

抽出型質問応答における 相対位置バイアスの除去

篠田一聡^{1,2} 菅原朔² 相澤彰子^{1,2}

¹東京大学大学院情報理工学系研究科

²国立情報学研究所



東京大学
THE UNIVERSITY OF TOKYO

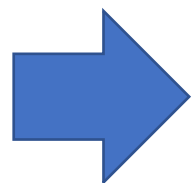


大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Background

抽出型質問応答におけるバイアス

- **語彙の重複**に関するバイアス
 - 質問と最も語彙が重複している文に答えがない時に精度が下がる (Sugawara et al., 2018)
 - 質問生成モデルは文章と語彙の重複の多い質問ばかり生成して、語彙の重複が少ない質問での精度を悪化させる (Shinoda et al., 2021)
- **回答の位置**に関するバイアス
 - 回答が1番目の文にしか現れないデータで訓練すると、2番目以降の文に回答が含まれるデータで精度が悪化する (Ko et al., 2020)



バイアスに頼った解き方を学習することで、
抽出型質問応答モデルの汎化性能の低下につながる

Pilot Study

相対的な位置に関するバイアス

定義: 相対的な位置 (Relative Position) d

- **回答スパン**に最も近い文章質問間の重複語彙の**回答スパン**から見た相対的な位置

例:

- 文章: ... The American Football Conference (**AFC**) champion **Denver Broncos** defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third **Super Bowl** title. ...
- 質問: Which NFL team represented the **AFC** at **Super Bowl** 50?
- Relative Position d : **−3** (単語レベル)

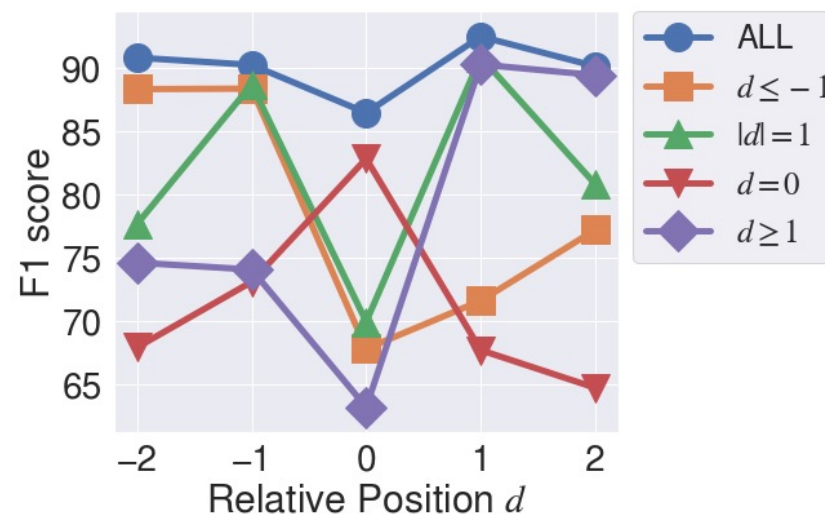
−3

Pilot Study

相対的な位置に関するバイアス

実験:

- モデル:
 - BERT-base
- 訓練セット:
 - SQuAD訓練セット
 - ALL
 - Relative Position d に関して以下の条件を満たすサブセット
 - $d \leq -1$
 - $|d| = 1$
 - $d = 0$
 - $d \geq 1$



結果:

- 訓練時に見たことのある d では精度を保てるが、見たことのない d では精度が低下



相対的な位置に関するバイアスが
モデルの汎化性能に悪影響を及ぼす

目的

相対位置バイアスの除去

目的:

- 訓練時に見たことのない相対位置で汎化性能が低下する問題を解決すること

意義:

- 実応用において開発者が独自に作成した訓練セットの分布が何らかの観点で意図せず偏ってしまうことで、汎化性能が低下することは十分に考えらる。
- そのため、モデルが利用しうるバイアスについては、事前に対処方法を研究する必要がある。

手法

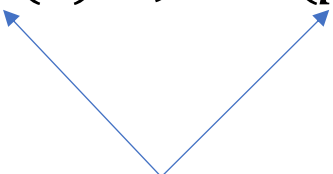
抽出型質問応答の損失関数

入力: 文章、質問

出力: 回答スパンの始点と終点

損失関数:

- 回答スパンの始点 s と終点 e の交差エントロピー誤差の和

$$L = L(\hat{p}(s), s) + L(\hat{p}(e), e)$$


相対位置に関して頑健な \hat{p} の訓練方法を本研究で提案

手法

相対位置のためのバイアス除去手法

Step 1:

- Biased model の準備
 - 相対位置を利用した解き方をするモデルを用意する

Step 2:

- Main model の訓練
 - Biased model の予測結果を利用して、相対位置を利用しない解き方の学習を促進

手法

相対位置のためのバイアス除去手法

Step 1:

- Biased model の準備
 - 相対位置を利用した解き方をするモデルを用意する

Step 2:

- Main model の訓練
 - Biased model の予測結果を利用して、相対位置を利用しない解き方の学習を促進

手法

Main model の準備

- BiasProduct (Clark et al., 2019)

$$\hat{p} = \text{softmax}(\log p + \log b)$$

- LearnedMixin (Clark et al., 2019)

$$\hat{p} = \text{softmax}(\log p + g(c, q) \log b)$$

p : main modelの出力確率

b : biased modelの出力確率

g : 学習可能な関数 (≥ 0)

訓練時: b は固定して p **のみ**を更新
テスト時: p **のみ**を使って推論

Biased model にバイアスを利用した解き方を意図的に学習させることで、

- Biased modelが正しい予測をした時=データがバイアスを含む時
 - Main model の学習が進まない
- Biased modelが誤った予測をした時=データがバイアスをあまり含まない時
 - Main model の学習が進む



Biased model に相対位置に関するバイアスを学習させることが必要

手法

相対位置のためのバイアス除去手法

Step 1:

- Biased model の準備
 - 相対位置を利用した解き方をするモデルを用意する

Step 2:

- Main model の訓練
 - Biased model の予測結果を利用して、相対位置を利用しない解き方の学習を促進

手法

Biased model の準備

- Answer Prior (AnsPrior)

- 重複語彙の周辺に出力確率を割り当てるヒューリスティック

👤 コストがかからない

👤 訓練セットの偏り方に依存しており柔軟性に欠ける

例) 重複語彙の右側が回答になりやすい時 $d \leq -1$

■ : 質問と文章で重複している語彙

出力確率 b :



入力文章:

... w_3 w_4 w_5 w_6 w_7 w_8 ...

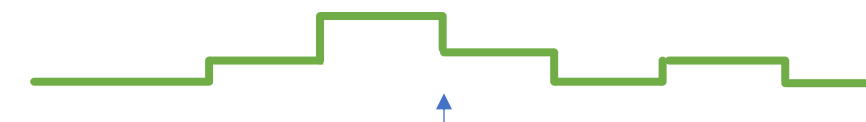
- Position-only model (PosOnly)

- 入力を各単語が重複語彙が否かの情報のみを与えてモデルを訓練

👤 訓練セットがどのように偏っていてもそのまま適用可能

👤 訓練にコストがかかる

出力確率 b :



Position-only model

入力文章:

... 0 1 0 0 1 0 ...

実験・結果

実験設定:

- モデル:
 - BERT-base
 - 提案手法
 - BiasProduct / LearnedMixin
 - AnsPrior / PosOnly
- 訓練セット:
 - SQuAD訓練セット (ALL)
 - Relative Position d に関して以下の条件を満たすサブセット
 - $d \leq -1$
 - $|d| = 1$
 - $d = 0$
 - $d \geq 1$

実験・結果

結果:

- BiasProduct よりも LearnedMixin の方が精度改善に効果的
- ➡ データごとにどれくらいbiased model の予測を main modelの訓練に利用するかを学習することが重要
- LearnedMixin + AnsPriorに比べてPosOnlyは、見たことのないdでの精度を最も向上できるが、見たことのあるdでの精度を下げてしまう。
- ➡ 訓練時と同じ分布、違う分布での精度の間にトレードオフが見られる。

Trained on	Model	Evaluated on						
		$d \leq -3$	$d = -2$	$d = -1$	$d = 0$	$d = 1$	$d = 2$	$d \geq 3$
ALL	BERT-base	82.19	90.82	90.25	86.47	92.49	90.14	81.43
$d \leq -1$	BERT-base	78.17	88.34	88.38	67.82	71.62	77.22	69.54
$d \leq -1$	BiasProduct-AnsPrior	73.00	84.34	85.61	46.32	25.23	64.91	59.06
$d \leq -1$	LearnedMixin-AnsPrior	79.07	89.27	89.01	68.52	72.35	80.43	70.31
$d \leq -1$	BiasProduct-PosOnly	75.04	83.90	83.22	73.80	81.35	81.79	73.27
$d \leq -1$	LearnedMixin-PosOnly	77.00	86.72	86.25	74.26	82.66	82.81	75.94
$ d = 1$	BERT-base	65.62	77.69	88.70	69.96	90.88	80.84	66.42
$ d = 1$	BiasProduct-AnsPrior	60.44	75.07	56.44	49.32	52.37	72.85	57.98
$ d = 1$	LearnedMixin-AnsPrior	73.42	83.39	88.70	74.24	90.47	85.51	73.52
$ d = 1$	BiasProduct-PosOnly	72.41	80.59	84.01	73.34	87.61	83.11	72.09
$ d = 1$	LearnedMixin-PosOnly	73.76	80.63	86.10	74.50	89.64	82.98	72.04
$d = 0$	BERT-base	60.75	67.94	73.11	82.85	67.72	64.74	52.88
$d = 0$	BiasProduct-AnsPrior	56.25	65.15	69.05	81.07	65.10	62.95	49.43
$d = 0$	LearnedMixin-AnsPrior	59.66	69.62	72.53	83.06	68.04	66.03	53.29
$d = 0$	BiasProduct-PosOnly	62.97	67.88	70.22	78.66	66.69	69.12	59.88
$d = 0$	LearnedMixin-PosOnly	65.09	70.47	72.51	81.32	68.29	68.47	59.54
$d \geq 1$	BERT-base	68.03	74.63	74.08	63.21	90.28	89.44	75.42
$d \geq 1$	BiasProduct-AnsPrior	58.63	63.13	29.08	39.22	88.53	88.34	72.29
$d \geq 1$	LearnedMixin-AnsPrior	70.71	77.22	76.82	66.67	90.87	89.75	76.31
$d \geq 1$	BiasProduct-PosOnly	68.54	78.13	78.58	70.72	85.17	81.59	72.90
$d \geq 1$	LearnedMixin-PosOnly	71.17	80.41	79.97	71.33	87.53	84.33	74.24

見たことがあるd / 見たことがないd

結論

貢献

- 重複語彙の回答から見た相対的な位置が偏っていると、モデルの汎化性能に悪影響を与えることを示した。
- バイアスを利用するモデルとのアンサンブルによってこの問題を解決する手法を提案し、有効性を示した。

今後の展望

- 訓練時と同じ分布と異なる分布での精度の間のトレードオフを解消する
- モデルの中間表現や学習過程の分析によって相対位置バイアスを学習するメカニズムを理解する