Minimal-pair Paradigm データセットにおけるトークン長バイアスの分析と改善

上田直生也 [1]　三田雅人 [2,1]　小町守 [3]
[1] 東京都立大学　[2] 株式会社 サイバーエージェント　[3] 一橋大学
ueda-naoya@ed.tmu.ac.jp　mita_masato@cyberagent.co.jp
mamoru.komachi@r.hit-u.ac.jp

概要

教師なし手法で言語モデルの言語能力を測るために，Minimal-pair paradigm（MPP）データセットがベンチマークとして用いられる．MPP では，容認可能な文に対して容認不可能な文より高い対数尤度を予測したミニマルペアの割合によって，言語能力を評価する．この対数尤度は文長に影響されるため，容認可能な文と容認不可能な文の単語数を揃えることが通例である．しかしながら，近年の言語モデルは文をトークン化するため，単語数を揃えるだけでは不十分である可能性がある．本研究は，容認可能な文と容認不可能な文の間でトークン長が異なることで起きるバイアスが言語・データセット横断的に存在しており，トークン長を揃えることでバイアスが改善できることを示す．

1　はじめに

容認性判断タスクは与えられた文が人間にとって容認可能かを判定するタスクであり，言語モデルの言語知識を測定する際に用いられる [1, 2]．最も広く使われている容認性判断コーパスは CoLA（Corpus of Linguistic Acceptability） [3] であり，言語知識の測定には分類器の教師あり学習が必要である．しかしながら，そのような分類器を用いた測定は，言語モデルが言語知識を事前学習時に獲得したのか，それとも分類器の教師あり学習時に獲得したのかの区別が難しいという問題がある [4]．

この問題を解決するために，教師なし手法による容認性判断が注目されている．その中でも，Minimal-pair Paradigm（MPP）データセットを利用したアプローチが主流となっている． [4, 5, 6]．MPP データセットは，一単語のみ異なるような容認可能な文と容認不可能な文のペアであるミニマルペアの集合で構成される [7]．言語モデルの言語知識は，モデルが容認可能な文に対して容認不可能な文よりも高い容認度を割り当てるミニマルペアの割合に基づいて評価する．評価では，各ミニマルペアで容認度を予測する必要があり，文の対数尤度が一般的に使用されている．この指標は，言語モデルが正しい言語知識を獲得しているならば，容認可能な文に対して容認不可能な文よりも高い対数尤度を推定するという仮定の下で使用されている．

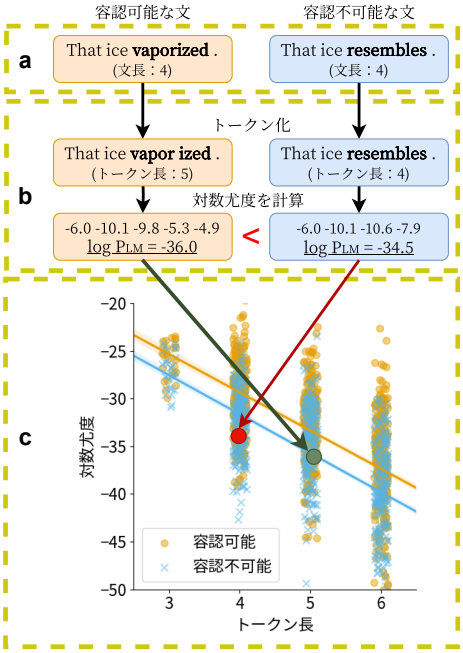

図 1　MPP データセットにおけるトークン長バイアスの例．（a）同じ文長である元のミニマルペア，（b）トークン長バイアスの例，（c）対数尤度とトークン長の比例関係．

文の対数尤度が容認度として MPP データセットで用いられているが，各ミニマルペアにおける容認可能な文と容認不可能な文の単語数を揃える必要があるという制約が存在する．これは，文の対数尤度は文の長さに比例して小さくなることが知られてい

るためであり，正しい性能評価を行うためには必要不可欠な制約である（図 1-a）[4]．しかし，GPT-2 [8] や BERT [9] のような現在使用される事前学習済み言語モデルを評価するには，不十分な制約である可能性がある．なぜなら，事前学習済み言語モデルは文をサブワード単位でトークン化するため，単語数は同一な容認可能な文と容認不可能な文においても，トークン長では異なる可能性がある（図 1-b）．そして，対数尤度はトークン長にも影響されることが知られているため [10]，このようなトークン長の差はトークン長バイアスを引き起こし，評価結果に影響を与える可能性がある（図 1-c）．しかしながら，MPP データセットにトークン長バイアスを引き起こすようなミニマルペアが存在するのか，また評価結果にどのような影響が生じるかは自明でない．

そこで本研究は，トークン長バイアスが MPP データセットを用いた評価に与える影響について分析した．まず，MPP データセットの言語や種類を問わず，ミニマルペアの容認可能な文と容認不可能な文のトークン長が異なると，評価に悪影響を及ぼすことを示す．加えて，トークン長で正規化した対数尤度は，トークン長バイアスを軽減できるか検証する．これは，対数尤度はトークン長と負の相関があるため [10]，トークン長で対数尤度を正規化することでトークン長バイアスの影響を軽減できる可能性があるからである．そのため本論文では，「トークン長バイアスは MPP データセットの評価結果に影響するか？」と「トークン長で正規化した対数尤度はトークン長バイアスを軽減できるのか？」という 2 つの研究課題を明らかにすることを目的に分析を行った．本研究の貢献は以下の通りである：

- MPP データセットではトークン長バイアスが生じており，モデルの言語能力を正しく評価できないことを明らかにした．
- 正規化対数尤度はトークン長バイアスを軽減できず，誤った評価結果になることを示した．
- ミニマルペアのトークン長を揃えることでトークン長バイアスを改善する手法を提案し，ケーススタディとして BLiMP データセットの改善を行うことでその有効性を分析した．

2 関連研究

MPP データセットは，BLiMP [4] のようなモデルの文法知識を評価するものが多かったが，現在では社会的偏見の度合いを評価する CrowS-Pairs [5] や上位・下位概念の認識度合いを評価する COMPS [6] などが登場している．また，MPP データセットの多くは，英語の言語知識の評価に焦点を当てている．しかし，英語以外の言語モデルの言語知識を測定する必要性から，様々な言語の MPP データセットが提案されている．例えば，SLING [11] や JBLiMP [12]，Arabic minimal pairs（Arabic MP）[13]，French CrowS-Pairs [14] などがある．

MPP データセットにおいて，一般的に文の対数尤度が容認度として使用される [4, 5]．GPT-2 のような自己回帰言語モデルでは，連鎖規則を適用して各トークンの確率を合計することで，文の対数尤度を容易に推定することができる．文 S が与えられたとき，文の対数尤度 $logP_{LM}(S)$ は，すでに出力されたトークン $S_{<t} := (w_1, ..., s_{t-1})$ から各トークン s_t を予測する条件付き対数確率の和として計算される．これは次のように表される：

$$\log P_{LM}(S) = \sum_{t=1}^{|S|} \log P_{LM}(s_t|S_{<t}) \tag{1}$$

一方，BERT のような双方向文脈表現を使用するマスク言語モデルでは，文の対数尤度を直接推定することができない [9]．代わりに，Pseudo-log-likelihood（PLL）が用いられる [15]．PLL では，トークン s_t をマスクし，その前後のトークン $S_{\setminus t} := (s_1, ..., s_{t-1}, s_{t+1}, ..., s_{|S|})$ を用いてトークンの対数確率を予測する．マスク言語モデルにおいて，文の PLL$logP_{MLM}(S)$ は，各トークンの条件付き対数確率 $\log P_{MLM}(s_t|S_{\setminus t})$ の和として計算でき，次式で表される：

$$\log P_{MLM}(S) = \sum_{t=1}^{|S|} \log P_{MLM}(s_t|S_{\setminus t}) \tag{2}$$

PLL はマスク言語モデルにおいて文の対数尤度を推定することを可能としたが，語彙外の単語の PLL を過大評価してしまうという問題がある．この問題を克服するために，Kauf ら [10] は PLL に代わる計算手法として PLL-word-l2r を提案した．対数尤度を予測したいサブワードトークン s_{w_t} のみをマスクするのではなく，語彙外の単語を構成する将来のサブワードトークン $s_{w_{>t}}$ をマスクすることで，PLL を推定する．PLL-word-l2r は文の対数尤度を以下のように推定する：

$$\log P_{MLMl2r}(S) = \sum_{w=1}^{|S|} \sum_{t=1}^{|w|} \log P_{MLM}(s_{w_t}|S_{\setminus s_{w_{t' \geq t}}}) \tag{3}$$

3 研究課題の調査方針

3.1 調査観点 1：トークン長バイアス

本研究では，容認可能な文と容認不可能な文におけるトークン長の差が MPP データセットを用いた評価に対する影響を調査する．そのために，各 MPP データセットにおいて，容認可能な文のトークン長（A）が容認不可能な文のトークン長（U）と比較して，同じか，長いか，短いかでミニマルペアを分類した（以降それぞれの分類を A=U，A>U，A<U と示す）．もし，トークン長バイアスが MPP データセットを用いた評価に影響しないのであれば，全ての分類において精度は同じであり，変化しないはずである．そのため，A>U と A<U の精度を A=U の精度と比較することで MPP データセットにおけるトークン長バイアスを分析する．具体的には，容認可能な文から容認不可能な文を引いたものをトークン長の差とし，トークン長の差と精度の変化を観察する．

3.2 調査観点 2：対数尤度の正規化

対数尤度はトークン長に比例するため，対数尤度をトークン長で正規化することで，トークン長バイアスを軽減できる可能性がある．実際に先行研究 [16, 17, 12] では，トークン長の影響を軽減するために，正規化した文の対数尤度を使用している．彼らは MeanLP，PenLP，SLOR [18] のような正規化手法を用いて文の対数尤度を正規化した．しかしながら，このような正規化手法が MPP タスクにおいて有効であり，文の対数尤度を正しく正規化できるかどうかは不明である．そこで本研究では，MPP データセットにおいて文の対数尤度をトークン長で正規化することが有効かどうかを分析する．

本研究では，正規化手法として MeanLP と PenLP を用いる．MeanLP と PenLP はどちらもトークン長で対数尤度を正規化するが，PenLP はトークン長を係数 α でスケーリングする．本研究では $\alpha = 0.8$ に設定した．MeanLP と PenLP はそれぞれ以下のように計算される：

$$MeanLP = \frac{\log P_{LM}(S)}{|S|} \quad (4)$$

$$PenLP = \frac{\log P_{LM}(S)}{((|S|+5)/(5+1))^{\alpha}} \quad (5)$$

本研究では，正規化した文対数尤度とトークン長の相関を観察することにより，正規化した対数尤度がトークン長バイアスを緩和できるか調査した．

4 実験

4.1 モデル

トークン長バイアスがモデルの種類によらず生じる現象であることを示すために，自己回帰言語モデルとして GPT-2 [8]，マスク言語モデルとして BERT [9] と RoBERTa [19] を使用した．使用した各モデルの詳細を付録 A の表 1 に示す．

マスク言語モデルの対数尤度は，PLL-word-l2r [10] で算出した．文の対数尤度を予測するのには，minicons [16] を使用した．

4.2 データ

英語の MPP データセットとして，BLiMP [4] と CrowS-Pairs [5]，COMPS [6] を使用した．また，英語以外の他言語の MPP データセットとして，JBLiMP [12] と French CrowS-Pairs [14]，Arabic MP [13] を使用した．各モデルにおける MPP データセットの A=U，A>U，A<U のデータ数を付録 B の表 2 に示す．

4.3 実験結果

トークン長の差と評価 MPP データセットにおけるトークン長の差が精度に対してどのような影響を与えるのか分析した．実験結果を図 2 に示す．結果として，容認可能な文が容認不可能な文よりも長いとき（A>U），モデルは容認不可能な文に対して高い対数尤度を付与しやすく，精度が大きく落ちてしまうことが判明した．一方で，容認可能な文が容認不可能な文よりも短いとき（A<U），モデルは容認可能な文に対して高い対数尤度を付与しやすく，精度が大きく上がることが示された．この現象はすべての MPP データセットにおいて共通であるため，構成している単語や文の構造の問題ではないことが示唆される．これらの結果から，ミニマルペアにおけるトークン長に差が存在する場合，文の容認性に関わらず，トークン長が短い文の方が高い対数尤度が割り当てられやすいことが確認された．このバイアスは，トークン化手法が評価結果に影響を与えるため，MPP データセットが言語モデルの言語能力を正確に評価することを妨げるという問題がある．したがって，現在利用可能な MPP データセットはトークン長バイアスに苦しんでおり，モデルの言語能力を正しく評価できていないと結論づける．

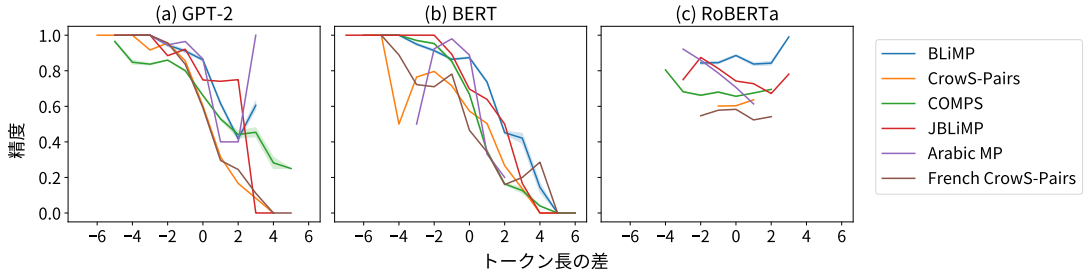

図 2　各モデルを MPP データセットで評価した際のトークン長の差と精度の変化を表した図．トークン長の差がプラスであることは，容認可能な文が容認不可能な文よりも長いこと（A>U）を示しており，トークン長の差がマイナスであることは，容認可能な文が容認不可能な文よりも短いこと（A<U）を示している．

また，GPT-2 と BERT モデルは全体的にトークン長バイアスの影響を受けやすかった．一方で，RoBERTa モデルでは影響が小さいことが示された．しかしながら，なぜ RoBERTa モデルがこのような特性を持つのかは不明であり，より詳細な検証は今後の課題である．

トークン長による正規化　セクション 3.2 で述べたように，トークン長で正規化した文対数尤度がトークン長バイアスを緩和できるか分析した．図 3 は，正規化した対数尤度を使用した場合のトークン長差と精度の変化を示す．MeanLP はトークン長バイアスを軽減するために用いられているが，トークン長に差が存在すると精度が変化することを示している．そのため，トークン長バイアスの影響は MeanLP では軽減できない．一方，PenLP の結果は，BLiMP と CrowS-Pairs などのデータセットにおいて，トークン長の差に関わらず精度が比較的一定であり，LP や MeanLP に比べてトークン長バイアスを軽減できている．しかし，COMPS ではトークン長の差に精度が影響されていることが確認される．従って，混乱を招くような評価結果をもたらす可能性があるため，PenLP は全ての MPP データセットにおいて一貫して使用できる手法ではない．上記の結果から，トークン長で正規化する手法は MPP データセットに用いることができないと結論づける．

5　MPP データセットの改善

トークン長バイアスを除去する手法として，ミニマルペアにおける容認可能な文と容認不可能な文のトークン長を揃える方法が考えられる．そこで本研究は，BLiMP データセットの改善を MPP データセットにおけるトークン長バイアスの除去のケーススタディとして行った．付録 C の表 3 は，元の BLiMP とバイアス除去した BLiMP を比較した実験

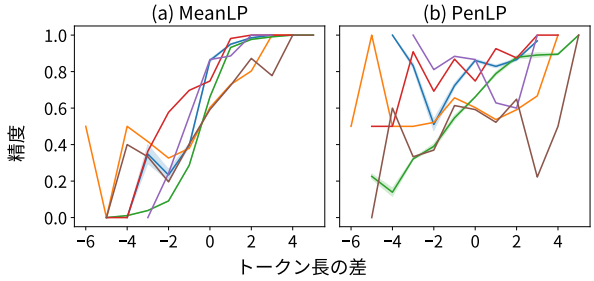

図 3　正規化した対数尤度を用いた場合の GPT-2 におけるトークン長の差と精度の変化を表した図．

結果である．全体として，モデル間の性能差が縮まっており，一部のサブセットではモデル性能の順位に変動が起きていることが確認できる．特に，“Expletive It Object Raising” の場合，元の BLiMP を用いた性能評価では，GPT-2 のモデルは 81.2 %であったのに対して，バイアス除去すると 90.1 %となっている．これは，トークン長バイアスによりモデル性能を過少評価してしまった例である．このことから，現状の MPP データセットでは，異なるトークン化手法ーを用いたモデルの性能を正しく比較することは不可能であり，トークン長を調整することにより公正に比較することが可能である．

6　おわりに

本研究では，MPP データセットにおけるトークン長バイアスについて分析した．実験の結果，MPP データセットでは容認可能な文と容認不可能な文においてトークン長に差が存在すると，精度に大きく影響するということが判明した．このことから，現在の単語単位での長さを揃える構築手法では不十分であり，トークン長単位で長さを揃える構築手法が必要であることを示した．今後は，対数尤度に代わる容認度の模索と RoBERTa のトークン長バイアスに関する頑健性に関する調査を検討する．

参考文献

[1] Noam Chomsky. **Syntactic Structures**. Mouton and Co., The Hague, 1957.

[2] C.T. Schutze. **The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology**. University of Chicago Press, 1996.

[3] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 625–641, 2019.

[4] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[5] 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.

[6] Kanishka Misra, Julia Rayz, and Allyson Ettinger. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 2928–2949, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[7] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 521–535, 2016.

[8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1(8):9, , 2019.

[9] 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.

[10] Carina Kauf and Anna Ivanova. A better way to do masked language model scoring. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 925–935, Toronto, Canada, July 2023. Association for Computational Linguistics.

[11] Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. SLING: Sino linguistic evaluation of large language models. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 4606–4634, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[12] Taiga Someya and Yohei Oseki. JBLiMP: Japanese benchmark of linguistic minimal pairs. In **Findings of the Association for Computational Linguistics: EACL 2023**, pp. 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[13] Wafa Abdullah Alrajhi, Hend Al-Khalifa, and Abdulmalik AlSalman. Assessing the linguistic knowledge in Arabic pre-trained language models using minimal pairs. In **Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)**, pp. 185–193, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[14] 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.

[15] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2699–2712, Online, July 2020. Association for Computational Linguistics.

[16] Kanishka Misra. minicons: Enabling flexible behavioral and representational analyses of transformer language models. **arXiv preprint arXiv:2203.13112**, 2022.

[17] Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. RuCoLA: Russian corpus of linguistic acceptability. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 5207–5227, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[18] Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 296–310, 06 2020.

[19] 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:2203.13112**, 2019.

A　使用したモデル

表 1　使用した言語モデルのトークン化手法や使用したデータセットの詳細．

言語	モデル	トークン化手法	語彙サイズ	使用したデータセット
英語	GPT-2 medium	BPE	50,256	BLiMP，CrowS-Pairs，COMPS
	BERT base (cased)	WordPiece	28,996	BLiMP，CrowS-Pairs，COMPS
	RoBERTa base	BPE	50,263	BLiMP，CrowS-Pairs，COMPS
日本語	japanese-gpt2-medium	Unigram	32,000	JBLiMP
アラビア語	Arabic GPT2	Unigram	64,000	Arabic MP
	AraBERT v2	Unigram	64,000	Arabic MP
フランス語	gpt2-wechsel-french	BPE	50,257	French CrowS-Pairs
	roberta-base-wechsel-french	BPE	50,265	French CrowS-Pairs
多言語	BERT multilingual base (cased)	WordPiece	119,547	JBLiMP，French CrowS-Pairs
	XLM-RoBERTa base	Unigram	119,547	JBLiMP，Arabic MP

B　各データセットにおけるデータ数

表 2　それぞれのデータセットのサブセットにおいてトークン長が容認可能な文と容認不可能な文で異なるようなミニマルペアを含む数．トークン化手法が異なるため，それぞれのモデルにおいてデータの数が異なる．

データセット	GPT-2			BERT			RoBERTa		
	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U
BLiMP	22,368	3,822	4,810	22,384	4,488	4,128	23,992	3,182	3,826
CrowS-Pairs	997	282	229	1,021	283	204	1,124	256	128
COMPS	29,592	16,727	16,917	24,691	19,584	18,961	20,608	6,784	35,844
JBLiMP	147	65	119	208	58	65	128	107	96
Arabic MP	702	251	547	852	174	474	480	476	544
French CrowS-Pairs	1132	309	235	855	495	326	1324	224	128

C　BLiMP データセットの改善

表 3　元の BLiMP とバイアス除去を行った BLiMP でモデルの性能評価した実験結果．トークン長バイアスによる影響が大きかった上位 10 個の文法現象サブセットを実験結果として載せている．

文法現象サブセット	元の BLiMP			バイアス除去した BLiMP		
	GPT-2	BERT	RoBERTa	GPT-2	BERT	RoBERTa
Animate Subject Passive	67.5	**79.1**	74.9	73.3	**82.4**	76.6
Causative	75.6	73.1	**83.6**	81.6	82.1	**86.1**
Drop Argument	58.7	58.8	**64.8**	60.1	**64.0**	63.7
Inchoative	65.3	59.8	**76.2**	72.5	73.3	**76.4**
Passive 2	88.7	**90.1**	**90.1**	91.3	92.8	**93.5**
Expletive It Object Raising	81.2	80.0	**81.5**	**90.1**	79.8	81.2
Tough vs. Raising 1	76.4	68.4	**87.6**	84.0	77.7	**87.5**
Det. Noun Agr. Irregular 1	93.4	97.7	**98.6**	98.0	**99.6**	99.2
Left Branch Island Echo Question	51.8	66.2	**69.7**	40.9	68.4	**69.1**
Matrix Question NPI Lic. Pres.	58.4	**92.7**	89.9	59.1	**94.3**	90.0