

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, 菅原朔, 相澤彰子

@shino__c

kazutoshi.shinoda0516@gmail.com



Codes



arXiv

背景：質問応答モデルのショートカット学習

- ✓ 微調整された事前学習済み言語モデルに基づく質問応答モデルは、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しやすい。
- ✓ データ拡張、損失関数、モデル機構の工夫などの手法が提案されてきたが、ショートカットの種類に応じた特性を考慮していない。

貢献

何をしたか？

- ✓ ショートカットの学習可能性（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。

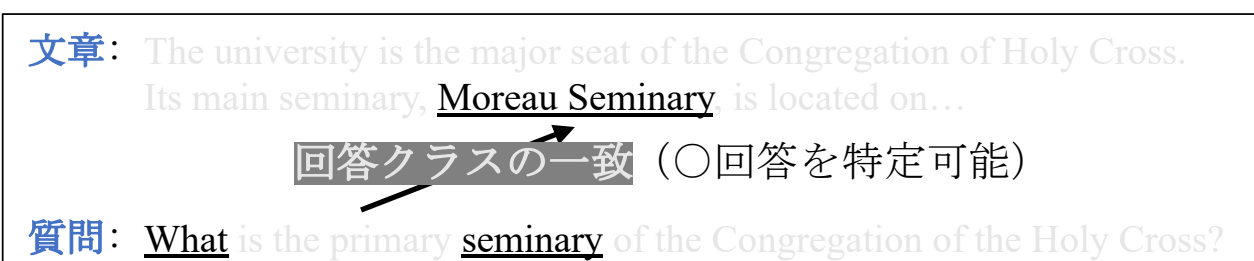
何がわかったか？

- ✓ 抽出型・多肢選択型質問応答において、ショートカットの学習可能性とショートカットの学習を回避するために必要なデータ比率の間には相関があることを示す。

ショートカット

抽出型質問応答では以下の3つのショートカットを分析する。

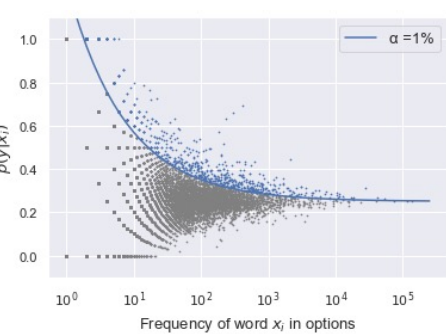
- **Answer-Position** [1]: 最初の文から回答を抽出するショートカット
- **Word Matching** [2]: 質問と最も語彙的に類似した文から回答を抽出するショートカット
- **Type Matching** [3]: 回答の固有表現タイプを予測して対応するスパンを抽出するショートカット



多肢選択型質問応答では、NLIに倣って以下の2つを分析する。

- **Word-label Correlation (Top-1)**: ある単語が選択肢に含まれるときに正解と予測するショートカット (z-statistics [4] を用いて単語を特定)

RACE		ReClor	
w	z*	w	z*
and	23.6	a	6.7
above	20.7	result	5.3
may	20.7	an	5.1



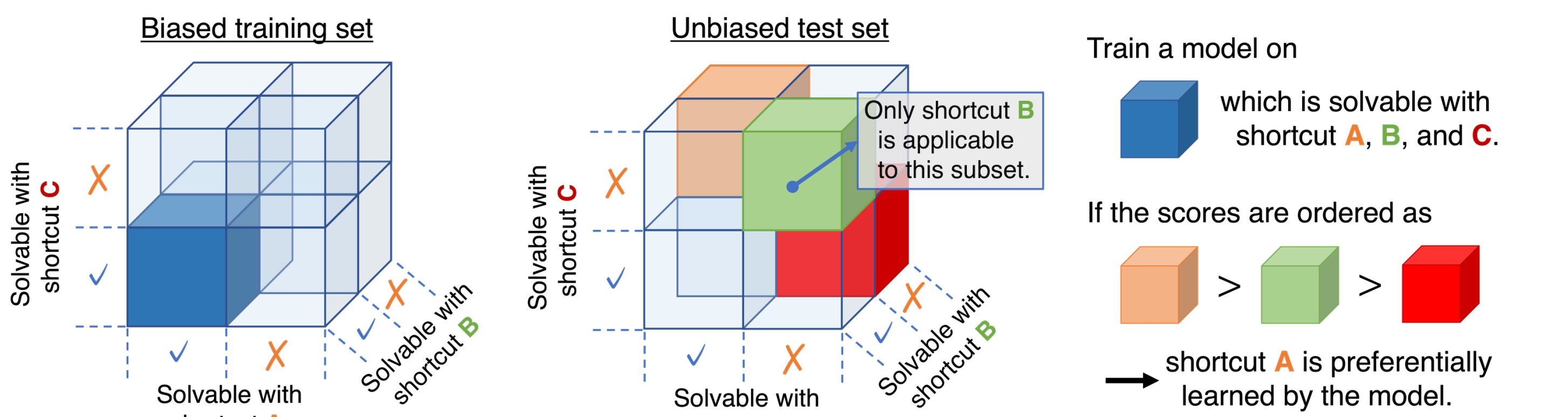
- **Lexical Overlap**: 入力文章と最も語彙の重複が多い選択肢を正解と予測するショートカット

各ショートカットについて、データセットをショートカットが使えるショートカット例 \mathcal{D}_k と使えない反ショートカット例 $\overline{\mathcal{D}_k}$ に分割する関数を定義。

実験

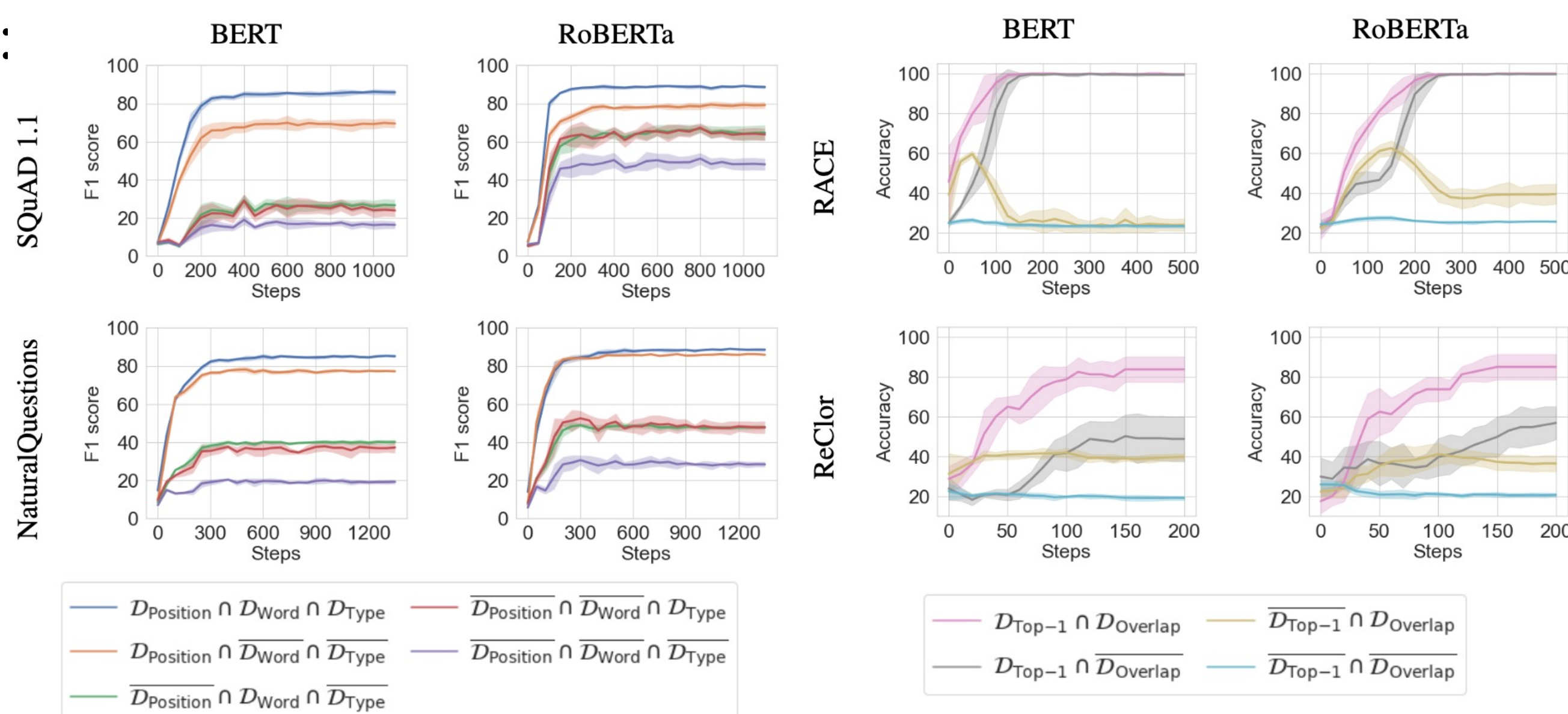
仮説検証のために、まず抽出型と多肢選択型読解における代表的なショートカットの学習可能性を行動テスト (RQ1), 質的分析 (RQ2), 量的分析 (RQ3) によって比較する。そして各ショートカットの学習の緩和に必要な反ショートカット例の割合を調べて (RQ4), 学習可能性との相関を見る。

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？



(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の図解)

結果：





➡ 訓練終了時の精度から、以下の順に優先されることが分かる

抽出型QA : **Position** > **Word** ≒ **Type**

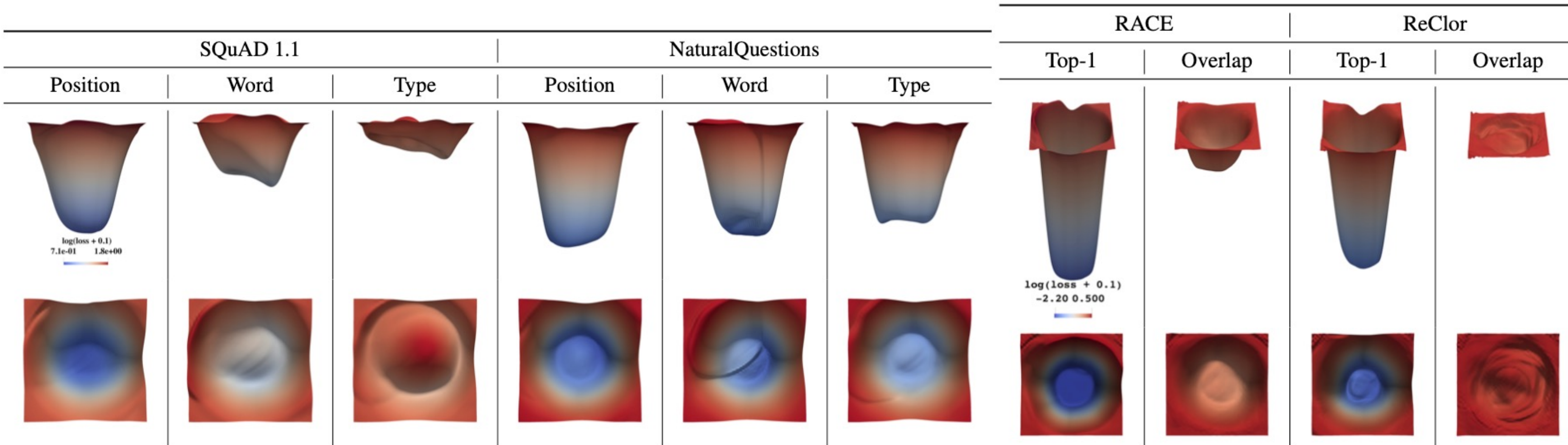
多肢選択型QA : **Top-1** > **Overlap**

多肢選択型QAで訓練の初期では **Overlap** > **Top-1** であり、Transformer の self-attention が語彙の重複を捉え易いからと解釈できる。

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？

ある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のみが有効なサブセット  で訓練することで得られると仮説を立て実験的に支持される。そして、RQ1で訓練セットに使われた  での損失を計算し、可視化して比較する。

結果：



➡ 優先して学習されたショートカット (**Position** and **Top-1**) はより平坦で深い曲面に位置する傾向がある。これが優先された理由の可能性。

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？

ショートカットの学習可能性を定量的に比較するために、**Rissanen Shortcut Analysis (RSA)** を提案する。RSAでは、あるショートカットのみが有効なサブセットの最小記述長 (≒ 学習しやすさ) を online code [5] によって近似する。

結果：

Shortcut	BERT	RoBERTa
<i>SQuAD 1.1</i>		
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
<i>NaturalQuestions</i>		
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
<i>RACE</i>		
Top-1	0.52 ± 0.34	0.41 ± 0.29
Overlap	4.16 ± 0.55	3.55 ± 0.10
<i>ReClor</i>		
Top-1	0.33 ± 0.07	0.28 ± 0.03
Overlap	0.55 ± 0.03	0.52 ± 0.02

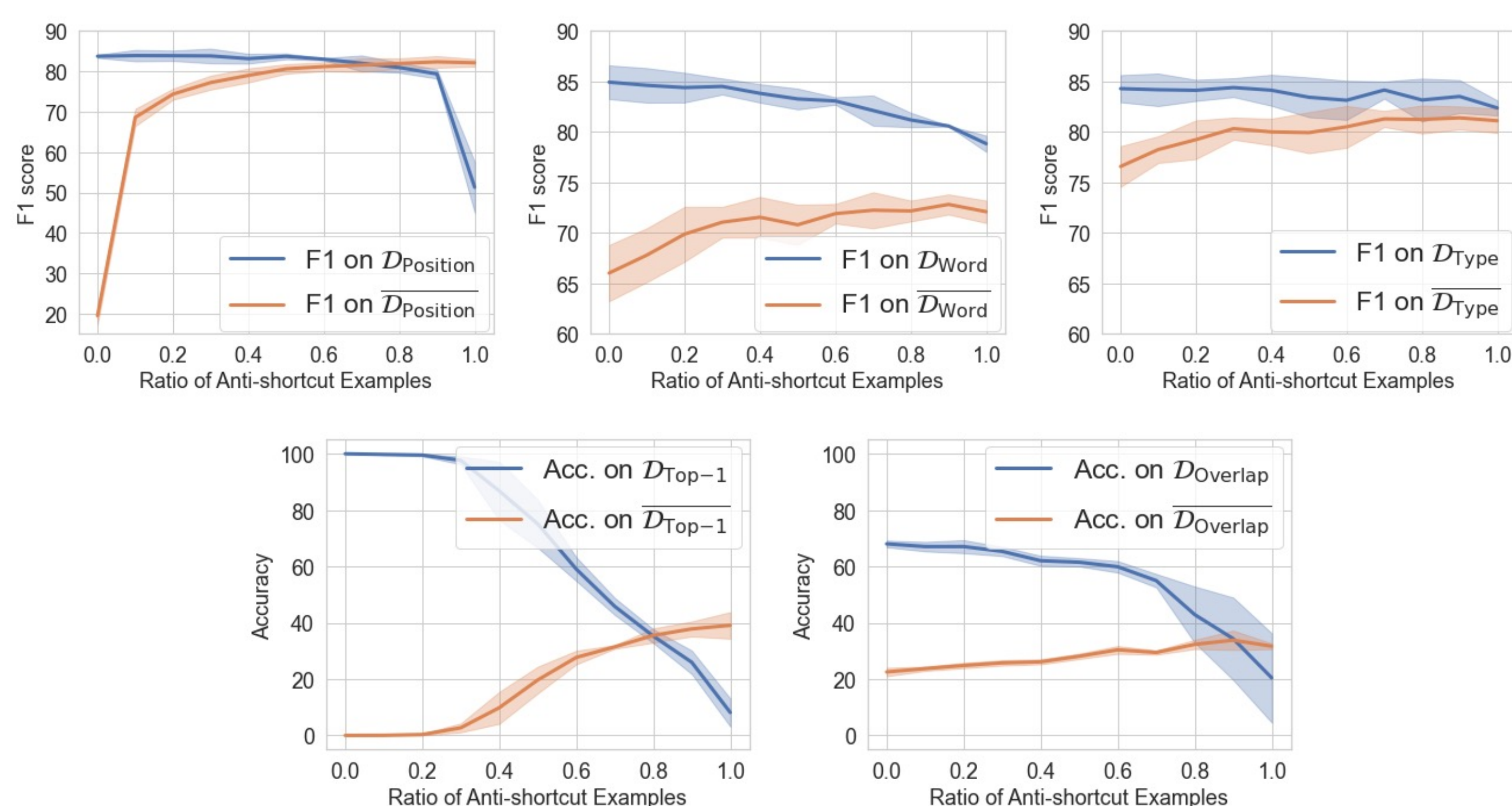
➡ RQ1で優先されたショートカット (**Position** and **Top-1**) はタスクを学習しやすく傾向にある。

➡ MDL の大小関係は RQ1,2 の結果と概ね一致している。

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はどのくらい必要か？それはショートカットの学習可能性と関係があるか？

訓練セットのサイズは固定した上で、反ショートカット例の比率を0から1まで変えた時のショートカット例 \mathcal{D}_k と反ショートカット例 $\overline{\mathcal{D}_k}$ での精度を報告する。

結果：



➡ 反ショートカット例の比率が 0.7, 0.8, 0.9 の時に **Position**, **Top-1**, **Overlap** で \mathcal{D}_k と $\overline{\mathcal{D}_k}$ での精度の差がなくなる。訓練セットの \mathcal{D}_k と $\overline{\mathcal{D}_k}$ の比率を変えるだけでは、**Word** と **Type** での精度差はなくすることができなかった。

➡ 精度差をなくするために必要なデータの比率の要件は、RQ1/2/3で明らかにしたショートカットの学習可能性と相関している。

結論

- ✓ ショートカットの学習可能性は、ショートカットの学習を緩和するための手法の設計に有用であると主張する。
- ✓ 学習しにくいショートカットについては、データ比率の調整だけでなく損失関数やモデル機構を工夫する必要がある可能性がある。

References

- [1] Ko et al. 2020. Look at the First Sentence: Position Bias in Question Answering. In EMNLP.
- [2] Sugawara et al. 2018. What Makes Reading Comprehension Questions Easier? In EMNLP.
- [3] Weissenborn et al. 2017. Making Neural QA as Simple as Possible but not Simpler. In CoNLL.
- [4] Gardner et al. 2021. Competency Problems: On Finding and Removing Artifacts in Language Data. In EMNLP.
- [5] Perez et al. 2021. Rissanen Data Analysis: Examining Dataset Characteristics via Description Length. In ICML.