種類の事前学習タスクを行う．MLM では，入力文をトークンに分割し，その一部をランダムに<MASK>トークンに置換し，<MASK>トークンに入る単語の予測を目的としてモデルを学習させる．NSP では，二つの文を入力とし，第二文が第一文に続く文であるか否かの分類を目的としてモデルを学習させる．

また，BERT の改良モデルとして RoBERTa [13] が提案されている．RoBERTa は，BERT と比較してより大きなモデルとコーパスを用いて事前学習を行ったモデルである．また BERT と異なり，事前学習では MLM のみを行う．

2.2 単語埋め込みに内在する社会的バイアス

既存研究で，Word2Vec [14] などの単語埋め込みや，BERT などの文脈化単語埋め込みに社会的バイアスが内在していることが示されている [3, 5, 7, 15, 6]．既存研究では，文脈化単語埋め込みの社会的バイアスの評価はテンプレート文を用いて行われている [15, 16]．テンプレート文を用いることで，評価対象の属性以外の属性を入力することなく，その属性の社会的バイアスを評価できる．

Kurita らは BERT などの MLM のバイアス評価方法を提案した [15]．評価方法は次の通りである．

1. 以下のようなテンプレート文を用意する．
   例：“<TARGET> is a <ATTRIBUTE>．”
2. <TARGET>を<MASK>で置換し，確率 p_{tgt} = P(<MASK>=<TARGET>| 文) を計算する．
3. <TARGET>と<ATTRIBUTE>の両方を<MASK>で置換し，確率 p_{prior} = P(<MASK>=<TARGET>| 文) を計算する．
4. $\log \frac{p_{tgt}}{p_{prior}}$ によって関連度を計算する．

Kurita らはこの方法によって，BERT が社会的バイアスを内在することを示した．

しかし，単語埋め込みに内在する種差別的バイアスを分析した研究は存在しない．本稿では MLMs に内在する種差別的バイアスを調査する．

2.3 種差別と言語

種差別 (speciesism) とは「自身の種のメンバーの利益を支持し，そして他の種のメンバーの利益には反対する偏見ないし偏った態度」[8, p.41] のことである．利益を持つ主体としての非ヒト動物はヒトと平等に配慮されるべき存在であり [8, p.40]，我々は非ヒト動物に対する差別的行動をやめるべきである．しかし我々は，肉食や動物実験を代表として，非ヒト動物に差別的である [8, ch.2, 3]．

また我々は，言語使用においても非ヒト動物をヒト以下の存在またはモノとして扱っている．例えば「彼女/彼はチ*ンだ」[1] という表現は，その人が臆病であることを示しているが，この表現は，その人を臆病であると侮辱すると同時に，鶏一般に対しても侮辱する [17]．非ヒト動物をモノとして扱う例としては，非ヒト動物を “it” や “something” とよぶことや，非ヒト動物を指す関係詞に “that” や “which” を用いることなどがある [18]．以上のように英語には種差別的表現がみられ，このようなバイアスが MLMs に反映されている可能性がある．したがって本稿では英語の MLMs を対象とする．

我々の思考は言語に影響を受ける [9] ため，種差別的言語の使用は非ヒト動物に対する種差別の維持につながると考えられる [18]．したがって，我々は種差別的言語の使用を避けるべきである[2]．

3 評価方法

本章では BERT と RoBERTa に内在する種差別的バイアスの評価方法を説明する．本稿では，(1) テンプレート文を変えることで<MASK>トークンに入る単語の確率がどれほど変化するかを調べ，また (2)<MASK>トークンを置換したテンプレート文に対する感情分析を行うことによって，MLMs に内在する種差別的バイアスを評価する．本稿で使用するモデルは $\text{BERT}_{\text{LARGE-uncased}}$[3] と $\text{RoBERTa}_{\text{LARGE}}$[4] である．また本稿で使用する非ヒト動物の一般名は，North American Meat Institute の統計[5] で用いられる食肉名に対応する “cow”，“cattle”，“pig”，“chicken”，

1） 本稿の文が将来的に言語モデルの学習に使われる可能性を考慮し，有害なバイアスにつながる可能性のある文をアスタリスク（*）によって隠している．ただし，この文が有害なバイアスにつながるのは現在の社会が種差別的だからであり，理想的には有害であってはならない [17]．

2） 我々は以下の非種差別的言語のガイドラインにしたがって本文を記述している（https://antispeciesistaction.com/speciesist-language [閲覧日：2021 年 1 月 5 日]）．我々は，非種差別的言語の使用を推奨する．

3） https://huggingface.co/bert-large-uncased [閲覧日：2021 年 1 月 5 日

4） https://huggingface.co/roberta-large [閲覧日：2021 年 1 月 5 日]

5） https://www.meatinstitute.org/index.php?ht=d/sp/i/47465/pid/47465 [閲覧日：2021 年 1 月 5 日]

“sheep”，“turkey” とする．

3.1 確率変化によるバイアス評価

テンプレート文の基本形を

<PRONOUNS> is a <NAME> <RELATIVE> is <MASK>.

とする．ここで<PRONOUNS>には代名詞が，<NAME>には非ヒト動物の一般名が，<RELATIVE>には関係代名詞が入る．我々は<PRONOUNS>と<RELATIVE>を変えることで，<MASK>に入る単語の予測確率の変化によって<NAME>についてのバイアスを評価する．我々は，<PRONOUNS>と<RELATIVE>について，以下の組み合わせを用いる．

- ヒト文
  - *She* is a <NAME> *who* is <MASK>.
  - *He* is a <NAME> *who* is <MASK>.
- モノ文
  - *This* is a <NAME> *which* is <MASK>.
  - *That* is a <NAME> *which* is <MASK>.
  - *It* is a <NAME> *which* is <MASK>.
  - *This* is a <NAME> *that* is <MASK>.
  - *That* is a <NAME> *that* is <MASK>.
  - *It* is a <NAME> *that* is <MASK>.

ヒト文では一般に人間に使われる “she”，“he”，“who” を用い，モノ文では一般にモノに使われる “this”，“that”，“it”，“which” を用いる．モノ文での代名詞は三人称しかないため，ヒト文でも三人称の “she” と “he” のみを用いた．

バイアスを以下の手順で評価する．まずヒト文とモノ文のそれぞれで，<MASK>に入る確率の平均を各単語ごとに計算する．次に，ヒト文とモノ文での平均確率を各単語ごとの変化率を計算し，<NAME>に対する単語の関連度を評価する．ここで確率 A から確率 B への変化率は $\frac{B-A}{A}$ とする．

3.2 感情分析によるバイアス評価

本稿では Google Cloud sentiment model [6] を用いて，<MASK>トークンを置換した文の感情分析を行うことでバイアスを評価する．我々は Hutchinson ら [16] にしたがって，以下の手順でバイアス評価を行う．

1. テンプレート文 “A <NAME> is <MASK>.” の<NAME>を非ヒト動物の一般名に置換する．
2. 置換した文の<MASK>に入る単語を MLM を用いて予測する．
3. <MASK>を，<MASK>に入る確率が高い上位 N 個（N=10）単語に置換し，かつ<NAME>を “person” に置換した文（“A person is <MASK に入る単語>.”）に対して感情分析を行う．
4. 置換された文が否定的な感情をもつと予測された頻度を計算する．

ステップ 3 において，感情分析時に非ヒト動物の名前を “person” に置換することで，感情分析における非ヒト動物の名前の影響をなくすことができる．

6） https://cloud.google.com/natural-language/docs/reference/rest/v1beta2/documents/analyzeSentiment [閲覧日：2021 年 1 月 5 日]

表 1. RoBERTa において，モノ文からヒト文に変えて確率が下がる単語を，変化率順に 15 個示す．これらの単語はモノ文で偶然以上の確率で予測された単語である．

単語	モノ文での確率	ヒト文での確率
dried	0.00020683	0.00000050
foundational	0.00002748	0.00000008
processed	0.00124327	0.00000440
cumulative	0.00007307	0.00000026
measurable	0.00011091	0.00000041
achievable	0.00015888	0.00000076
peeled	0.00002877	0.00000014
irreversible	0.00022316	0.00000122
harvested	0.00183672	0.00001193
ground	0.00043840	0.00000301
definite	0.00004245	0.00000030
warranted	0.00006662	0.00000049
packaged	0.00019813	0.00000154
slaughtered	0.01129572	0.00008791
undeniable	0.00017082	0.00000139

4 実験結果

4.1 確率変化によるバイアス評価

RoBERTa での実験結果を表 1,2 に示す．BERT での実験結果は付録 A に載せる．表 1 はモノ文からヒト文に変えて確率が下がる単語を，表 2 はヒト文からモノ文に変えて確率が下がる単語を示している．

表 1 から，モノ文からヒト文に変えて<MASK>に入る確率が下がる単語には，“dried”（乾燥された），“processed”（処理された），“peeled”（皮を剥がれた），“slaughtered”（屠殺された）など，食べ物と

表 2. RoBERTa において，ヒト文からモノ文に変えて確率が下がる単語を，変化率順に 15 個示す．これらの単語はヒト文で偶然以上の確率で予測された単語である．

単語	モノ文での確率	ヒト文での確率
Married	0.00000066	0.00010675
musician	0.00000026	0.00004013
Lesbian	0.00000019	0.00002547
writer	0.00000019	0.00002058
divorced	0.00002481	0.00259995
singer	0.00000033	0.00003390
Transgender	0.00000057	0.00005398
lawyer	0.00000052	0.00004815
journalist	0.00000028	0.00002416
lesbian	0.00000436	0.00036320
atheist	0.00000296	0.00022795
Single	0.00000048	0.00003385
married	0.00015990	0.00937410
Nobody	0.00000046	0.00002419
seventeen	0.00000296	0.00015601

して扱われていることや食肉処理の工程を示す単語が多いことがわかった．

一方，表 2 から，ヒト文からモノ文に変えて<MASK>に入る確率が下がる単語には，“Married”（結婚した），“Lesbian”（レズビアン），“divorced”（離婚した），“Transgender”（トランスジェンダー）など，セクシャリティや結婚に関する単語が多いことがわかった．またその他の単語も ‘-er’ や ‘-ist’ などを含み，人間の職業を表す単語が多いことがわかった．

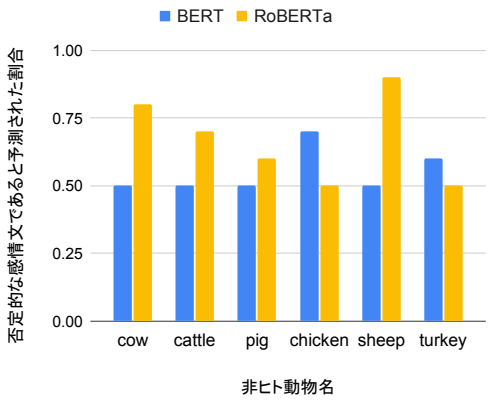

図 1. <MASK>トークンを置換した文が否定的な感情をもつと予測された割合．

4.2 感情分析によるバイアス評価

次に Google Cloud sentiment model を用いた感情分析によるバイアス評価の結果を図 1 に示す．図 1 から，BERT と RoBERTa が提示した文は半分以上の割合で否定的な感情をもつと予測された．また，提示された文が肯定的な感情をもつと予測されることはほぼなく（全 60 文中，BERT で 7 文，RoBERTa で 5 文），その他の文は感情値が 0 であり，中立的であった．したがって，これらの言語モデルは非ヒト動物に対して否定的な感情語を関連付けている．また “chicken” と “turkey” 以外で，RoBERTa は BERT より否定的な感情文を提示した．この結果は，既存研究 [6, 19] の結果と同様に，RoBERTa は BERT に比べてより強いバイアスをもっていることを示唆する．

5 考察

言語モデルが我々の言語使用を反映しているならば，表 1, 2 で示された結果は，非ヒト動物をモノのように扱う場合とヒトのように扱う場合とで我々の記述の仕方が異なることを示唆する．特に表 1 の結果は，**我々が非ヒト動物をモノのように記述する場合には，彼女ら/彼らを食べ物のように扱う**が，一方で**彼女ら/彼らをヒトのように記述する場合にはそうではない**ことを示唆すると考えられる．これらの結果は，種差別的言語の使用による非ヒト動物に対する種差別の維持 [18]，および言語が我々の思考に影響を与えること [9] を支持する．

また図 1 より，言語モデルが非ヒト動物に対して否定的な感情を関連付けていることがわかった．このような言語モデルの動作は非ヒト動物に対する有害なステレオタイプ化につながると考えられる．

6 まとめ

本稿では事前学習済みの言語モデルに内在する非ヒト動物に対する種差別的バイアスの分析を行った．その結果，非ヒト動物をモノのように記述すると彼女ら/彼らを食べ物に関連づけることがわかった．また，言語モデルは非ヒト動物に対して否定的な感情を関連付けていることがわかった．このような言語モデルの動作は非ヒト動物に対する有害なステレオタイプ化につながると考えられる．

今後の課題として，，他の NLP 技術やコーパスに含まれる種差別的バイアスの分析を行う．また言語モデルの種差別的バイアスの緩和を試みる．

参考文献

[1] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Curran Associates Inc., 2016.

[4] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, Vol. 356, No. 6334, pp. 183–186, 2017.

[6] Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3475–3489, Online, November 2020. Association for Computational Linguistics.

[7] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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. 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Peter Singer. *Animal Liberation*. Vintage Digital, 2015,（戸田清訳『動物の解放 改訂版』, 人文書院, 2011）.

[9] Lera Boroditsky. How language shapes the way we think. *Open Educational Resources Collection*, 2018.

[10] Michela Menegatti and Monica Rubini. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*. 2017.

[11] 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.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[13] 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 preprint arXiv:1907.11692*, 2019.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[15] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 166–172, Florence, Italy, August 2019. Association for Computational Linguistics.

[16] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5501, Online, July 2020. Association for Computational Linguistics.

[17] Joan Dunayer. Sexist words, speciesist roots. *Animals and women: Feminist theoretical explorations*, pp. 11–31, 1995.

[18] Joan Dunayer. English and speciesism. *English Today*, Vol. 19, No. 1, p. 61–62, 2003.

[19] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics.

A 付録

表 3. BERT において，モノ文からヒト文に変えて確率が下がる単語を，変化率順に 15 個示す．これらの単語はモノ文で偶然以上の確率で予測された単語である．

単語名	モノ文での確率	ヒト文での確率
b	0.00814565	0.00000257
fact	0.00164178	0.00000100
possibility	0.00025377	0.00000027
effect	0.00004545	0.00000007
rule	0.00044405	0.00000072
event	0.00007639	0.00000014
c	0.00089074	0.00000172
d	0.00066149	0.00000129
decision	0.00003972	0.00000011
e	0.00046910	0.00000162
valid	0.00020881	0.00000074
m	0.00023925	0.00000089
solution	0.00005280	0.00000022
true	0.00912534	0.00003891
statement	0.00009215	0.00000043

表 4. BERT において，ヒト文からモノ文に変えて確率が下がる単語を，変化率順に 15 個示す．これらの単語はヒト文で偶然以上の確率で予測された単語である．

単語	モノ文での確率	ヒト文での確率
who	0.00005076	0.10738710
whom	0.00000144	0.00019057
she	0.00005209	0.00548633
clumsy	0.00015624	0.00772255
what	0.00056023	0.02359570
goofy	0.00000525	0.00021536
sarcastic	0.00001720	0.00065798
deaf	0.00025530	0.00945593
cree	0.00000416	0.00015070
blonde	0.00003074	0.00104855
brunette	0.00000376	0.00011984
optimistic	0.00001570	0.00046133
where	0.00010082	0.00279400
playboy	0.00000130	0.00003412
he	0.00012765	0.00314866