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　[2] 国立情報学研究所
shinoda@is.s.u-tokyo.ac.jp　{saku,aizawa}@nii.ac.jp

概要

抽出型質問応答において，質問–文章間で重複している語彙の回答スパンから見た相対位置の分布が訓練セットにおいて偏っている時，質問応答モデルはそれを利用した解き方を学習してしまう問題を発見した．本研究では訓練セットで相対位置がどのような偏り方をしていても適用可能なバイアス除去手法を提案する．訓練時に見たことのない相対位置のデータへの汎化性能の評価によって提案手法の有効性を示した．

1　はじめに

深層学習に基づく自然言語理解モデルが訓練セット中のバイアスを利用した解き方を学習してしまい，汎化性能が悪化してしまうことが自然言語理解分野における重大な課題となっている [1, 2]．特に意図的に分布が偏った訓練セットで訓練された自然言語理解モデルは，入力と出力の間の因果関係ではなく，バイアスを利用した解き方をより学習しやすい傾向があることが知られている [3, 4, 5]．実応用においても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偏ってしまう恐れは十分にある．ゆえに，分布が偏った訓練セットからバイアスを利用しない解き方を学習するための方法を開発することが重要である．

文章から回答スパンを抽出して質問に答える抽出型質問応答 [6] においては，質問–文章間で重複している語彙（重複語彙）の回答スパンから見た相対位置が偏っている時に，質問応答モデルがその相対位置に関するバイアスを学習する傾向があることを新たに発見した．例えば，訓練セット中で重複語彙と回答スパンが隣接している（相対位置の絶対値が 1 の）データのみで構成されている時，読解モデルはテストセットで同様のデータでは精度を高く保てるが，そうでないデータでは精度を大幅に落としてしまう現象が確認された（図 1）．これはモデルが重複語彙の隣接箇所から回答を探す解き方を優先して学習しているとも捉えられる．訓練セットの相対位置の分布が異なる時も同様の現象が観測された．

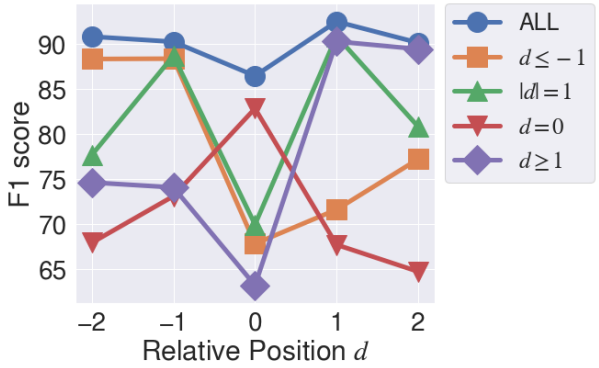


図 1　SQuAD 開発セットにおける回答に最も近い重複語彙の回答から見た相対位置 d ごとの F1 スコア．凡例の ALL は SQuAD 訓練セットの全てのデータで訓練されたモデルを指し，その他はそれぞれの条件式が満たされるデータでのみ訓練されたモデルを指す．モデルにはいずれも BERT-base を用いた．訓練時に見たことがある d のデータでは ALL と比べて精度を保てるが，その他では精度が悪化する．d の定義は式 3 を参照されたい．

本研究では，抽出型質問応答の訓練セットが相対位置に関して偏っていても，モデルが相対位置のみを利用した解き方の学習をせずに，相対位置の分布が異なるテストセットでも回答スパンも正しく予測するための学習方法の開発を目的とする．そのために，相対位置バイアスを利用して予測をするバイアスモデルと，実際に予測をするメインモデルの product-of-experts [7] を用いた手法を提案する．これによってメインモデルが相対位置バイアスだけを利用した解き方を学習しないよう促進する．実験において，提案手法は相対位置に関してバイアスがかかっていない時に近い精度を達成することを示す．さらに，提案手法は訓練セットが相対位置に関して異なる偏り方をしていても汎用的に効果的であることを示す．

2　相対位置バイアス

2.1 定式化

抽出型質問応答タスクにおいて回答スパンから最も近い重複語彙の回答から見た相対位置を d とおく．正確には w を単語，文章を $c = \{w_i^c\}_{i=0}^N$，質問を $q = \{w_i^q\}_{i=0}^M$，回答を $a = \{w_i^c\}_{i=s}^e$（$0 \le s \le e \le N$）としたとき，相対位置 d は以下のように定義する．

$$f(j,s,e) = \begin{cases} j-s, & \text{for } j < s \\ 0, & \text{for } s \le j \le e \\ j-e, & \text{for } j > e \end{cases} \tag{1}$$

$$D = \{f(j,s,e) | w_j^c \in q\} \tag{2}$$

$$d = \text{argmin}_{d' \in D} |d'| \tag{3}$$

ここで $0 \le j \le N$ は文章中の単語 w_j^c の位置，$f(i,s,e)$ は w_i^c の a から見た相対位置，D は q と c で重複している全ての単語の相対位置の集合を表す．[1] 式 3 で最も絶対値が小さいものを相対位置として定義して用いるのは，質問応答モデルが重複語彙から近い位置のスパンをより優先して予測する傾向があること [8] や，回答スパンと重複語彙の絶対的な距離が大きい時に精度が悪化すること [9] が知られているからである．[2]

2.2 相対位置の分布

図 2 に SQuAD [6] の訓練セットにおける相対位置 d の度数分布を示す． d の値は 0 の周辺に偏っていることがわかる．SQuAD 以外の質問応答データセットにおいても 0 の周辺に偏る傾向は一貫している一方で，データセット間で分布の形に違いが見られる．（詳細は付録 A を参照されたい．）この違いはデータセットの集め方や文章のドメインによって生じたものと考えられる．よって，特定の相対位置の分布に特化した学習は避け，相対位置のバイアスを学習しない質問応答モデルの構築が求められる．

3 手法

3.1 バイアス除去手法

バイアス除去のためのアルゴリズムには，product-of-experts [7] を応用した手法である BiasProduct と

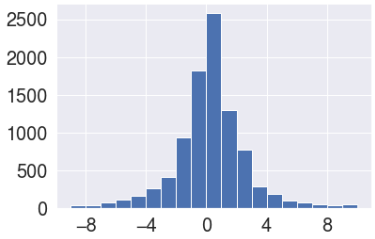


図 2 SQuAD 訓練セットの相対位置 d の度数分布

LearnedMixin [10, 11] を用いる．これらの手法では，共にバイアスを利用した予測をする Biased model をまず用意し，それと Main model を混合した product-of-experts を用いて損失関数を求める．Main model の訓練時は Biased model は固定した上で誤差関数を最小化する．そしてテスト時には Main model のみを用いて予測を行う．既存研究 [12, 13] に倣い，モデル $\hat{p}$ は回答スパンの始点 s と終点 e の確率 $\hat{p}(s)$ と $\hat{p}(e)$ を出力し，誤差関数は始点と終点の交差エントロピー誤差の和とする．以降簡単のため $\hat{p}(s)$ と $\hat{p}(e)$ の区別はせず $\hat{p}$ と表記する．

3.1.1 BiasProduct

BiasProduct では以下のように Biased model の出力確率 b と Main model の出力確率 p の対数の和を softmax 関数に入力して $\hat{p}$ を得る．

$$\hat{p} = \text{softmax}(\log p + \log b) \tag{4}$$

これによって，Biased model が正しく予測をできるようなバイアスを含むデータよりも， Biased model が予測を誤るようなバイアスを含まないデータで Main model の学習が強く促進される．

3.1.2 LearnedMixin

BiasProduct は Biased model の出力確率に強く依存した学習手法である．Main model が Biased model の予測を信頼できるかどうかを各データごとに予測する LearnedMixin を用いることで Main model がより頑健なモデルになることが知られている．

$$\hat{p} = \text{softmax}(\log p + g(c,q) \log b) \tag{5}$$

ここで $g \ge 0$ は学習可能な関数である．

3.2 Biased model

3.1 節で述べた Biased model の構築方法を述べる．

1) 内容語だけでなく機能語も読解をする上で重要な手がかりとなりうるため，式 2 ではあらゆる単語を対象として重複判定を行う．
2) 式 3 で d が一つに定まらない場合が少数だが存在するが，このようなデータは訓練セットと評価セットから必要に応じて除くこととする．

3.2.1 Answer prior

訓練セットの相対位置 d の分布に応じて，回答スパン a の始点と終点の事前確率を経験的に定めた Answer Prior (AnsPrior) を Biased model として用いる．具体的には，図 1 の凡例に示した 4 つの条件を満たす訓練セットのサブセットそれぞれに対して，文章中の単語 w_i^c が始点または終点である事前確率 b_i を以下のように定める．

$$b_i = \begin{cases} \mathbb{1}\left[w_{i+1}^c \in q\right]/Z, & \text{for } d \leq -1 \\ \mathbb{1}\left[(w_{i+1}^c \in q) \vee (w_{i-1}^c \in q)\right]/Z, & \text{for } |d| = 1 \\ \mathbb{1}\left[w_i^c \in q\right]/Z, & \text{for } d = 0 \\ \mathbb{1}\left[w_{i-1}^c \in q\right]/Z, & \text{for } d \geq 1 \end{cases} \tag{6}$$

ここで Z は正規化定数である．これらの事前確率は訓練セットで回答になりやすい箇所に等しく確率を割り当てる発見的手法に基づいている．そのため，特定の訓練セットの相対位置 d の分布に特化した事前確率である点で柔軟性に欠ける．

3.2.2 Position-only model

訓練セットで相対位置 d がどのような偏り方をしていても適用可能な Position-only model (PosOnly) を Biased model として提案する．PosOnly は，文章中の各単語の絶対的な位置と，各単語が質問と重複しているか否かのみの情報を入力として，回答スパンの始点と終点を出力として訓練する．PosOnly が回答スパンを予測するためには重複単語からの相対的な距離と絶対的な位置のみしか利用できる情報がないため，訓練セットの相対位置がどのように偏っていたとしても相対位置バイアスを利用した解き方を学習することが期待される．

4 実験

4.1 実験設定

データセット　データセットには SQuAD 1.1 [6] を用いた．モデルが相対位置 d が偏ったデータから見たことのある相対位置に依存しない解き方を学習できているかを評価するために，SQuAD 訓練セットから 4 つの異なる方法で意図的に相対位置バイアスを含むサブセットを取得して別々に訓練に用いた．4 つのサブセットは，相対位置 d が条件 $d \leq -1$，$|d| = 1$，$d = 0$，$d \geq 1$ を満たすデータのみを抽出してそれぞれ構築した．サイズはそれぞれ 33,256，30,003，21,266，25,191 である．比較のために全ての訓練セットを用いた場合のスコアも報告する．評価には SQuAD 開発セットの相対位置ごとの性能を評価した．抽出型質問応答タスクの評価指標には，回答スパンの完全一致を測る Exact Match スコアと部分一致を測る F1 スコアが用いられてきた [6] が，紙面の節約のため F1 スコアのみ報告する．

比較手法　バイアス除去のための学習手法として BiasProduct と LearnedMixin の 2 種類，それぞれに使われる Biased model として AnsPrior と PosOnly の 2 種類の全ての組み合わせ 4 通りと，何も工夫せずに訓練した場合の計 5 通りの訓練方法を比較してこれらの有効性を評価した．Main model と PosOnly にはいずれも BERT-base [13] を用いた．

4.2 結果

表 1 に結果を示す．まず，訓練セットを全て用いた場合（ALL），BERT-base の性能は $|d| = 1, 2$ の時に 90 ポイントを超える一方で $|d| \geq 3$ の時には 8 ポイント程度低下していた．2.2 節で示したように訓練セットの相対位置 d の分布が 0 付近に偏っていたことから，**訓練セットで頻度が高い d では精度が高くなり，逆に頻度が低い d では精度が低くなるという仮説**が立てられる．

相対位置 d の分布を意図的に偏らせた訓練セットを使って BERT-base の通常の訓練を行なった結果はこの仮説の信頼性を高めるものであった．相対位置 d が $|d| = 1$ を満たすデータのみで通常の訓練を行なったところ，$|d| = 1$ では F1 スコアが ALL に比べて 2 ポイント未満しか低下しなかったのに対して，$|d| \neq 1$ では F1 スコアが 10~15 ポイント低下した．他のサブセットを用いた通常の訓練でも同様の傾向が見られた．これが示唆するのは，同じ相対位置 d の値を持つデータで成り立つ擬似相関をモデルが利用して推論を行なっているということである．

訓練セットのサブセットが満たす d の条件が同じ下で，提案する 4 種類のバイアス除去手法と通常の訓練をそれぞれ比較して得られる結論を述べる．まず学習手法としては多くの場合 BiasProduct よりも LearnedMixin の方が高い F1 スコアが得られた．この結果は各データで Biased model の予測結果を Main model の訓練に反映させる度合いを学習することの有効性を示している．また，Biased model としては，訓練時にモデルが見たことのない相対

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

表 1 SQuAD 開発セットの各サブセットにおける F1 スコア．灰色 で示した箇所では訓練時と評価時の d が重複している．灰色 では全ての訓練セットを用いた場合 (ALL) に比べてスコアが高く保たれる傾向にある．

位置 d のデータへの汎化性能—すなわち表 1 の白い箇所のスコア—を向上させるためには，AnsPrior よりも PosOnly の方が優れていることがわかった．特に訓練時に $d = 0$ でテスト時に $d \leq -3$ の場合や，訓練時に $d \leq -1$ でテスト時に $d \geq 3$ の場合は，LearnedMixin-PosOnly は LearnedMixin-AnsPrior よりも 5 ポイント程度優っていた．一方で，訓練時に見たことのある相対位置 d のデータへの汎化性能—すなわち表 1 の 灰色 の箇所のスコア—の観点では，LearnedMixin-AnsPrior が最も優れていた．LearnedMixin-AnsPrior 以外の提案モデルは通常の訓練と比べてスコアを悪化させてしまうことが多いこともわかった．バイアス除去の既存研究 [14] でも指摘されているが，相対位置のバイアス除去でも訓練時と同じ分布と違う分布のテストセットでの精度にトレードオフが生じるという問題が明らかになった．

5 関連研究

訓練セットに特有なバイアスを利用した解き方を自然言語理解モデルが学習していること，または訓練セットと同じ分布のテストセットで高い精度を出すにはバイアスを利用するだけで十分であることを指摘する研究が近年増えている．自然言語推論では特定の語彙 [15] や，語彙の重複 [1] とラベルとの擬似相関をモデルが利用してしまう問題が指摘された．抽出型質問応答では質問と回答のタイプ一致だけで解ける質問が多いこと [16, 9] や，回答の絶対的な位置のバイアスをモデルが学習しやすいこと [4] が知られている．モデルが人らしい読解スキルを学習できているかどうかを評価するには，訓練時とテスト時で違う分布のデータを使うことが最も有効な方法の一つである．本研究では相対位置という新しい観点でこれについて分析して問題を発見し，有効なバイアス除去手法を開発した．誤差関数は既存研究 [4] と同じものを用い，相対位置バイアスのための Biased model は独自に設計した．

6 おわりに

抽出型質問応答でモデルが相対位置に関するバイアスを利用しうることを示し，それを解決するための有効なバイアス除去手法を提案した．他のデータセットでの検証，モデルの中間表現や学習過程についての分析が今後の展望である．

謝辞

本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2108 の支援を受けたものです．本研究は JSPS 科研費 JP21H03502, JP20K23335 の助成を受けたものです．

参考文献

[1] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. **Nature Machine Intelligence**, Vol. 2, No. 11, pp. 665–673, November 2020.

[3] Mike Lewis and Angela Fan. Generative question answering: Learning to answer the whole question. In **International Conference on Learning Representations**, 2019.

[4] Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. Look at the first sentence: Position bias in question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1109–1121, Online, November 2020. Association for Computational Linguistics.

[5] Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. Predicting inductive biases of pre-trained models. In **International Conference on Learning Representations**, 2021.

[6] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[7] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. **Neural computation**, Vol. 14, No. 8, pp. 1771–1800, 2002.

[8] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[9] Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. What makes reading comprehension questions easier? In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 4208–4219, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 4069–4082, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11] He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. In **Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)**, pp. 132–142, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In **International Conference on Learning Representations**, 2017.

[13] 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.

[14] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8717–8729, Online, July 2020. Association for Computational Linguistics.

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

[16] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Making neural QA as simple as possible but not simpler. In **Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)**, pp. 271–280, Vancouver, Canada, August 2017. Association for Computational Linguistics.

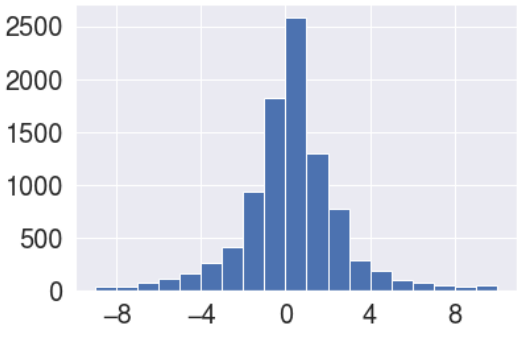
[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

A　相対位置の分布

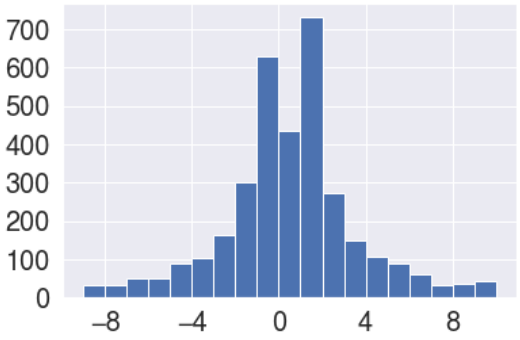
SQuAD, NewsQA, TriviaQA, NaturalQuestions の開発セットにおける相対位置の度数分布を図 3 に示す．データセットによって違いはあるものの，いずれも相対位置が 0 の周辺に偏っていることがわかる．

B　訓練の詳細

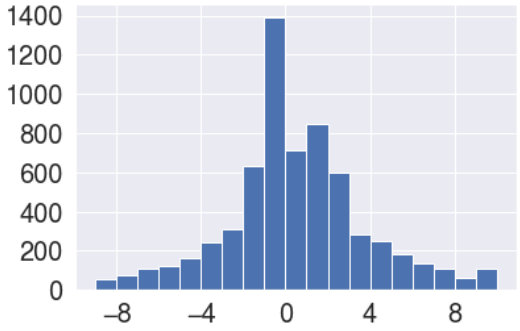
各モデルの訓練のエポック数は 2，バッチサイズは 32 とし，学習率は 3e-5 から線形に 0 まで減少させ，最適化には Adam [17] を用いた．



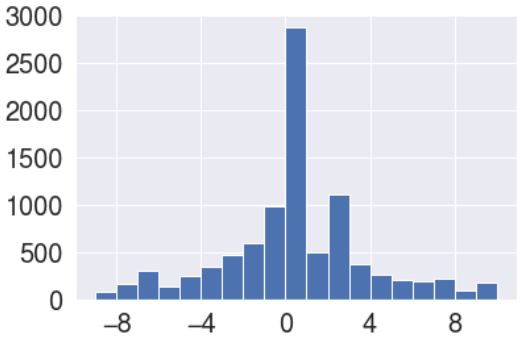
(a) SQuAD



(b) NewsQA



(c) TriviaQA



(d) NaturalQuestions

図 3　回答から見た重複語彙の相対位置の度数分布