コンテキストの量が質問応答モデルのショートカット推論に与える影響について

秋元康佑　竹岡邦紘　小山田昌史
NEC　データサイエンス研究所
{kosuke_a,k_takeoka,oyamada}@nec.com

概要

質問応答は近年活発に研究が進められている言語理解タスクであるが, 近年の深層学習を利用した質問応答モデルは訓練データセット中のバイアスを利用したショートカット推論に陥りやすいことが知られている. 一方この問題に関する既存研究の分析はショートカット推論を促す情報源文章の特徴に関するものがほとんどであり, 情報源文章の数（長さ）が変化した場合のショートカット推論の挙動については詳しくわかっていない. 本研究ではオープンドメイン質問応答の設定に注目し, 特に検索によって得られた情報源文章を利用する retrieve-then-read 方式の質問応答モデルが持つショートカット推論の挙動が情報源文章の数によってどのように変化するかを調べた. 実験の結果, そのような質問応答モデルは回答に必要ない不要な情報源文章の数が増えるほどショートカット推論を引き起こしやすくなり, この問題が主にモデルに悪影響を及ぼす一部の文章によって引き起こされていることがわかった.

1　はじめに

質問応答 (question answering) とは, 利用可能な情報源文章 (コンテキスト) を基にユーザーから与えられた質問文に対する回答を自動的に作成するタスクである. 近年では大規模なベンチマークデータセット [1] が公開されるなど活発に研究が行われており, 大規模言語モデルの利用などによって一部のベンチマークデータセットに対する性能が人間を超えるなど性能面での進歩も著しい [2].

一方で質問応答モデルが本来意図されていた言語理解能力ではなく, 訓練に利用したデータセットに含まれるバイアスなどを利用したショートカット推論 [3] を学習してしまう事例も報告されており, ベンチマークデータセット以外への実応用に向けた課題となっている. 例えば [4] では質問応答モデルが質問と単語の重複が大きいコンテキスト文に依存しており, そうした文に正答が含まれない場合に誤答が増える結果が報告されている.

本研究では特にコンテキストが明示的にシステムに与えられないオープンドメイン質問応答の設定に注目し, そのような設定で近年よく用いられる retrieve-then-read 方式 [5] と呼ばれるシステムが持つショートカット推論への脆弱性について研究する. retrieve-then-read 方式のシステムは質問文をクエリとしてコンテキストを外部コーパスから検索して得る. そのため質問文とコンテキストとの単語的・意味的な類似度が大きくなりやすい特徴があり, 前述した従来研究において指摘されているショートカット推論に対する脆弱性が大きな課題となりうる. さらにこうしたシステムでは検索の recall を確保するため比較的多くの文章 (100 個程度まで) を利用することが一般的 [6] であるため, モデルにショートカット推論を促す文が複数含まれる可能性がある.

しかしショートカット推論に関する従来研究はコンテキストに含まれる文章の内容（意味や単語）を対象に分析を行っており, コンテキストに含まれる文章の量に関連するショートカット推論の挙動は著者らの知る限り [7] における限定的な実験があるのみである.

さらにコンテキストの量は推論時だけでなく, retrieve-then-read 方式の質問応答モデルを訓練する際にも性能に影響を与えていることが示されている [6, 8]. しかし特に質問応答に不要なコンテキストの量が性能に与える影響が明らかになっていないなど, 訓練時のコンテキストの量に関する研究もまだ十分行われているとは言えない.

そこで本研究では retrieve-then-read 方式の質問応答システムの訓練および推論の挙動が, コンテキスト中の必要な文章 (positive passage) と必要ない文章

(negative passage) の量にどのように影響を受けるかを分析する. この研究は現在進行中であるが, 本稿では執筆時点で完了している推論時の挙動に関する実験の結果について報告する.

2 関連研究

与えられた自然言語文のコンテキストを利用して質問応答を行う機械読解の分野は近年 SQuAD データセット [1] などの大規模なベンチマークが公開されるなど活発に研究が行われており, 大規模言語モデルを利用した手法 [9] などが提案されている. またコンテキストが明に与えられないオープンドメイン質問応答の分野の研究も盛んに行われており, Wikipedia などの外部知識源から質問に関連する文章を検索してコンテキストとして利用する retrieve-then-read 方式 [5] が現在主流となっている. retrieve-then-read 方式ではコンテキストに含まれる情報の recall を増やすため比較的多くの文章をコンテキストとして利用することがあり, 効率的に複数の文章の情報を利用するために RAG[10] や FiD[6] などのモデルが提案されている.

一方で学習された質問応答モデルがデータセットに含まれるバイアスに頼ったショートカット推論 [3] を学習してしまっていることを示唆する事例も多数報告されている [11, 4, 12, 13, 14, 15, 16, 17, 7, 18]. 例えば [11, 4, 7] ではモデルが質問文と単語的な重複が大きい文に注目して誤答してしまう傾向があることを示す結果が報告されており, [16] では特にモデルが注目している質問文中の単語が重複している場合にこの傾向が強いことが示されている. また質問応答モデルはエンティティの型情報も利用しており, エンティティの型のみを考慮して正答が得られるバイアス [14] を利用したショートカット推論を学習してしまいうることが示されている [15]. これらの既存研究では主にモデルにショートカット推論を促しやすいコンテキストの内容に関する分析を行っているが, 一方でそのようなコンテキストの量に関連した分析も行っている研究は著者らの知る限り [7] のみで比較的少ない. ショートカット推論以外の分野では, コンテキストに含まれる矛盾した情報の量による質問応答モデルの挙動を分析した論文がつい最近発表されている [19, 20].

3 実験

本章ではまず retrieve-then-read 方式の質問応答モデルの推論時の性能が, コンテキスト[1)]中の回答に必要な情報を含む文章 (positive passage) と必要ない文章 (negative passage) の量に依存することを示し (§3.3), この性能低下が特に質問応答モデルに性能低下を促す一部の文章 (hard negative passage) によって引き起こされていることを示す.

3.1 データ

本稿における実験では, オープンドメイン質問応答のベンチマークデータセットである Natural Question[21] を利用した. 本実験では retrieve-then-read 方式の検索部分については分析の対象外とし, Wikipedia から DPR[22] によって検索された文章が各質問に対するコンテキストとして予め付与されている, [6] における前処理済みデータを利用した. これらのデータセットにおける各質問のコンテキストについて, アノテーションされている正答を文字列的に完全一致で含む文章を positive passage, 含んでいない文章を negative passage として取り扱った. 後述するモデルの学習には `train` サブセットを利用することとし[2)], モデルの性能評価には `dev` サブセットを利用した[3)].

3.2 質問応答モデル

本稿における実験では, 質問応答モデル (reader) として近年高い性能を示している FiD[6] と呼ばれる encoder-decoder モデルを対象とした. FiD は text-to-text タスクで利用される T5[23] を拡張したモデルであり, より効率的な推論を実現するため複数の入力文 $\{s_1, ..., s_n\}$ をそれぞれ独立に encoder f_{enc} によってエンコードする[4)]ことが特徴である. 各入力文 s_i のエンコード結果 $f_{\text{enc}}(s_i)$ は結合されて decoder f_{dec} に入力され, 自己回帰的に出力文 o が生成される[5)]. (すなわち $[x; y]$ をベクトル列 x と y の

1) 本稿では個々の情報源文章 (passage) と, それらの集合としてのコンテキストを用語として区別する.
2) 実際に訓練データとして利用されるデータと early stopping のための検証用データの 2 つにさらに分割した.
3) 全ての実験で同じ質問集合が評価に利用されるよう, このうち positive passage を 8 個以上, negative passage を 64 個以上持つような質問のみを利用した.
4) T5 と同様に, 入力文中の各トークンごとにベクトルが計算される.
5) FiD は与える入力文の順序の並べ替えによって出力がほとんど変化しないことが報告されており [8], 著者らの予備実験

結合として, $o = f_{\text{dec}}([f_{\text{enc}}(s_i); \ldots; f_{\text{enc}}(s_n)])$ である.) 質問応答モデルとして利用される際は, 一般的に質問文 q とコンテキスト中の i 番目の文章 p_i を用いて s_i = "question: q context: p_i" のようなテンプレートに従い入力文 s_i を作成し, 出力文 o として正答を出力させるように学習する[6]. また FiD を初期化するための T5 モデルとしては, transformers ライブラリ [24] の `t5-base` を利用した.

3.3 positive, negative passage の量による性能変化

本節では質問応答モデルの性能がコンテキスト中の positive passage と negative passage の個数によってどのように変化するかを調べる.

まず実応用に近い自然なコンテキストが入力された際の挙動を調べるため, 各質問毎に positive passage と negative passage をそれぞれ重複なしで指定した個数サンプル[7]した場合の正答率を測った. 結果は表 1 の上部のようになっており[8], 以下のような傾向が観察された.

(1) 同じ positive passage の集合を与えられていても, negative passage の個数が増えると正答率が低下する.

(2) (1) の傾向は, positive passage の個数が少なくなるとより顕著となる.

より個数変化のみによる影響を評価するため, ある 1 つの negative passage を指定した個数だけコピーしてコンテキストに加えた場合の正答率も評価した (以後この設定を「重複コピー設定」と呼ぶ)[9]. 結果は表 1 の下部のように前述した (1) と同様の結果が得られた. このことから (1) の結果は含まれる情報の内容に一切の変化が無く個数のみが変化する場合でも起こることが確認できた.

さらに (1) の結果がどの程度 negative passage 中の矛盾する情報[10]や正答の表記ゆれ由来の効果によるものなのかを明らかにするため, negative passage の追加によって回答が変化した事例[11]を 100 個サンプルして人手評価を行った. その結果, positive passage のみ与えられていた場合にモデルが正解できていた事例は 87 個存在し, そのうち変化後の回答が正答の表記ゆれだった事例は 16 個 (18%), negative passage 中に正答に矛盾する情報が含まれていた事例は 20 個 (22%) であった. この結果から, 残りのおよそ半数の事例については negative passage に矛盾する情報が含まれないにも関わらずモデルが誤答を出力してしまっていることが確認された.

表 1 positive および negative passage の個数の変化による質問応答モデルの正答率 (exact match) の変化. all は各質問毎に利用できる positive または negative passage を全て利用したことを示す.

# pos \ # neg	0	1	2	4	8	16	32	64	all
重複なしでサンプリングした場合									
1	0.468	0.405	0.369	0.318	0.267	0.216	0.168	0.117	0.108
2	0.545	0.511	0.485	0.447	0.393	0.339	0.275	0.208	0.190
4	0.612	0.594	0.579	0.554	0.516	0.468	0.407	0.329	0.314
8	0.664	0.657	0.653	0.638	0.621	0.590	0.545	0.474	0.460
all	0.732	0.729	0.726	0.720	0.712	0.693	0.669	0.623	0.612
単一の negative passage をコピーした場合									
all	0.731	0.729	0.726	0.719	0.707	0.686	0.654	0.608	-

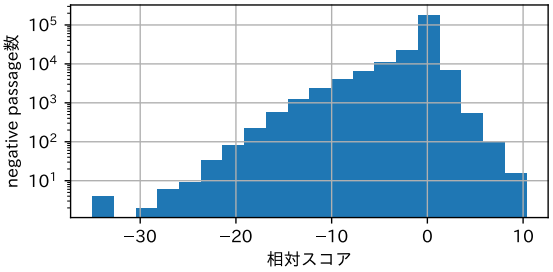

図 1 negative passage が持つ相対スコアのヒストグラム.

3.4 negative passage ごとの性能変化に及ぼす程度の違い

本節では §3.3 で確認された negative passage の悪影響が, 個々の negative passage によってどのように異なるかについて調べる.

3.4.1 hard negative passage

ある negative passage がモデルに悪影響を及ぼしている度合いの指標として, ここでは §3.3 の重複コピー設定でその negative passage を 64 個コピーして

でも同様の結果が確認できたことから, 本実験ではコンテキスト中の文章の順序については学習時, 推論時共に特に意識しないこととした.

6) 本実験ではコンテキスト中の各文章 p_i のタイトル t_i も用いて s_i = "question: q title: t_i context: p_i" とした.

7) 各質問に対しランダムに 5 通りサンプルして結果を平均した.

8) 各行, 各列はそれぞれ同じ positive, negative passage の集合に対する実験結果である.

9) 各質問ごとに, 全ての negative passage に対して正答率を評価し結果を平均した.

10) 例えば正答以外の別解が存在することを示す情報が挙げられる. このような矛盾を含むコンテキストに対する分析も行われているが [19, 20], そのようなコンテキスト中の矛盾に対するモデルの望ましい挙動は応用次第で変わりうる. そのため本稿ではコンテキスト中の矛盾に関しては解決の対象外として区別している.

11) 重複コピー設定 (negative passage の個数が 64 個) での結果を利用した.

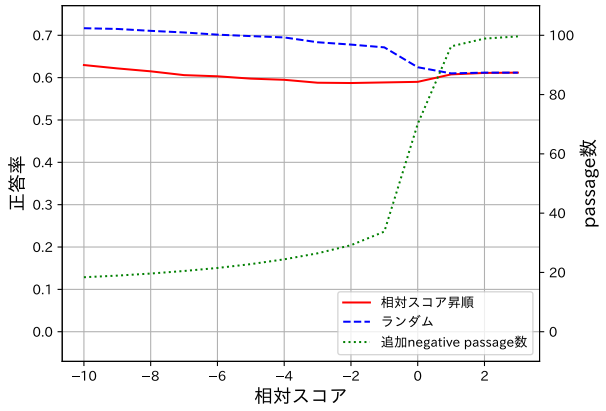

図 2　ある閾値以下の相対スコアを持つ negative passage のみをコンテキストに追加した際の正答率の変化 (赤実線). 青破線は各閾値について同数の negative passage をランダムに選んで追加した場合の正答率の変化を示す.

コンテキストに追加した場合に正答が出力される対数尤度[12]がどれだけ変化したかの「相対スコア」[13]を利用する. negative passage の相対スコア分布は図 1 のヒストグラムのようになっており, ほとんどの negative passage が 0 に近い相対スコアを持っている. 一方で一部の negative passage は大きな負の相対スコアを持っており, モデルが正答を生成することを強く阻害していることがわかる (以後このような negative passage を hard negative passage と呼ぶ).

そこで §3.3 で確認された negative passage が及ぼす悪影響が実際にはこの一部の hard negative passage の影響で引き起こされたのではないかという仮説を検証するため, 相対スコアの昇順に negative passage を追加した際に正答率がどのように変化するかを調べた (図 2). その結果, 相対スコアが昇順になるように negative passage を追加した場合, 20 個程度しか追加していない状態で全 negative passage を追加した場合と同程度の正答率低下を再現できることがわかった. このことから, 相対スコアが 0 に近い大部分の文章はモデルの推論に大きな影響を与えず、主に hard negative passage が性能低下の原因であると考えられる.

3.4.2　hard negative passage の識別可能性

最後に §3.4.1 でモデルの推論に対する悪影響が明らかになった hard negative passage を, その文章の内容から識別可能かどうかを調べた. §2 で述べたように既存研究においていくつかの文章特徴量にショートカット推論との関連性が指摘されている. ここではそれらのうち, 質問と文章の間の最長一致 n-gram 長 [11], 共通単語数 [4], および文ベクトル間のコサイン類似度 [14], そして retrieve-then-read システムの検索システムによる文章の検索順位, の 4 つの特徴量を計算し相対スコアとの間のケンドールの順位相関係数 τ_b[25] によって相関の程度を評価した. 結果は表 2 のようになっており, 弱い相関の存在は確認できたものの hard negative passage かどうかの識別が可能なほどの強い相関は確認できなかった[14].

表 2　各文章特徴量と相対スコアの間の相関係数.

特徴量	τ_b
最長一致 n-gram 長	0.061
共通単語数	0.082
コサイン類似度	0.113
検索順位	0.112

4　結言

本稿では retrieve-then-read 方式の質問応答モデルが, コンテキストとして与えられる文章の量によってどのような影響を受けるかを調べた. その結果質問応答モデルは質問への回答に不要な文章が多くなるほど正答率が低下すること, そしてこの影響が主にモデルの推論に悪影響を与えやすい一部の hard negative passage によって引き起こされていることがわかった. さらにある 1 つの不要な文章に注目したとき, コンテキスト中に何度も重複して出現させることで正答率を下げることができることも確認された.

今後の課題としては, 本稿における実験結果を踏まえて hard negative passage に対してよりロバストな質問応答モデルを学習する方法を検討することが挙げられる.

参考文献

[1] 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, 2016.

[2] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert:

12)　正答が複数存在する場合は, それらが生成される対数尤度のうち最大の値を利用する.

13)　すなわち負の相対スコアは negative passage を追加することで正答の生成確率が下がったことを意味する.

14)　特徴量による相対スコアの分布の変化は付録の図も参照されたい.

A lite bert for self-supervised learning of language representations. In **International Conference on Learning Representations**, 2019.

[3] 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, 2020.

[4] 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, 2018.

[5] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1870–1879, 2017.

[6] Gautier Izacard and Édouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 874–880, 2021.

[7] Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2726–2736, 2019.

[8] Shujian Zhang, Chengyue Gong, and Xingchao Liu. Passage-mask: A learnable regularization strategy for retriever-reader models. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 3931—-3943, 2022.

[9] 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.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474, 2020.

[11] 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, 2017.

[12] 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, 2020.

[13] Wee Chung Gan and Hwee Tou Ng. Improving the robustness of question answering systems to question paraphrasing. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 6065–6075, 2019.

[14] Mingzhu Wu, Nafise Sadat Moosavi, Andreas Rücklé, and Iryna Gurevych. Improving qa generalization by concurrent modeling of multiple biases. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 839–853, 2020.

[15] Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. Why machine reading comprehension models learn shortcuts? In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 989–1002, 2021.

[16] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. Did the model understand the question? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1896–1906, 2018.

[17] Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. Which shortcut solution do question answering models prefer to learn? **arXiv preprint arXiv:2211.16220**, 2022.

[18] Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. Look to the right: Mitigating relative position bias in extractive question answering. **arXiv preprint arXiv:2210.14541**, 2022.

[19] Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. **arXiv preprint arXiv:2210.13701**, 2022.

[20] Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. Defending against poisoning attacks in open-domain question answering. **arXiv preprint arXiv:2212.10002**, 2022.

[21] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 453–466, 2019.

[22] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6769–6781, 2020.

[23] 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. **Journal of Machine Learning Research**, Vol. 21, pp. 1–67, 2020.

[24] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. **arXiv preprint arXiv:1910.03771**, 2019.

[25] Maurice G Kendall. The treatment of ties in ranking problems. **Biometrika**, Vol. 33, No. 3, pp. 239–251, 1945.

A　実験設定に関する詳細情報

表 3　データセットのサイズ

train サブセット		dev サブセット
訓練データ	開発データ	評価データ
61133	8852	2898

モデルの学習には transformers ライブラリの `Seq2SeqTrainer` クラスを利用し, 学習率は 5×10^{-5}, ステップ数は 15000, バッチサイズは 64, weight decay の係数は 0.01, warmup のステップ数は 1000 とした. 開発データでの評価は 500 ステップごとに行い, 評価指標としては開発データでの正答率を利用した. 学習および評価に利用したデータセットのサイズは表 3 の通りである.

B　hard negative passage の普遍性について

ある相対スコアの negative passage を持つ質問の割合を調べた結果は図 3 のようになっており, hard negative passage が一部の質問に限らず多くの質問について存在していることがわかる. また異なるシード値で学習した 5 つのモデルの相対スコアのばらつきは図 4 のようになっており, モデルによってばらつきがあるもののおおむね相関していることが確認された.

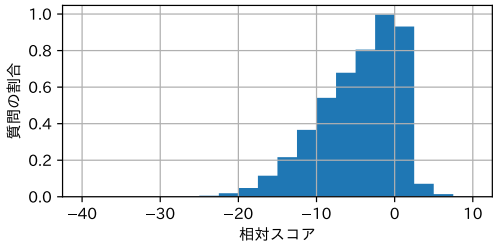

図 3　ある相対スコアの negative passage を持つ質問の割合.

C　negative passage の特徴量による相対スコア分布変化

§3.4.2 の実験における negative passage の各特徴量による相対スコアの分布の変化は図 5,6 のようになっている. ここで青点線, 黒線, 赤破線, 赤点線はそれぞれ 95, 50, 20, 5%分位点である.

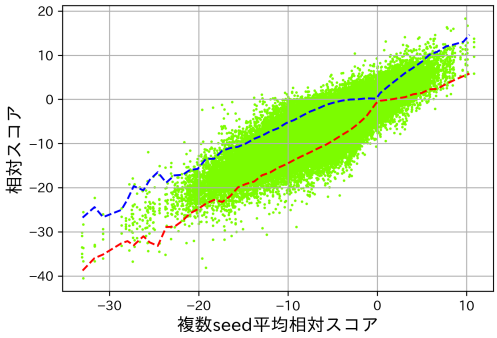

図 4　異なる seed で学習したモデルによる相対スコア分布. 赤, 青破線はそれぞれ 5,95%分位点.

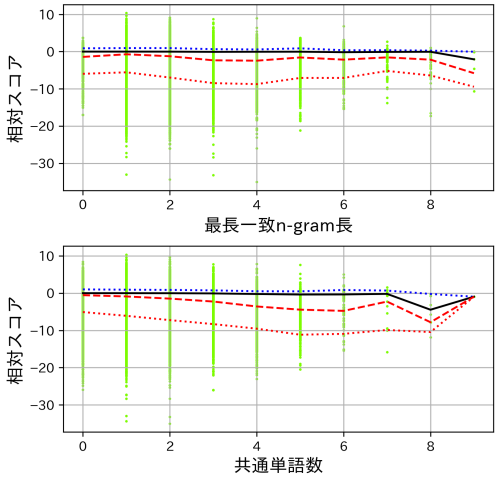

図 5　最長一致 *n*-gram 長および共通単語数による相対スコア分布変化.

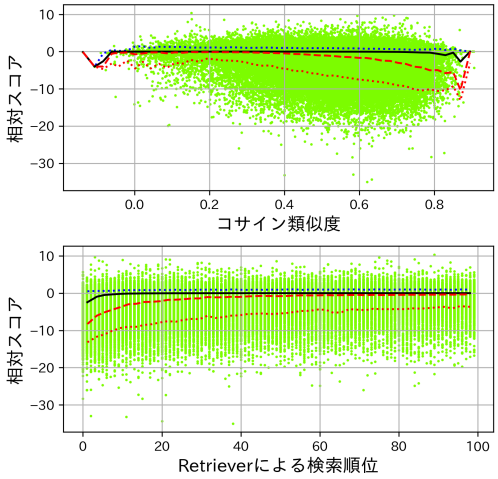

図 6　コサイン類似度および検索順位による相対スコア分布変化.