因果的プロンプトによる NLI の敵対的ロバスト性の強化

Pride Kavumba[1,2] Ana Brassard[2,1] Benjamin Heinzerling[2,1]
坂口 慶祐 [1,2] 乾 健太郎 [1,2]
[1] 東北大学 [2] 理化学研究所
kavumba.pride.q2@dc.tohoku.ac.jp
{ana.brassard, benjamin.heinzerling}@riken.jp
{keisuke.sakaguchi, kentaro.inui}@tohoku.ac.jp

概要

因果的プロンプトは，{*label*} *because* {*explanation*} というテンプレートを用いることで，与えられた入力に特定のラベルを割り当てるだけでなく，このラベルをサポートする説明を生成することができる．この種のプロンプトはもともとモデルの解釈可能性を向上させる目的で導入されたが，本論文では，因果的プロンプトが自然言語推論ベンチマークにおける敵対的摂動に対して，頑健性を向上させる効果があることを示す．

1 はじめに

因果的プロンプトは，言語モデルに与えられた入力に特定のラベルを割り当てるだけでなく，そのラベルをサポートする説明を言語モデルに生成させる．前提 “A soccer game with multiple males playing” と仮説 “Some men are playing a sport” を持つ自然言語推論（NLI）の例を見てみよう．predict-only の設定では，モデルは entailment のようなラベルを生成する必要がある．因果的プロンプトでは，モデルもラベルとこのラベルをサポートする説明，例えば “It is entailment because playing soccer is playing a sport” を生成しなければならない．

因果的プロンプトの当初の目的は，モデルの解釈可能性を向上させることであった [1]．この論文では，因果的プロンプトが敵対的な設定におけるモデル性能の向上にもつながる点を調査する．Adversarial NLI[2] のような敵対的ベンチマークは，SNLI[3] のような従来のベンチマークに含まれる表層的手がかり (superficial cue) による「ショートカット」要因を取り除いたものである．

敵対的ベンチマークにおける因果的プロンプトの利点を調査するにあたり，本論文では因果的プロンプトで自然言語推論タスクを解くために言語モデルを細かく設定し，その性能を Predict-only の設定と比較した．実験の結果，4 つの敵対的な NLI データセットと 2 つの非敵対的な NLI データセットにおいて，因果的プロンプトが一貫してパフォーマンスを向上させることを発見した (§4)．この性能向上は，異なるアーキテクチャサイズや異なるプロンプトのバリエーションにおいても一貫している．さらに分析を進めた結果，ラベルの具体的な言語化と説明とラベルの因果関係の両方が，モデルの性能に重要であることが明らかになった．最後に，Predict-only の設定においてモデルがショートカットを可能にする表層的なな手がかりに頼っていないことを確認した．

2 因果的プロンプト

自然言語推論（NLI）では，与えられた入力に entailment や contradiction といったラベルを割り当てるようにモデルが学習される [3]．NLI データセットはモデル性能を評価する観点で優れている．しかし，これまでの研究では，NLI や他のデータセットには情報があることが示されている [4, 5, 6]．例えば，Gururangan et al.[4] は SNLI において，「not」などの否定語が contradiction ラベルと強く結びついていることを発見した．したがって，入力に「not」が含まれる場合に contradiction と予測することが多いモデルは，高い精度を得ることができるが，実際には自然言語推論に関する言語能力を獲得していることにはならない．その結果，表層的な手がかりに依存するモデルは元々訓練されたデータセットでは高い精度を出すが，表層的な手がかりを含まない入力に対しては精度が低下するということが起きてしまう．

情報の問題に対処するためには，いくつかのアプ

ローチがある．一般的には，既存のデータセットから表面的な手がかりを削除するなど，データセットを慎重に作成するアプローチが多い．これは，敵対的フィルタリング（adversarial filtering）アドバーセーリエル フィルテリングによってサンプルを削除するか [7, 8, 9]，データセットを反実仮想的な例で補強することで実現できる [6, 10]．その他には敵対的トレーニングなどのトレーニング方法に着目したアプローチもある [11, 12, 13]．しかし，この方法では複雑な学習設定と高い計算量が必要になる．また，解説付きのマルチタスク学習をベースにしたアプローチもある [14]．

Predict-only パラダイムとは異なり，因果的プロンプトでは予測に対するフリーテキストの説明をモデルに生成させる必要がある．モデルの解釈性を向上させるために，Narang et. al. [1] は予測や自由記述の説明を求めるプロンプトを導入した．本書では，このパラダイムを因果的プロンプトと呼ぶことにする．

これまでの研究では，様々な条件下でモデルの解釈性を向上させるために，因果的プロンプトを採用する様々な方法が検討されてきた．[15] は因果的プロンプトの説明の忠実度を，[16, 17] はモデルから質の高い説明を抽出する方法を研究している．

このように，因果的プロンプトの解釈可能性は盛んに研究されているが，敵対的な頑健性については特に調査されていない．因果的プロンプトは元のラベルと説明文からなる出力形式を持つため，説明文を生成する際に表層的な手がかりがモデルによって利用される可能性は低いと推測される．したがって，我々は因果的プロンプトが敵対的頑健性に正の効果をもたらすと仮説を立て，以下の節でこの仮説を検証する．

3 手法

プロンプトフォーマット 各 NLI 問題は premise と hypothesis を提供し，モデルは entailment，neutral，contradiction の 3 つのラベルのうち 1 つを出力する必要がある．これに対して，以下のようなプロンプト形式を因果的プロンプトと呼ぶことにする．

- `INPUT`: Is this true and why? {premise} implies {hypothesis}
- `OUTPUT`: {Yes or No} it is {label} because {explanation}

ここでは，入力と出力の両方が（ラベルではなく）自然文になっており，タスクは自然言語の質問と回答という形式をとっている．予測は大きく 2 つのステップに分かれる．まず，与えられた前提と仮説に含意関係があるかという分類問題に対し entailment，neutral，contradiction のいずれかのラベルを予測する．そしてラベルに応じて，entailment の場合は “Yes it is ...” が，neutral と contradiction の場合は “No it is ...” が出力する．本稿ではこれを「マルチステップ・バーバライザー」と呼ぶ．これは，一段階または一語のバーバライザーを使用した従来の研究とは異なる．例えば，entailment は “yes”，contradiction と neutral は “no”，“maybe” と一語で表されることが多い [19]．最後に，予測の根拠を説明することで出力は完了する．本研究では，因果的プロンプトが敵対的頑健性に及ぼす効果に着目し，プロンプトが性能へ与える効果を検証した．

ベンチマーク e-SNLI [14], Adversarial NLI (ANLI) [2], SNLI Hard [4], NLI Diagnostic [20], Heuristic Analysis for NLI Systems (HANS) [21], Counterfactually-Augmented NLI (Counter-NLI) [10] でモデルを評価する．

トレーニングセットアップ モデルのトレーニングには，3/4 の e-SNLI とすべての ANLI データが使用される．なお，ANLI では一部の説明のみで，足りない部分はプロンプトテンプレートを修正して対応した．

4 実験結果

因果的プロンプトは敵対的ロバスト性を向上させるか？ 因果的プロンプトを用いた学習により，ほぼすべてのベンチマークですべてのモデルの性能が向上し，これまで報告されている State-of-the-art の精度を上回るケースも見られた (表 1)．T5-3B [22] はサイズが小さいにもかかわらず全体的に高い性能を示したが，これは T0 [23] がメモリの制限により，他のモデルで使用するバッチサイズとシーケンス長の 1/4 を使用したためである．また，HANS データセットは、T5 と T0 モデルにとって挑戦的であることが証明された。T5 と T0 モデルは、Subsequence(Sub) や Constituent(Cons) の敵対的攻撃に対してまだ脆弱な可能性がある。このモデルの Lexical Overlap (Lex) と subsequence、Constituent の格差は，さらなる分析のための興味深い道であると思われるが本稿の範囲外である．他のすべてのベンチマークにおいて，因果的プロンプトモデルは，ドメ

		Current	T5-Small (60M)		BART-Base (125M)		BART-Large (400M)		T5-Large (770M)		T5-3B (3B)		T0* (11B)	
		SOTA	PR	EXP	PR	EXP	PR	EXP	PR	EXP	PR	EXP	PR	EXP
e-SNLI		92.3	82.4	88.8	88.7	92.1	90.4	93.8	90.9	94.4	91.7	**95.1**	91.0	91.9
SNLI Hard		80.7	68.5	82.2	78.1	84.3	81.5	84.9	82.1	88.7	84.0	**89.7**	83.0	84.5
ANLI	R1	75.5	46.5	52.5	56.8	53.0	64.9	65.9	66.1	77.2	74.9	**81.8**	69.6	75.6
	R2	51.4	37.6	56.4	41.5	50.3	44.4	57.1	49.2	67.8	58.9	**72.5**	53.7	60.6
	R3	49.8	40.4	59.1	40.9	54.0	46.5	59.6	49.4	68.0	57.9	**74.8**	55.0	59.9
HANS	Lex	94.1	2.6	0.0	71.2	69.6	85.0	90.2	82.9	81.3	94.2	94.2	**97.9**	95.9
	Sub	46.3	2.2	0.0	43.2	54.1	27.3	**63.7**	35.6	27.6	46.3	30.3	20.5	37.9
	Cons	38.5	2.5	0.0	34.7	51.9	22.4	**63.8**	19.6	9.9	38.6	17.1	24.3	53.9
Counter-NLI	RP	54.3	54.1	75.6	59.8	74.9	66.1	77.3	67.8	82.3	69.6	**83.0**	66.5	69.2
	RH	74.3	78.4	86.5	82.9	87.8	85.3	87.4	86.5	92.4	88.9	**93.5**	87.9	87.4
	RP&RH	64.3	66.3	81.1	71.3	81.3	75.7	82.3	77.1	87.3	79.3	**88.3**	77.2	78.3
NLI Diagnostic	Know	53.9	34.5	58.8	41.2	60.2	57.4	70.4	54.9	65.8	58.8	**76.4**	58.8	59.9
	Logic	58.7	45.3	59.6	45.6	67.0	54.9	67.0	57.4	70.3	63.7	**73.9**	60.7	64.5
	LS	66.5	49.5	63.3	49.2	62.2	62.2	69.6	63.9	76.1	69.6	**79.3**	63.0	70.4
	PAS	69.9	58.0	69.3	55.7	65.3	67.9	66.7	71.0	76.4	73.1	**80.9**	70.8	72.4

表 1: predict-only (PR) と因果的プロンプト (EXP) でトレーニングされたモデルの平均予測精度． 現在の最新技術は WT5 [1], BERT-Sup-ATT [18], InfoBERT [12], RoBERTa-AFLITE [9], BERT [10], RoBERTa-AFLITE [9] e-SNLI、ANLI、および SNLI-Hard は、ドメイン内のテスト セットである．

イン内設定 と敵対的ドメイン外設定の両方に対して明確な改善を示している．このように，因果的プロンプトは敵対的攻撃に対するモデルの頑健性を概ね向上させ，全体として NLI 予測性能を向上させたと結論付けることができる．

結果はアーキテクチャ/サイズに依存するか？ モデルサイズが性能に与える影響を調べるため，パラメータが 6,000 万から 30 億までの 6 種類のモデルを評価した．これらのモデルには，125M と 400M のパラメータを持つ BART [24] の 2 つのバリエーションと，60M，770M，3B のパラメータを持つ T5 の 3 つのバリエーションが含まれている．表 1 からわかるように，ANLI データセットでは，モデルサイズと性能の間に明確な関係があり，より大きなモデルがより良い結果を出していることがわかる．

モデルは表層的手がかりの影響を受けているか？ モデルが表層的な手がかりの影響を受けているか調べるため，先行研究に従って仮説のみで学習したモデルの性能を比較した [4]．仮説のみで学習する、つまり、前提がない不完全なタスク設定になるため，表層的な手がかりの影響を受けないモデルはランダムな精度になると考えられる．

本実験ではラベル予測のみを学習し他モデルと因果的プロンプトで学習したモデルを比較した．その結果，predict-only モデルはランダム性能を上回っている（63.7% vs 33.3% ）一方、因果的プロンプトモデルはランダムな精度になっており，表面的な手がかりの影響を受けていないことがわかった．

因果的プロンプトはトークンとラベルの関連性を弱めるか？ 入力の単語と学習セットのラベルの関係を理解するために，それらの間の自己相互情報量 (PMI) を計算する．これにより，特定のラベルと強く結びついている言葉や，因果的プロンプトを使用したときにその結びつきがどのように変化するかを特定することが可能になる．

$$PMI(word, label) = log \frac{p(word, label)}{p(word, \cdot)p(\cdot, label)}$$

Gururangan et al. [4] と同様に，ラベルと最も強く関連する単語を強調するため，PMI の計算には add-100 スムージングを適用した．

実験の結果，因果的プロンプトを用いることで入力語とラベルの関連性が低下することがわかった．例えば，「frowning」というネガティブな単語は contradiction というラベルと強く結びついているが，因果的プロンプトを使うとこの関連はなくなる (付録 A，図 1)．この結果は，仮説のみのモデルの結果とも一致する．

モデルはプロンプトの微小な違いに敏感か？ これまでの研究で，モデルはプロンプトに対して非常に敏感であることが示されている [19] ため，我々もクラウドソーシングのプロンプトソースプロジェ

Dataset	Prompt ID 1	2	3	4	5	Mean$_{(stddev)}$
e-SNLI	91.5 / **94.4**	91.6 / **94.7**	91.8 / **94.6**	91.6 / **94.5**	91.8 / **94.4**	91.7$_{(0.1)}$ / **94.5**$_{(0.1)}$
CNLI (RP)	70.8 / **82.5**	73.0 / **83.3**	71.5 / **83.0**	72.1 / **83.0**	70.1 / **82.8**	71.5$_{(1.1)}$ / **82.9**$_{(0.3)}$
CNLI (RH)	82.0 / **92.3**	82.8 / **93.0**	81.9 / **92.8**	82.0 / **92.2**	83.1 / **92.5**	82.4$_{(0.6)}$ / **92.6**$_{(0.3)}$
CNLI (RP&RH)	76.4 / **87.4**	77.9 / **88.4**	76.7 / **87.9**	77.0 / **87.6**	76.6 / **87.7**	76.9$_{(0.6)}$ / **87.8**$_{(0.4)}$

表 2: e-SNLI と CounterNLI (CNLI) の開発セットに関するプロンプトの感度．数値は predict-only/因果的プロンプトの平均精度である．因果的プロンプトの標準偏差が低いほど，安定性が高いことを示している．

Explanation	Accuracy	BLEU
Random characters	21.00	0.02
Random words	0.00	0.90
Low-sim. sentences	0.04	0.03
High-sim. sentences	59.1	1.73
None	88.4	-
Original (e-SNLI)	**91.6**	36.1

表 3: 切除した説明文を用いて学習させた T0 モデルの平均予測精度

クトから入手した 5 種類のプロンプトを用いて比較を行った．各モデルについて，3 種類のランダムシードで 3 回の実験を行った．なおリソースの制限から，この実験では T0 を除外し，e-SNLI からランダムにサンプリングした 2 万個のインスタンスのみを使用した．

実験の結果，因果的プロンプトモデルはより優れた汎化性を示し，敵対的な攻撃に対してより頑健である (表 2) ことがわかった．

説明には因果関係が必要か？ 説明における因果関係の有無の影響を調べるため，因果的プロンプトの説明文を，全くランダムな文から類似の文まで，無関係の文に置き換えたモデルで実験を行った．具体的には，(i) ランダムな文字，(ii) ランダムな単語，(iii) BookCorpus [25] の低類似度文，(iv) BookCorpus の高類似度文の設定を比較し，元の説明文との類似度は SentenceBERT [26] を用いた．すべてのモデルは e-SNLI からランダムにサンプリングした 20K 個のインスタンスで学習させた．実験の結果，元の説明文を用いた生成が最も高い精度を示した．ランダムな説明や無関係な説明では性能が低下するため，因果関係のある説明を予測するようにモデルを訓練することで，敵対的な頑健性が向上することが確認された (表 3)．

因果的プロンプトはモデルのパフォーマンスを向上させるか？ このことを確認するために，1 段階の発話によるプロンプト ({**label**} **because** {**explanation**}) と多段階の発話によるプロンプト (**Yes/No it is** {**label**} **because** {**explanation**}) を用いて，説明を加えた場合と加えない場合の両方を T0 に学習させた．predict-only モデルは，シングルステップ・バーバライザーで 87.2%，マルチステップ・バーバライザーで 88.4%の精度を達成した．一方，因果的プロンプトモデルはシングルステップ・バーバライザーで 90.9%，マルチステップ・バーバライザーで 91.6%という精度を達成した．predict-only と因果的プロンプトのいずれの設定でも，多段階のバーバライザーを追加することで，1 段階のものよりも改善が見られた．

5 おわりに

本研究では，因果的プロンプトが自然言語処理モデルの敵対的頑健性に与える影響について検討した．実験の結果から，因果関係のプロンプトを使用することで，敵対的な攻撃に対するモデルの頑健性を向上させることができることが示された．具体的には，因果的プロンプトモデルは，(1) 表層的な手がかりによる影響を受けなくなったこと，(2) 因果関係の説明や多段階の言語化で最も効果的であること，(3) プロンプト形式の違いに頑健であること，(4) モデルサイズが大きくなると性能が向上すること，そして (5) 因果的プロンプトの使用により入力単語とラベルの関連性が低下すること，を示した。

6 謝辞

本研究は，JST，CREST，JPMJCR20D2 の支援を受けたものである．本研究は JSPS 科研費 JP21K21343,

21K17814 の助成を受けたものです.

参考文献

[1] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions, 2020.

[2] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In **58th Annual Meeting of the ACL**. ACL, 2020.

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In **2015 EMNLP**, pp. 632–642, Lisbon, Portugal, September 2015. ACL.

[4] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In **2018 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 2 (Short Papers)**, pp. 107–112, New Orleans, Louisiana, June 2018. ACL.

[5] Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. What makes reading comprehension questions easier? In **2018 EMNLP**, pp. 4208–4219, Brussels, Belgium, October-November 2018. ACL.

[6] Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. When choosing plausible alternatives, clever hans can be clever. In **First Workshop on Commonsense Inference in Natural Language Processing**, pp. 33–42, Hong Kong, China, November 2019. ACL.

[7] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In **2018 EMNLP**, pp. 93–104, Brussels, Belgium, October-November 2018. ACL.

[8] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In **2018 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 2 (Short Papers)**, pp. 8–14, New Orleans, Louisiana, June 2018. ACL.

[9] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In Hal Daumé III and Aarti Singh, editors, **37th International Conference on Machine Learning**, Vol. 119 of **PMLR**, pp. 1078–1088. PMLR, 13–18 Jul 2020.

[10] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually augmented data. **ICLR**, 2020.

[11] Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. On adversarial removal of hypothesis-only bias in natural language inference. In **Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)**, pp. 256–262, Minneapolis, Minnesota, June 2019. ACL.

[12] Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. Infobert: Improving robustness of language models from an information theoretic perspective. In **ICLR**, 2021.

[13] Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. Adversarial training for large neural language models, 2020.

[14] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, **Advances in Neural Information Processing Systems 31**, pp. 9539–9549. Curran Associates, Inc., 2018.

[15] Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In **2021 EMNLP**, pp. 10266–10284, Online and Punta Cana, Dominican Republic, November 2021. ACL.

[16] Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. Few-shot self-rationalization with natural language prompts. In **Findings of the ACL: NAACL 2022**, pp. 410–424, Seattle, United States, July 2022. ACL.

[17] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In **2022 Conference of the North American Chapter of the ACL: Human Language Technologies**, pp. 632–658, Seattle, United States, July 2022. ACL.

[18] Joe Stacey, Yonatan Belinkov, and Marek Rei. Supervising model attention with human explanations for robust natural language inference, 2021.

[19] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In **2021 Conference of the North American Chapter of the ACL: Human Language Technologies**, pp. 2339–2352, Online, June 2021. ACL.

[20] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. ACL.

[21] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In **57th ACL**, pp. 3428–3448, Florence, Italy, July 2019. ACL.

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **JMLR**, Vol. 21, No. 140, pp. 1–67, 2020.

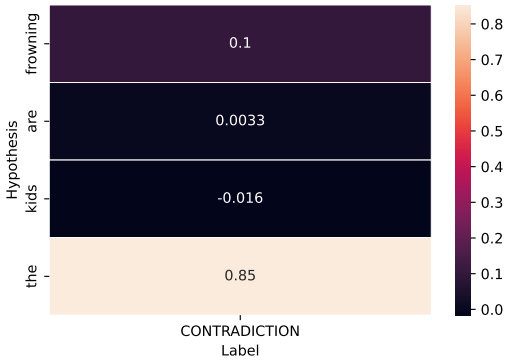
[23] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2021.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **58th Annual Meeting of the ACL**, pp. 7871–7880, Online, July 2020. ACL.

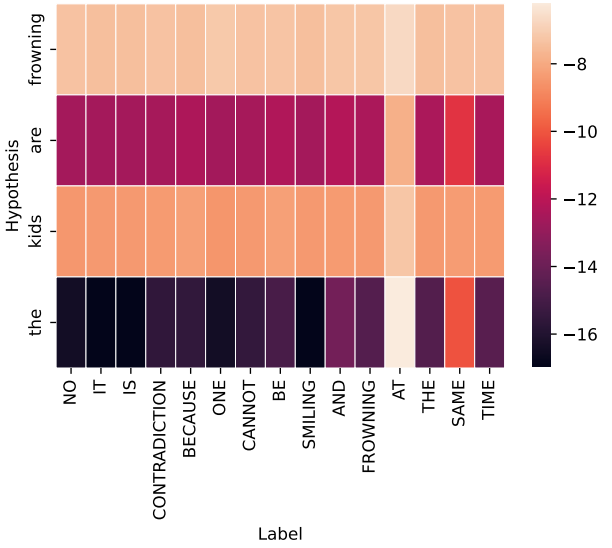
[25] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In **The IEEE International Conference on Computer Vision (ICCV)**, December 2015.

[26] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **2019 EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. ACL.

A PMI



(a) Predict-only の PMI



(b) 因果的プロンプトの PMI

図 1: 仮説語句の PMI 統計．(a) Predict-only 設定での PMI．(b) 因果的プロンプト設定での PMI． 仮説に含まれる単語は、Predict-only ラベルと強く結びついている． 因果的プロンプトでは、仮説語とラベルの関連は負の関連に低下する．

入力の単語と学習セットのラベルの関係を理解するために，それらの間の自己相互情報量 (PMI) を計算する．これにより，特定のラベルと強く結びついている言葉や，因果的プロンプトを使用したときにその結びつきがどのように変化するかを特定することが可能になる．

$$PMI(word, label) = log \frac{p(word, label)}{p(word, \cdot)p(\cdot, label)}$$

Gururangan et al.[4] と同様に，ラベルと最も強く関連する単語を強調するため，PMI の計算には add-100 スムージングを適用した．

実験の結果，因果的プロンプトを用いることで入力語とラベルの関連性が低下することがわかった．例えば，「frowning」というネガティブな単語は contradiction というラベルと強く結びついているが，因果的プロンプトを使うとこの関連はなくなる (図 1)．この結果は，仮説のみのモデルの結果とも一致する．