質問応答モデルはどのショートカットを優先して学習するか？

篠田一聡 [1,2*] 菅原朔 [2] 相澤彰子 [1,2]
[1] 東京大学 [2] 国立情報学研究所
shinoda@is.s.u-tokyo.ac.jp {saku,aizawa}@nii.ac.jp

概要

読解のための質問応答モデルは，訓練セット内の擬似相関を利用した解き方であるショートカットを学習する傾向がある．この問題を緩和するために様々な手法が提案されてきたが，それらはショートカット自体の特性を十分に考慮していない．本研究ではショートカットの学習可能性（ショートカットをどれくらい学習し易いか）が緩和手法の設計に有用であるという仮説を立て，実験的に検証する．

1 はじめに

微調整された事前学習済み言語モデルは，本来の解き方ではなく訓練セット内の擬似相関を利用したショートカットを学習しやすい [1]．この傾向はショートカットが有効な“ショートカット例”への汎化に寄与する一方，ショートカットが無効な“反ショートカット例”への汎化を阻害する [2, 3]．

読解のための質問応答モデルは複数の種類のショートカットを学習しうることが示唆されている [4, 5, 6]．この問題を緩和するために，データ拡張 [7, 8] や損失関数 [6, 9, 10] 等の緩和手法が提案されてきたが，**既存手法はショートカットの種類に応じた特性を十分に考慮していない**．

本研究ではショートカットの学習可能性（ショートカットをどれくらい学習し易いか）は緩和手法の設計に有用であると仮説を立てる．仮説検証のために，まず抽出型と多肢選択型読解における代表的なショートカットの学習可能性を行動テスト (§3.1)，質的分析 (§3.2)，量的分析 (§3.3) によって比較する．そして各ショートカットの学習の緩和に必要な反ショートカット例の割合を調べる (§3.4)．実験結果より，学習し易いショートカットほど学習の緩和に必要な反ショートカット例の割合が少ないことを示す．よって，**緩和手法を設計する際にショートカットの学習可能性を考慮すべきだと主張する**．

* 現在は NTT 人間情報研究所に所属する

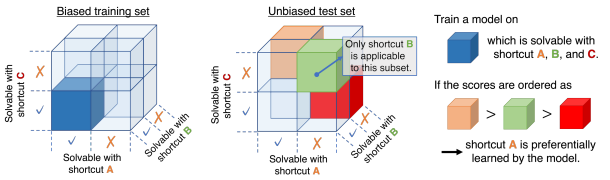

図 1 質問応答モデルがどのショートカットを優先して学習するかを明らかにするための行動テスト (§3.1) の図解．

2 ショートカット

実験 (§3) のために各ショートカット k に対してデータ集合 $\mathcal{D}$ をショートカット例 $\mathcal{D}_k$ と反ショートカット例 $\overline{\mathcal{D}_k}$ に分割するルールを定義する．

2.1 抽出型読解

文脈から回答を抽出して質問に答える抽出型読解では，以下の 3 つのショートカットを分析する．

Answer-Position [6] 文脈の最初の文から回答を抽出するショートカット．回答が最初の文に含まれない時に無効とする．(k = Position)

Word Matching [5] 質問と最も語彙的に類似した文[1)]から回答を抽出するショートカット．回答がそれ以外の文に含まれる時に無効とする． (k = Word)

Type Matching [11] 質問から回答の固有表現タイプを予測して対応するスパンを抽出するショートカット．回答と同じタイプの固有表現が文脈内で 2 つ以上含まれる場合に無効とする．[2)3)] (k = Type)

2.2 多肢選択型読解

文脈と質問をもとに複数の選択肢から回答を選択する多肢選択型読解では，NLI の研究に倣い，2 つのショートカットを定義して分析した．[4)]

Word-label Correlation 多肢選択型読解タスクでは選択肢のみから正しい回答を予測できる例が多

1) ここでは質問と共通の最長の n-gram を含む文と定義する．
2) 固有表現でない回答は分析の対象から除外した．
3) 固有表現抽出には spaCy (https://spacy.io/) を用いた．
4) 多肢選択型読解と自然言語推論は，文脈と質問 (前提) をもとに選択肢 (仮説) が正しいかを予測する点で類似している．

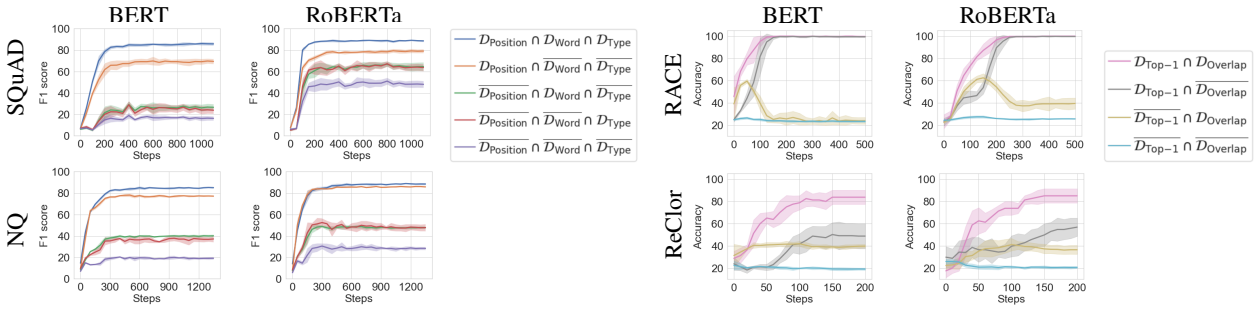

図 2　抽出型（左）と多肢選択型読解（右）のテストセットの各サブセットにおける訓練中の精度の推移．5 つのランダムシードでの平均 ± 標準偏差を折れ線で示す．

い [12, 13] ことから，仮説のみから正しい予測ができる例が多い NLI [14] と同様，ラベルとの間に擬似相関がある単語が選択肢に含まれると仮定した．Gardner ら [15] の主張[5] に従い，z-statistic [15] を用いて最もラベルと相関する単語を簡単のために 1 つ特定する．詳細は付録 A を参照のこと．ここで，特定された単語が正解選択肢に含まれるときにショートカットが有効だと定義する．(k = Top-1)

Lexical Overlap　NLI モデルは語彙の重複に基づくショートカットを学習し易い [2] ため，多肢選択型読解でも同様のショートカットが学習され易いと仮定する．ここでのショートカットは，文脈・質問と語彙の重複[6] が最大となる選択肢が正解であるときに有効とする．(k = Overlap)

3　実験

データセット　抽出型読解では SQuAD 1.1 [16] と NaturalQuestions (NQ) [17]，多肢選択型読解では RACE [18] と ReClor [13] を用いた．RACE と ReClor では選択肢のみからラベルを予測するモデルがランダムよりも良い精度を出しており [12, 13]，選択肢とラベルの間に擬似相関があると考えられる．

モデル　エンコーダには抽出型・多肢選択型読解で広く採用されている BERT-base [19] と RoBERTa-base [20] を用いた．タスクごとに出力層をエンコーダの上に追加する．抽出的読解ではモデルは文脈内の回答スパンの始点と終点の確率分布を出力する．多肢選択型読解では 4 つの選択肢から正しい選択肢の確率分布を予測する．訓練ステップ以外は，既存研究のハイパーパラメータに従った．[7]

3.1　行動テスト

RQ1:　各ショートカットが訓練セットのすべての質問で有効な時，質問応答モデルはどのショートカットを優先して学習するか？

この質問に答えるため，全てのショートカットが有効な訓練セットで訓練し，いずれかのショートカットのみが有効なテストセットで評価を行う行動テスト（図 1）を行った．この訓練セットではショートカット例の頻度が平等なため，学習可能性がモデルの予測に与える純粋な影響を比較できる．

設定　まず訓練セットから抽出した $\mathcal{D}_{\text{Position}} \cap \mathcal{D}_{\text{Word}} \cap \mathcal{D}_{\text{Type}}$ または $\mathcal{D}_{\text{Top-1}} \cap \mathcal{D}_{\text{Overlap}}$ でモデルを微調整した．次にテストセットから抽出した $\mathcal{D}_{\text{Position}} \cap \overline{\mathcal{D}_{\text{Word}}} \cap \overline{\mathcal{D}_{\text{Type}}}$ などのサブセットで評価し，どのショートカットが優先的に学習されたかを明らかにした．ショートカットの学習過程を明らかにするため，訓練中のスコアの推移も報告する．

結果：抽出型読解　結果は図 2（左）の通り．一貫して訓練中は $\mathcal{D}_{\text{Position}} \cap \overline{\mathcal{D}_{\text{Word}}} \cap \overline{\mathcal{D}_{\text{Type}}}$ でのスコアが $\overline{\mathcal{D}_{\text{Position}}} \cap \mathcal{D}_{\text{Word}} \cap \overline{\mathcal{D}_{\text{Type}}}$ と $\overline{\mathcal{D}_{\text{Position}}} \cap \overline{\mathcal{D}_{\text{Word}}} \cap \mathcal{D}_{\text{Type}}$ よりも高かった．つまり Position ショートカットが最も学習されやすかった．また $\overline{\mathcal{D}_{\text{Position}}} \cap \mathcal{D}_{\text{Word}} \cap \overline{\mathcal{D}_{\text{Type}}}$ や $\overline{\mathcal{D}_{\text{Position}}} \cap \overline{\mathcal{D}_{\text{Word}}} \cap \mathcal{D}_{\text{Type}}$ では $\overline{\mathcal{D}_{\text{Position}}} \cap \overline{\mathcal{D}_{\text{Word}}} \cap \overline{\mathcal{D}_{\text{Type}}}$ よりもスコアが高いことから，モデルは複数のショートカットを複合的に学習することがわかる．

結果：多肢選択型読解　結果は図 2（右）の通り．訓練終了時，Overlap ショートカットよりも Top-1 ショートカットが一貫して優先的に学習された．興味深いことに，Overlap の学習は，訓練の初期で Top-1 よりも優先された．このことから，単語とラベルの相関を認識するためには数百ステップの訓練が必要な一方，Transformer [21] に基づく言語モデルは自己注意機構を介して語彙の重複を認識しやすい帰納バイアスがある可能性がある．

5)　言語理解タスクにおいて入力の単一の特徴がラベルの情報を含むべきではないという主張．
6)　選択肢の単語数に対する文脈・質問にも含まれる共通の unigram の数の比率と定義する．
7)　本研究で使用したコードは一般に公開している．https://github.com/KazutoshiShinoda/ShortcutLearnability

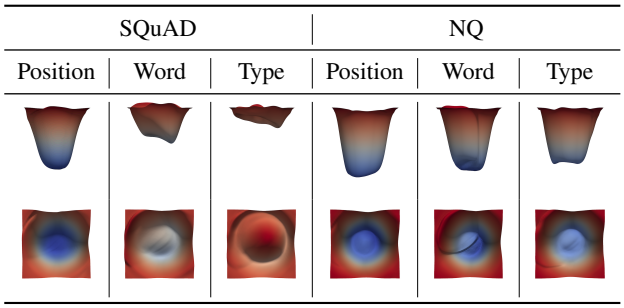

図 3　抽出型読解における各ショートカットの周辺の損失関数の可視化．X・Y 方向はパラメータ空間内でランダムにサンプリングした．曲面を横と上から見た図を示す．

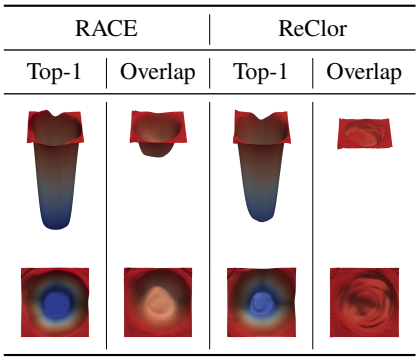

図 4　多肢選択型読解における各ショートカットの周辺の損失関数の可視化．可視化の方法は図 3 と同様．

3.2　質的分析

RQ2:　なぜ特定のショートカットが他のショートカットよりも優先して学習されるのか？

画像分類において Scimeca ら [22] が行ったように，損失関数の可視化によってこの質問に答える．具体的には各ショートカットを学習させたモデルパラメータの周辺の損失を可視化する．ショートカット間で損失曲面の平坦さ[8] と深さを比較し，選好性について洞察を得ることを目的とする．

設定　まず各ショートカットを独占的に学習したモデルを用意する．これはそのショートカットだけが有効なサブセットで訓練することで得られると仮定する．例えば $\mathcal{D}_{\text{Position}} \cap \overline{\mathcal{D}_{\text{Word}}} \cap \overline{\mathcal{D}_{\text{Type}}}$ で訓練したモデルは Position ショートカットを学習する可能性が高い．テストセットにおいて訓練セットと同じサブセットでモデルが最も高いスコアを達成したため，この仮定は支持された．§3.1 で訓練に使用した $\mathcal{D}_{\text{Position}} \cap \mathcal{D}_{\text{Word}} \cap \mathcal{D}_{\text{Type}}$ と $\mathcal{D}_{\text{Top-1}} \cap \mathcal{D}_{\text{Overlap}}$ で損失を計算し，[24] に従って可視化をした．紙面の関係で BERT-base の結果のみを報告する．

結果　結果を図 3 と 4 に示す．曲面の中央が各ショートカットを学習したモデルに対応する．§3.1 で優先して学習されたショートカット（Position と Top-1）は，より平坦で深い曲面に位置する傾向があることがわかった．損失曲面の平坦さと深さの順序は，§3.1 の行動テストにおけるショートカットの学習の優先順序とほぼ相関している．損失曲面の平坦さと汎化性能には相関がある [25] ことから，これらの観察は，$\mathcal{D}_{\text{Position}} \cap \mathcal{D}_{\text{Word}} \cap \mathcal{D}_{\text{Type}}$ と $\mathcal{D}_{\text{Top-1}} \cap \mathcal{D}_{\text{Overlap}}$ で学習したモデルがそれぞれ Position と Top-1 を優先して学習した理由を説明できる．

8）　平坦さの定義は，損失がほぼ一定に保たれるパラメータ空間の連結領域の大きさ [23] とする．

3.3　量的分析

RQ3:　各ショートカットの学習可能性は定量的にどの程度違うか？

この質問に答えることでショートカットの選好性を定量的に説明することを目指す．そのために，あるショートカットのみが適用可能なサブセット（例えば $\mathcal{D}_{\text{Position}} \cap \overline{\mathcal{D}_{\text{Word}}} \cap \overline{\mathcal{D}_{\text{Type}}}$）についてラベルの最小記述長（MDL）[26] を推定し，ショートカットごとに比較した．この方法を MDL 原理の生みの親に因んで Rissanen Shortcut Analysis (RSA) と名付ける．直感的には RSA は訓練セットにおいて各ショートカットが利用可能なことがどの程度タスクを学習しやすくするかを定量的に比較できる．

設定　既存研究 [27, 28] に従って MDL を推定するために Online Code [29] を用いた．Online Code は直感的には訓練損失の曲線の下の面積の計算に相当する．詳細は付録 B を参照のこと．

結果　結果を表 1 に示す．一部の例外を除き，SQuAD 1.1 と NQ では Position ショートカットが有効な時に最も学習しやすいことがわかった．この例外は §3.1 で示すように SQuAD で RoBERTa が Word ショートカットを学習しやすいためだと考えられる．RACE と ReClor では Top-1 ショートカットによって Overlap よりも低い MDL が得られた．これらの結果は一部の場合を除き，行動テスト（図 2）や可視化（図 3 と 4）の結果と一貫している．[9] また RoBERTa は全ての条件で BERT と比較して MDL を低下させた．図 2 で RoBERTa は BERT よりも反ショートカット例に対して頑健であったことから，MDL はショートカットの特性だけでなくモデル元来の汎化能力も反映していると考えられる．

9）　Online Code は式 2 に示すようにデータセットのサイズに依存するため，データセット間で比較することはできない．

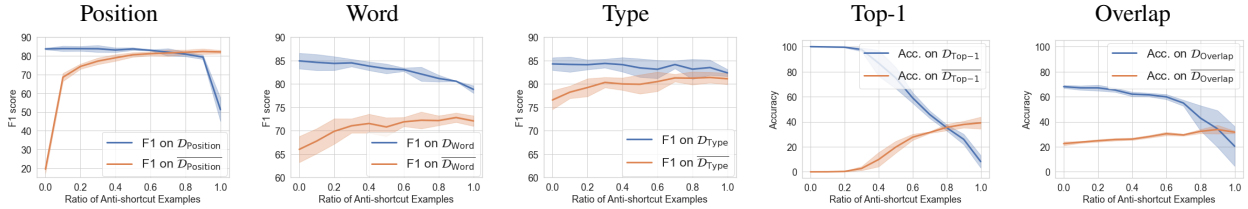

図 5　SQuAD（左 3 つ）と RACE（右 2 つ）のテストセットにおける $\mathcal{D}_k$（青）と $\overline{\mathcal{D}_k}$（橙）での精度．横軸は訓練セットにおける $\overline{\mathcal{D}_k}$ の割合．訓練セットのサイズは一定に保った．5 つのランダムシードでの平均 ± 標準偏差を折れ線で示す．

3.4　反ショートカット例の割合の制御

RQ4:　ショートカットの学習を避けるために反ショートカット例の割合はどの程度必要か？それはショートカットの学習可能性と関係があるか？

ショートカット学習を緩和する最も単純な方法の 1 つは，反ショートカット例を手動または自動で訓練セットに追加して，データセットのバイアスを軽減することである [8, 30]．なぜなら訓練セットの大半で何らかのショートカットが有効な場合，そのショートカットを学習したモデルは低い訓練損失を達成できるためである．また反ショートカット例の割合とショートカットの手がかりの抽出可能性がショートカット学習の要因であると文法タスクで報告されており [30]，読解タスクにおいても同様の傾向があると考えられる．もしあるショートカットを学習を回避するために反ショートカット例がどの程度必要か分かれば，質問応答モデルに頑健な解き方を学習させるために最適な訓練セットの構築やデータ拡張手法の設計に役立てられる．

結果　結果を図 5 に示す．訓練セットがショートカット例のみで構成されている場合（x 軸の値が 0 の時），$\mathcal{D}_k$ と $\overline{\mathcal{D}_k}$ のスコアの差分は全てのケースで有意である．Position，Top-1，Overlap において反ショートカット例の割合が 0.7, 0.8, 0.9 のときに $\mathcal{D}_k$ と $\overline{\mathcal{D}_k}$ の間の精度差が解消され，ショートカットの学習を緩和できていると捉えられる．

§3.1,3.2,3.3 の結果を考慮すると，より学習可能なショートカットほどショートカット学習の緩和により少ない割合の反ショートカット例を必要とすることが分かる．さらに Word や Type のような学習しにくいショートカットでは，反ショートカット例の割合を制御するだけではショートカット学習を緩和するには不十分であることを示唆している．学習しにくいショートカットに関してさらに精度差を緩和するためには，損失関数 [31] やモデル構造の改善 [32, 33] をする必要があることを示唆している．

Dataset	Shortcut	BERT	RoBERTa
SQuAD	Position	4.65 ± 0.12	4.22 ± 0.23
	Word	4.94 ± 0.24	3.73 ± 0.17
	Type	5.75 ± 0.30	4.52 ± 0.06
NQ	Position	6.28 ± 0.15	5.37 ± 0.24
	Word	12.24 ± 0.14	9.08 ± 0.20
	Type	11.76 ± 0.55	8.83 ± 0.38
RACE	Top-1	0.52 ± 0.34	0.41 ± 0.29
	Overlap	4.16 ± 0.55	3.55 ± 0.10
ReClor	Top-1	0.33 ± 0.07	0.28 ± 0.03
	Overlap	0.55 ± 0.03	0.52 ± 0.02

表 1　各ショートカットのみが有効なサブセットで推定した最小記述長 (kbits)．5 つのランダムシードでの平均 ± 標準偏差を報告する．

4　おわりに

代表的なショートカットの学習可能性を一連の実験で比較することにより，抽出型と多肢選択型読解におけるショートカット学習の理解を深めた．つまり特定のショートカット（Position と Top-1）が学習されやすく，これらはパラメータ空間においてより平坦で深い損失曲面に位置する傾向があり，情報理論的にタスクを学習しやすくすることがわかった．そしてショートカットの学習可能性は学習の緩和に必要な反ショートカット例の比率と相関していることがわかった．このことから，学習可能性は緩和手法の設計に有用だと主張する．

本研究では読解タスクに特化したパラメータ更新を伴うモデルを扱った．一方，LLM の文脈内学習はタスクごとのパラメータ更新を伴わないため，訓練セットに特有の擬似相関を学習しない利点があると GPT-3 の論文では主張されている [34]．しかし，近年 LLM の文脈内学習においても，プロンプトで与える例群における偏りや擬似相関が LLM による推論時のショートカットの使用と精度劣化を引き起こすという報告もある [35, 36]．文脈内学習に限らず，LLM が擬似相関を利用する原理の理解 (e.g., [37]) とその抑制 (e.g., [38]) は今後の課題である．

謝辞

本研究は JSPS 科研費 JP21H03502, 22J13751, 22K17954, NEDO SIP-2 “Big-data and AI-enabled Cyberspace Technologies” の助成を受けたものです.

参考文献

[1] Robert Geirhos et al. Shortcut learning in deep neural networks. **Nature Machine Intelligence**, Vol. 2, No. 11, pp. 665–673, November 2020.

[2] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In **ACL**, 2019.

[3] Matt Gardner et al. Evaluating models' local decision boundaries via contrast sets. In **Findings of EMNLP**, 2020.

[4] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In **EMNLP**, 2017.

[5] Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. What makes reading comprehension questions easier? In **EMNLP**, 2018.

[6] Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. Look at the first sentence: Position bias in question answering. In **EMNLP**, 2020.

[7] Kazutoshi Shinoda et al. Improving the robustness of QA models to challenge sets with variational question-answer pair generation. In **ACL SRW**, 2021.

[8] Kazutoshi Shinoda et al. Can question generation debias question answering models? a case study on question–context lexical overlap. In **MRQA Workshop**, 2021.

[9] Mingzhu Wu et al. Improving QA generalization by concurrent modeling of multiple biases. In **Findings of EMNLP**, 2020.

[10] Kazutoshi Shinoda et al. Look to the right: Mitigating relative position bias in extractive question answering. In **BlackboxNLP Workshop**, 2022.

[11] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Making neural QA as simple as possible but not simpler. In **CoNLL**, 2017.

[12] Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. In **AAAI**, 2020.

[13] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In **ICLR**, 2020.

[14] Suchin Gururangan et al. Annotation artifacts in natural language inference data. In **NAACL**, 2018.

[15] Matt Gardner et al. Competency problems: On finding and removing artifacts in language data. In **EMNLP**, 2021.

[16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **EMNLP**, 2016.

[17] Tom Kwiatkowski et al. Natural questions: A benchmark for question answering research. **TACL**, Vol. 7, pp. 452–466, 2019.

[18] Guokun Lai et al. RACE: Large-scale ReAding comprehension dataset from examinations. In **EMNLP**, 2017.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL**, 2019.

[20] Yinhan Liu et al. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[21] Ashish Vaswani et al. Attention is all you need. In **NIPS**, 2017.

[22] Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will DNNs choose? a study from the parameter-space perspective. In **ICLR**, 2022.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. **Neural Computation**, Vol. 9, No. 1, pp. 1–42, 1997.

[24] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In **NeurIPS**, 2018.

[25] Pierre Foret et al. Sharpness-aware minimization for efficiently improving generalization. In **International Conference on Learning Representations**, 2021.

[26] Jorma Rissanen. Modeling by shortest data description. **Automatica**, Vol. 14, No. 5, pp. 465–471, 1978.

[27] Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In **EMNLP**, 2020.

[28] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. Rissanen data analysis: Examining dataset characteristics via description length. In **ICML**, 2021.

[29] Jorma Rissanen. Universal coding, information, prediction, and estimation. **IEEE Transactions on Information theory**, Vol. 30, No. 4, pp. 629–636, 1984.

[30] Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. Predicting inductive biases of pre-trained models. In **ICLR**, 2021.

[31] Christopher Clark et al. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In **EMNLP**, 2019.

[32] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Occamnets: Mitigating dataset bias by favoring simpler hypotheses. In **ECCV**, 2022.

[33] Kazutoshi Shinoda et al. Improving the robustness to variations of objects and instructions with a neuro-symbolic approach for interactive instruction following. In **MMM**, 2023.

[34] Tom Brown et al. Language models are few-shot learners. In **NeurIPS**, 2020.

[35] Zihao Zhao et al. Calibrate before use: Improving few-shot performance of language models. In **ICML**, 2021.

[36] Ruixiang Tang et al. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In **Findings of ACL**, 2023.

[37] Mosh Levy et al. Guiding LLM to fool itself: Automatically manipulating machine reading comprehension shortcut triggers. In **Findings of EMNLP**, 2023.

[38] Josh Magnus Ludan et al. Explanation-based finetuning makes models more robust to spurious cues. In **ACL**, 2023.

A　ラベルと相関の高い単語の特定

Gardner ら [15] は，言語理解タスクにおいて単一の特徴がラベルの情報を含むべきではないという理由から，入力のある特徴が与えられた元でのラベルの確率は一様分布であるべきだと主張した．そこで，ラベルの条件付き確率が統計的に有意に一様分布から乖離していることを判断するために z-statistic が提案された [15]．z-statistic z^* は以下のように計算される．

$$z^* = \frac{p(y|w)}{\sqrt{p_0(1-p_0)/n}}, \tag{1}$$

ここで p_0 はラベル y の一様分布，n は単語 w の頻度，$p(y|w)$ は単語 w が含まれる選択肢のラベル y の確率とする．RACE と ReClor データセットでは各質問に対して回答の選択肢は 4 つあるため，p_0 は 1/4 である．z-statistic が高く，ラベルとの相関が十分高いと判断される単語の例は表 2 の通り．

RACE		ReClor	
w	z^*	w	z^*
and	23.6	a	6.7
above	20.7	result	5.3
may	20.7	an	5.1
b	16.5	the	4.9
c	13.5	motive	4.5
might	10.5	not	4.3
objective	10.0	stays	4.2

表 2　RACE と ReClor データセットの訓練セットで計算された z-statistic が高い上位 7 単語．

B　最小記述長の推定

Online Code [29] による MDL の推定では，訓練セットのサブセットがモデルに複数ステップにわたって与えられる．各ステップにおいて，モデルはその時点までに与えられたサブセットについてスクラッチから学習され，次のサブセットを予測するために使用される．時間ステップ $t_0, t_1, ..., t_S$[10] によってデータセットを S 個のサブセットに分割したとき，Online Code によって MDL は以下のように推定される．

$$L = \sum_{i=0}^{S-1} \sum_{n=t_i+1}^{t_{i+1}} -\log_2 p_{\theta_i}(y_n|x_n). \tag{2}$$

ここで θ_i は $(x_j, y_j)_{j=1}^{t_i}$ で学習した質問応答モデルのパラメータであり，p_{θ_0} は一様分布である．Online Code の詳細については [27, 28] を参照のこと．ショートカットに関係なくデータセットのサイズは，SQuAD 1.1，NaturalQuestions，RACE，ReClor でそれぞれ 1400，4000，3000，300 に統一して MDL の推定に用いた．

10）　時間ステップは Voita ら [27] に従ってデータセットの 0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.25, 12.5, 25, 50, 100%とした．