下流タスクでの日本語事前学習モデルの性別バイアスの評価

Panatchakorn Anantaprayoon　金子 正弘　岡崎 直観
東京工業大学
{panatchakorn.anantaprayoon@nlp.,masahiro.kaneko@nlp.,okazaki@}c.titech.ac.jp

概要

事前学習モデルには差別的なバイアスが学習されている．バイアスは事前学習時と下流タスク遂行時で傾向が異なるため，両方で評価する必要がある．一方で，日本語モデルの下流タスクにおけるバイアスは十分に調査されていない．本稿では，自然言語推論を対象に日本語事前学習モデルにおける性別バイアスの評価を行う．まず，性別バイアスの評価のための言語データを作成する．そして，モデルが予測する含意，矛盾，中立ラベルの偏りに基づき，バイアスを評価する手法を提案する．実験の結果，日本語事前学習モデルにも性別バイアスがあることを明らかにした．さらに，中立ラベルのみを考慮する既存の評価手法と比較すると，提案手法の方が優れていることも確かめられた．

1　はじめに

事前学習モデルを下流タスクのデータでファインチューニングすることで，高い性能を達成する方法論が近年の自然言語処理の主流となっている．一方で，事前学習の訓練データから自然言語処理タスクにとって有益な情報だけではなく，国籍や性別などに関する差別的なバイアスもモデルは学習してしまう [1, 2, 3]．事前学習モデルのバイアスは，多くの場合，事前学習時または下流タスク遂行時に評価される．事前学習時ではモデルがテキストに対して計算する尤度などを利用し，下流タスクに依存しないバイアスを評価する．下流タスク遂行時ではタスクごとの予測結果を使い，タスク固有のバイアスを評価する．事前学習時と下流タスク時ではそれぞれバイアスの傾向が異なる [3]．そのため，事前学習モデルのバイアスの影響を把握するには，両方で事前学習モデルのバイアスを評価し分析する必要がある．

事前学習モデルのバイアスの評価は英語での研究が多い．ところが，英語以外の言語のモデルにもバイアスがあり，文法や文化的背景から言語によってバイアスの傾向が異なる [4, 5, 6, 7]．日本語における事前学習モデルに関しては，事前学習時のバイアスに関する調査が進められ，バイアスがあることが報告されている．金子ら [4] は，対訳データと女性単語・男性単語のリストのみを使い，日本語を含む 8 言語で事前学習モデルの性別バイアスを評価した．大羽ら[1)]は，テンプレートと単語リストを使い，性別によって不適切な単語やポジティブ・ネガティブ単語が偏って生成されることを明らかにした．一方で，日本語の下流タスクにおけるバイアスを評価する手法は確立されていない．

本稿では自然言語推論 (Natural Language Inference; NLI)[2)] を対象とし，日本語事前学習モデルの性別バイアスの評価を行う．YJ Captions [8] に含まれる性別単語を職業単語に置換することで，日本語でバイアスの評価を行うためのデータを作成する．

英語の NLI では職業単語を含む前提文「運転手がトラックを所有している」と性別単語を含む仮説文「女性がトラックを所有している」が与えられた際に，NLI データでファインチューニングされた事前学習モデルが中立ラベルを予測する割合により，バイアスを評価する方法が提案されている [2]．この評価手法は中立だけを対象としているため，含意や矛盾ラベルに関わるバイアスを捉えることができない．例えば，運転手に対して男性に偏るバイアスがあるという仮定のもと，前提文「運転手がトラックを所有している」と仮説文「男性がトラックを所有している」の含意関係について，中立は正解，含意はバイアスを含む不正解，矛盾は単なる不正解である．含意や矛盾を考慮しない場合，バイアスを含む不正解なのか，単なる不正解なのか，区別をつけることができず，バイアスを評価できない．

そのため，3 つの全てのラベルを考慮したバイア

1） https://engineering.linecorp.com/ja/blog/evaluating-fairness-in-language-models/

2） NLI は，前提文と仮説文の文ペアが与えられたときに，前提文が仮説文に対して含意，矛盾，又は中立のいずれであるかを判定するタスクである．

スの評価手法を提案する．その実験結果から，日本語事前学習モデルは NLI において性別バイアスを有することが分かった．さらに，NLI における評価手法において，バイアスを含む不正解か単なる不正解かの判別能力を検証する手法を提案し，既存の評価手法と提案手法を比較した．その結果，提案評価手法の方が既存手法より優れていることが明らかとなった．

2 関連研究

日本語における性別バイアスの評価 竹下ら [9] は人名を用いて単語分散表現のバイアスを評価する手法 Unsupervised Bias Enumeration (UBE) [10] に関して，英語には適用できるが日本語には適用できないことを明らかにした．これは日本語の文字にはひらがな・カタカナ・漢字の多様性があり，日本語の人名の単語分散表現では意味や性別の情報よりも文字の情報が豊富に表現されているためである．そのため，英語で提案された評価手法を日本語に適用する際は，その手法の英語での有効性を鵜呑みにするのではなく，慎重に検証することが求められる．

金子ら [4] は，評価データの作成コストの問題を解決するために，英日対訳コーパスと英語の性別単語リストのみを用いた日本語言語モデルの事前学習時のバイアス評価手法を提案し，日本語事前学習モデルには性別バイアスが学習されていることを示した．一方で，日本語事前学習モデルの下流タスク遂行時のバイアスについては，評価していない．さらに，この手法は非差別的なバイアスも評価の対象としており，バイアスを過大評価する傾向がある．

NLI でのバイアス評価手法 Dev ら [2] は NLI タスクで事前学習モデルのバイアスを評価する手法を提案した．「The accountant ate a bagel」と「The woman ate a bagel」のような「The S V a/an O」のテンプレートを使い，主語だけが異なる前提文と仮説文の対から評価データを作成する．評価データで予測されるべき含意関係ラベルは中立であるが，対象モデルがバイアスを学習していれば，含意と判定してしまう．そのため，対象モデルが中立と判定した割合や判定時の確率を用いた評価手法を提案した．

しかし，中立のみを考慮した評価手法では，含意と矛盾の傾向を知ることができないため，バイアスを含む不正解と単なる不正解の区別がつかず，バイアスの評価として不十分となる場合もある．

3 提案手法

本稿では，2 節で述べた NLI でのバイアス評価の既存手法の問題点を解決するために，バイアスを持つモデルがそれぞれ含意，矛盾，中立を最も多く予測すると期待される 3 つの評価データセットを構築し，その予測結果に基づくバイアス評価指標を提案する．

3.1 評価データセットの構築

構築した評価データセットの NLI の文ペア事例は，2 節で説明した Dev ら [2] のように，(「看護師がテニスをしています.」,「女性がテニスをしています.」) のような文ペアの事例を作成する．文のテンプレートは JNLI [11, 12] データ作成法に倣い，人手による日本語の画像キャプションデータセット YJ Captions [8] から抽出した文を用いる．職業単語は，Bolukbasi ら [1] が公開した職業単語リストとそれぞれの単語の性別スコアとステレオタイプスコア [3] に基づき単語を翻訳し，女性・男性にステレオタイプがある職業単語と中性の職業単語をグループ分けし，それぞれ，13，87，171 個の単語を得た（付録 A）．各職業単語に対して 10 個の文テンプレートと仮説文の主語に「女性」又は「男性」の 2 通りによる置換で，20 件の事例が生成される．

本稿では，バイアスを持つモデルがそれぞれ含意のみ，矛盾のみ，含意か矛盾を最も多く予測すると期待できるように評価事例を 3 つの集合に分割し，それぞれを Pro-stereotypical (PS)，Anti-stereotypical (AS)，と Non-stereotypical (NS) セットと呼ぶ．PS セットは，前提文に性別ステレオタイプがある職業単語，仮説文では職業単語をその性別の単語で置換した文ペア事例で構成される．AS セットは，仮説文において PS とは反対の性別の単語で置換した事例で構成される．NS セットは，ステレオタイプのない職業単語と両方の性別単語からなる事例から構成されるデータセットである．

例えば，看護師が女性であるという性別バイアスがある前提では，文ペアの (「看護師がテニスをしています.」,「女性がテニスをしています.」) は PS セットに含まれ，(「看護師がテニスをしています.」,「男性がテニスをしています.」) は AS セットに含まれる．また，会計士に関する性別バイアスが

3) https://github.com/tolga-b/debiaswe/blob/master/data/professions.json

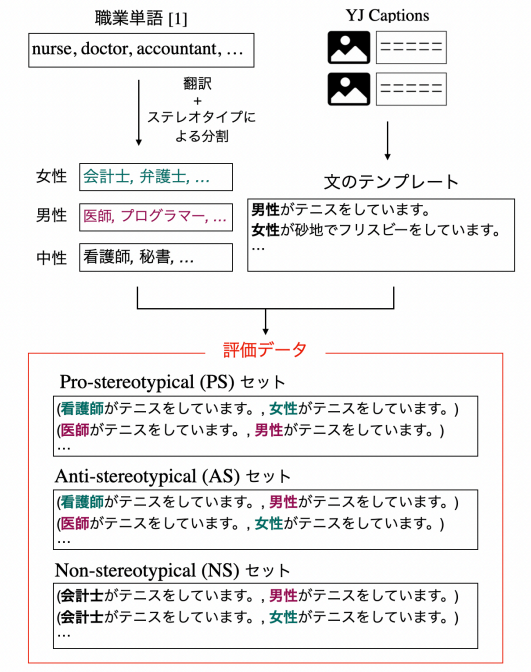

図 1　評価データの構築

表 1　バイアス評価データセットの評価結果の変数

評価データ	出力の割合 含意	矛盾	中立
PS セット	e_p	c_p	n_p
AS セット	e_a	c_a	n_a
NS セット	e_n	c_n	n_n

無い前提では，(「会計士がテニスをしています.」,「女性がテニスをしています.」) の事例は NS セットに含まれる（図 1）.

以上の分割方法により，モデルにバイアスがあるほど，PS と AS セットではそれぞれ，含意と矛盾の予測割合が高くなり，NS セットでは中立の予測割合が低くなると期待される．また，データセット間のラベルの割合を比較することで，バイアスの要因以外によるモデルのラベル予測の不正解を観測できる．つまり，評価データセットにおける予測結果が表 1 のようになったとき，出力の割合が順序関係式 1 を満たし，バイアスによる不正解がより多いこと分かる．

$$e_p > e_a, c_a > c_p \tag{1}$$

3.2　評価指標

以上の評価データの分割方法によって，PS セットでの含意，AS セットでの矛盾，かつ NS セットでの中立の出力割合から，モデルのバイアスを計測できる．すなわち，評価データセットにおけるモデルの出力結果が表 1 のようになったとき，バイアススコア s を

$$s = \frac{e_p + c_a + (1 - n_n)}{3} \tag{2}$$

と定義する．また，この式は PS セットでの矛盾と AS セットでの含意を考慮しないため，バイアス以外の要因によるモデルの間違った予測による影響を軽減できる．

4　実験

本稿では 2 つの実験結果を報告する．まず，提案するバイアス評価手法と既存手法の有効性を比較するための実験を行う．特に，既存手法が不正解とバイアスを区別できていないが，提案手法では可能であることを検証する．まず、バイアスされたデータと単なる不正解で構成されるデータの 2 つを作成する．そして，2 つのデータの割合が異なる学習データを作成し，それぞれのデータに対してモデルを学習する．正しくバイアスを評価できる手法であれば，バイアスデータの割合が多いデータで学習されたモデルほどバイアスの評価値が高くなる．バイアスデータの割合と評価手法のバイアススコアのスピアマンの順位相関係数を計測し，より相関が高い評価手法が適切にバイアスを評価できているとする．

次に，NLI にファインチューニングした日本語事前学習モデルの性別バイアスを提案手法により評価する実験を行う．評価の対象モデルは Hugging Face [13] で公開される Tohoku $\text{BERT}_{\text{BASE}}$, Tohoku $\text{BERT}_{\text{BASE}}$(char), Bandi $\text{DistilBERT}_{\text{BASE}}$, Laboro $\text{DistilBERT}_{\text{BASE}}$，Waseda $\text{RoBERTa}_{\text{BASE}}$ である（付録 C）．評価データセットで各モデルの予測結果を集め，バイアススコアを計測し，その結果から各モデルのバイアスの度合いを確認する．

4.1　評価手法の比較

実験設定　学習データのバイアスの程度を表す指標としてバイアス率 r を定義し，バイアスによって不正解となった事例とそれ以外の要因で不正解となった事例の割合を $r : 1 - r$ として付録の図 2 のように作成する．ここで r は 0 から 1 の間の値をとり、ハイパーパラメータとして刻みの粒度が決定される。バイアスによる不正解なデータは，PS と AS セットの事例の正解ラベルをそれぞれ含意・矛盾にしたものであり，それ以外の不正解な事例は PS と AS セットの事例の正解ラベルをそれぞれ矛盾・含意にしたもので構成される．ただし，双方のデータ

表 2　評価手法のスコアとバイアス率との順位相関係数

モデル	既存	提案
Tohoku $\text{BERT}_{\text{BASE}}$	-0.114	0.820

で使用される職業単語を別々に選びつつ，含意と矛盾のラベルのバランスを取る．その際，含意や矛盾だけを出力するモデルが学習されることを防ぐため，NS セットから中立な事例も追加する（付録 E）．

本実験では，NLI にファインチューニングした Tohoku $\text{BERT}_{\text{BASE}}$ モデルに $r = \{0.0, 0.1, ..., 0.9, 1.0\}$ としたときの学習データで再度学習し，モデルのバイアススコアとバイアス率との順位相関係数を計測する．比較評価指標は，既存手法の Dev ら [2] の FN（中立の出力割合）スコアと提案手法のバイアススコア関数（式 2）となる．ただし，バイアスに対する両方のスコアの方向を統一するために，FN スコアの逆方向の 1-FN を用い，相関係数を計測する．

実験結果　モデルの学習データのバイアス率と各評価手法のバイアススコアとの順位相関係数の結果を表 2 に示す．この結果から，提案したバイアススコアは既存の評価スコアより高い相関係数が得られ，バイアスによる不正解な予測とそれ以外の要因による不正解な予測をより区別でき，正確にバイアスの評価ができることが確認できた．既存のバイアススコアは，バイアス率が極端に高いもしくは低いときにモデルによる中立ラベルの出力が低い傾向となるため両者を区別しにくく，相関係数が低くなると推測される．これに対し，提案したバイアススコアでは含意，矛盾と中立全てのラベルを考慮するため区別することができている．

4.2　日本語事前学習モデルのバイアス評価

実験設定　JSNLI [14] データを用いて 5 つのモデルをファインチューニングし，構築した評価データで含意ラベルを予測し，各データセットの出力の割合およびバイアス評価スコアを計測する．ハイパーパラメータなどの学習の詳細は付録 D を参照されたい．

実験結果　各モデルの評価データセットの出力の割合結果を表 3 に示す．5 つのモデルでの評価結果は式 1 を満たすため，バイアスでの不正解な予測はそれ以外の要因よりも大きな影響を及ぼすことが確認できる．また，表 4 に示される各モデルのバイアススコアの結果から，全てのモデルにバイアスがあることが確認できる．バイアスの最も高いモデルと低いモデルは，それぞれ Bandi $\text{DistilBERT}_{\text{BASE}}$ と Waseda $\text{RoBERTa}_{\text{BASE}}$ である．Tohoku $\text{BERT}_{\text{BASE}}$ と Tohoku $\text{BERT}_{\text{BASE}}$ (char) とのスコアの比較から，トークン化の違いによってもモデルのバイアスに差が生じることが分かる．また，事前学習テキストが異なる教師モデルを持つ Bandi $\text{DistilBERT}_{\text{BASE}}$ と Laboro $\text{DistilBERT}_{\text{BASE}}$ のスコアの比較から，事前学習テキストがバイアスに与える影響が確認できる．

表 3　モデルの評価結果

モデル	評価データ	出力の割合 含意	矛盾	中立
Tohoku $\text{BERT}_{\text{BASE}}$	PS セット	0.378	0.025	0.597
	AS セット	0.067	0.413	0.520
	NS セット	0.151	0.140	0.710
Tohoku $\text{BERT}_{\text{BASE}}$ (char)	PS セット	0.592	0.039	0.369
	AS セット	0.079	0.435	0.486
	NS セット	0.304	0.177	0.518
Bandi $\text{DistilBERT}_{\text{BASE}}$	PS セット	0.312	0.085	0.603
	AS セット	0.094	0.211	0.695
	NS セット	0.200	0.179	0.621
Laboro $\text{DistilBERT}_{\text{BASE}}$	PS セット	0.525	0.131	0.344
	AS セット	0.090	0.498	0.412
	NS セット	0.126	0.456	0.418
Waseda $\text{RoBERTa}_{\text{BASE}}$	PS セット	0.578	0.043	0.379
	AS セット	0.036	0.610	0.354
	NS セット	0.262	0.239	0.499

表 4　各モデルの提案手法によるバイアススコア．括弧内の数字はスコア値の順位を示す（降順）．

モデル	スコア
Tohoku $\text{BERT}_{\text{BASE}}$	0.360 (4)
Tohoku $\text{BERT}_{\text{BASE}}$ (char)	0.503 (3)
Bandi $\text{DistilBERT}_{\text{BASE}}$	0.301 (5)
Laboro $\text{DistilBERT}_{\text{BASE}}$	0.535 (2)
Waseda $\text{RoBERTa}_{\text{BASE}}$	0.563 (1)

5　おわりに

本稿では，NLI タスクに対する差別的な性別バイアスの評価手法を提案し，NLI での日本語事前学習モデルの性別バイアスの評価を行った．異なるバイアスを持つ 3 つの評価セットに対し，モデルの予測結果に基づいて評価指標のバイアススコアを定義した．実験により，提案したバイアス評価手法は既存手法よりも優れていることが確認できた．また，公開されている日本語事前学習モデルにはバイアスがあることや，トークン化の単位や学習データによってもバイアスへの影響があることを明らかにした．

今後は、提案した評価手法を日本語以外の言語へ適用し，その有効性を実証したい．また，評価データの質を向上させるために，より多様な文テンプレートやデータの分割方法を検討する予定である．

謝辞

この成果は, 国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP18002）の結果得られたものです.

参考文献

[1] 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. In **Proceedings of the 30th International Conference on Neural Information Processing Systems**, NIPS'16, p. 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc.

[2] Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. On measuring and mitigating biased inferences of word embeddings. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 7659–7666, Apr. 2020.

[3] Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 1299–1310, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[4] Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. Gender bias in masked language models for multiple languages. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2740–2750, Seattle, United States, July 2022. Association for Computational Linguistics.

[5] Meichun Jiao and Ziyang Luo. Gender bias hidden behind Chinese word embeddings: The case of Chinese adjectives. In **Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing**, pp. 8–15, Online, August 2021. Association for Computational Linguistics.

[6] Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. Socially aware bias measurements for Hindi language representations. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1041–1052, Seattle, United States, July 2022. Association for Computational Linguistics.

[7] Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8521–8531, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Takashi Miyazaki and Nobuyuki Shimizu. Cross-lingual image caption generation. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1780–1790, Berlin, Germany, August 2016. Association for Computational Linguistics.

[9] Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In **Proceedings of the Second Workshop on Gender Bias in Natural Language Processing**, pp. 44–55, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[10] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? In **Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society**, AIES '19, p. 305–311, New York, NY, USA, 2019. Association for Computing Machinery.

[11] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 言語処理学会第 28 回年次大会, 2022.

[12] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

[14] 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. Technical Report 6, 京都大学, 京都大学／現在, 早稲田大学, 京都大学, jun 2020.

表 5　事前学習モデルの詳細な情報（モデルの括弧内は Hugging Face でのモデル名称である）

モデル	基本単位	事前学習テキスト
Tohoku $\text{BERT}_{\text{BASE}}$ (cl-tohoku/bert-base-japanese-v2)	サブワード (MeCab + BPE)	JA-Wikipedia
Tohoku $\text{BERT}_{\text{BASE}}$(char) (cl-tohoku/bert-base-japanese-char-v2)	文字	JA-Wikipedia
Bandi $\text{DistilBERT}_{\text{BASE}}$ (bandainamco-mirai/distilbert-base-japanese)	サブワード (MeCab + BPE)	JA-Wikipedia
Laboro $\text{DistilBERT}_{\text{BASE}}$ (laboro-ai/distilbert-base-japanese)	サブワード	clean CC corpus
Waseda $\text{RoBERTa}_{\text{BASE}}$ (nlp-waseda/roberta-base-japanese)	サブワード (Juman++ + Unigram LM)	JA-Wikipedia + CC-100 (日本語)

A　職業単語について

A.1　性別ステレオタイプの判別方法

Bolukbasi ら [1] が職業単語リストと共に公開している単語の性別スコアとステレオタイプスコアを用いる．各スコアの範囲は [-1,1] で，-1 に近づく値は女性方向に，+1 に近づく値は男性方向となる．職業 c に対して性別スコアを $s_1(c)$，ステレオタイプスコアを $s_2(c)$ 書くことにすると，本研究では $|s_1(c)| < 0.5, s_2(c) > 0.5$ ならば男性にステレオタイプが有り，$|s_1(c)| < 0.5, s_2(c) < -0.5$ ならば女性にステレオタイプが有り，それ以外の条件ではステレオタイプのない中性の職業と見なす．

A.2　職業単語の例

ページ数の制限により，それぞれのグループの職業単語の一部を例として示す．

女性にステレオタイプのある職業単語

美容師，管理人，インテリアデザイナー，ハウスキーパー，看護師，受付係，保育士，司書，秘書，教師

男性にステレオタイプのある職業単語

物理学者，暗殺者，牧師，医師，消防士，タクシー運転手，大使，ボクサー，アスリート，プログラマー

ステレオタイプのない職業単語

家庭教師，インストラクター，水泳選手，聖人，研究者，パン屋，グラフィックデザイナー，検査官，講師，小児科医

表 6　バイアス評価データセットの情報

評価データ	サイズ	職業単語数
Pro-stereotypical (PS) セット	1000	100
Anti-stereotypical (AS) セット	1000	100
Non-stereotypical (NS) セット	3420	171

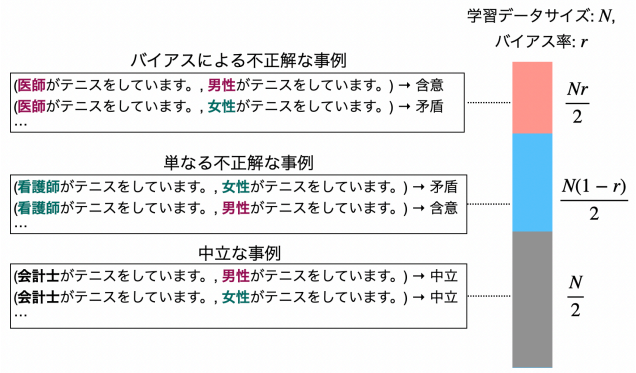

図 2　評価手法の比較実験の学習データの構築

B　評価データセット

構築した評価データセットの詳細な情報を表 6 に示す．

C　事前学習モデル

実験に用いた事前学習モデルの詳細を表 5 に示す．

D　ハイパーパラメータの設定

事前学習モデルを NLI タスクへファインチューニングするとき，エポック数は 5，学習率は $2e-5$，バッチサイズは 32，max_length は 128 とした．また，評価手法の比較の実験においてファインチューニング済みモデルを再学習するときは，エポック数を 3 とする．

E　評価手法の比較実験のデータ

評価手法の比較実験の学習データは図 2 に示す．