対話型質問応答における質問書き換えのためのターン強調

小堀 智祥　小林 哲則　林 良彦
早稲田大学理工学術院
kobori@pcl.cs.waseda.jp

概要

質問応答を含む対話システムでは，対話内容に沿った質問の解釈が重要である．先行研究では，すべての対話履歴と質問を一つの繋がった文章として入力し，対話履歴に依存しない質問へ書き換えることが行われているが，書き換えに関係のあるターンを考慮することが難しい．本研究では，対話中の質問応答における各ターンの重要度を推定し，その重要度に応じて強調した表現を用いて質問書き換えを行う TurnRewrite モデルを提案する．質問書き換えの上で重要なターンへルールによるラベル付けを行ない，各ターンの重要度を推定するターン強調モデルを学習する．CANARD データセットを用いた実験結果により，TurnRewrite モデルの有効性を確認した．

1　はじめに

質問応答モデルの精度が向上し，与えられた文書に関する単一の質問に答えるタスクは人間の性能を超えるものも出てきている [1]．近年，より現実に即したタスクとして，対話型質問応答 (Conversational Question Answering: CQA) タスクと呼ばれる会話の中で現れる質問に対する回答を生成するタスクが提案されている．このタスクは質問内容の解釈が対話履歴に依存するという点で，単一の質問に答えるタスクと比較して難しい [2]．このタスクのサブタスクとして，対話履歴に依存して解釈の決まる質問を対話履歴に依存しない質問 (以降，書き換えられた質問) に書き換える質問書き換え (Question Rewriting) タスクが提案され，CANARD と呼ばれるデータセットが構築されている [3]．

CANARD におけるデータの例を図 1 に示す．この例において Q_i, A_i はそれぞれ質問と回答のターンを表す．Q_1 は与えられた文書に対する最初の質問であり，Q_2 以降はそれ以前の対話に依存した質問になっている．CANARD データセットでは書き換えるべき質問 Q_i と，それに先行するすべての対話履歴 $\{Q_j, A_j\}_{j<i}$ が与えられ，クラウドワーカーによって対話履歴に依存しない形式に書き換えられた質問（図 1 右）が正解として付与されている．

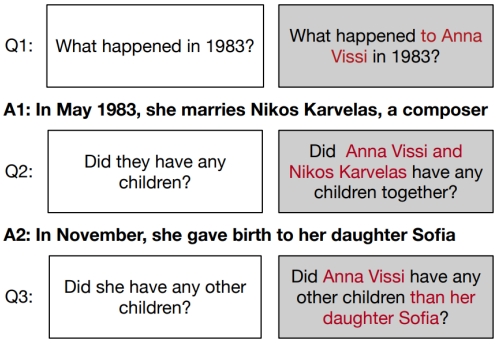

図 1　書き換えデータセットの例 [3]

このタスクに取り組んだ先行研究では，事前学習済み言語モデル（Pretrained Language Model: PLM）に対話履歴と質問を 1 つの繋がった単語系列として入力しているため，対話履歴に含まれるターンがそれぞれ別の内容を持つ文章として区別する表現力が低いことが考えられる．その結果，ターン単位のトピック遷移をうまく学習することができず，複雑な履歴参照が必要な質問の書き換えにおいて誤ったエンティティの参照などが発生していると考えられる．

そこで本研究では，対話履歴に含まれるそれぞれのターンの重要度を推定し，この重要度を用いて質問書き換えを行うことを提案する．既存の CANARD データセットには，対話履歴のターン毎に質問書き換え上の重要度が付与されていないため，本研究ではルールによりターンへの重要度のラベル付けを行い，ターンの重要度を推定するターン推定モデルを学習した．

2　関連研究

CQA タスクの代表的なデータセットとして，物語についての質問応答系列を収集した CoQA デー

タセット [4] や Wikipedia の記事について収集した QuAC データセット [5] が存在する．これらの CQA データセットの各データは，a.) 答えを探す根拠になる文書，b.) 対話履歴，c.) 質問という 3 つの構成要素から成る．対話履歴を考慮する場合，その解釈が必要となる (History Modeling [2]) ．また，質問応答システムへの入力文長の制限が問題となる [6]．これらの課題は，BERT や T5 といった近年の事前学習済み言語モデル (PLM) においても問題となる．

対話履歴の扱いに関する先行研究として，初期には RNN や Attention 機構を用いたモデルが提案されており [4, 7]，またより対話履歴を重視したモデルとしては，対話履歴内の前の質問に答えた際のモデル内部の潜在表現を次の質問に渡すモデル [8, 9, 10] が提案されてきた．また別のアプローチとして，対話履歴とそれに依存している質問から，対話履歴に依存しない質問を生成することで CQA タスクを一問一答形式の質問応答タスクに帰着させる研究がある [3]．これらはモデルが解釈したユーザの質問意図をユーザにフィードバックでき，誤りをユーザが認識しやすいという点と，単一の質問に答える精度の高い質問応答システムやデータセットを利用できる点で他のアプローチよりも優位である．

対話履歴に依存しない質問を生成するタスクを質問書き換え (Question Rewriting: QR) タスクといい，本研究ではこのタスクに取り組む．このタスクを代表するデータセットに CANARD データセット [3] がある．このデータセットを扱った先行研究では，RNN ベースのモデル [3, 11] や PLM を利用したモデル [12]，共参照解消に着目したモデル [13]，CQA と QR タスクを同時に学習させるモデル [14] などが提案されている．先行研究が対話履歴と質問を単語系列として処理するのに対し，本研究では対話履歴に含まれるそれぞれのターン間の関係性を学習するためのアーキテクチャを導入する，すなわち，ターン単位で履歴を扱う点がこれらの先行研究と異なる．

3 提案手法: TurnRewrite モデル

本研究が提案する TurnRewrite モデルの構成図を図 2 に示す．本モデルは，複雑な履歴参照が必要な質問の書き換えに対して，対話履歴に含まれるターンごとの潜在表現を利用してターン単位の重要度を考慮することで書き換え精度向上を図る．具体的には，単語単位による処理に加え履歴ターン単位で対話履歴の重要度を評価し，その結果を書き換え質問生成に利用する．

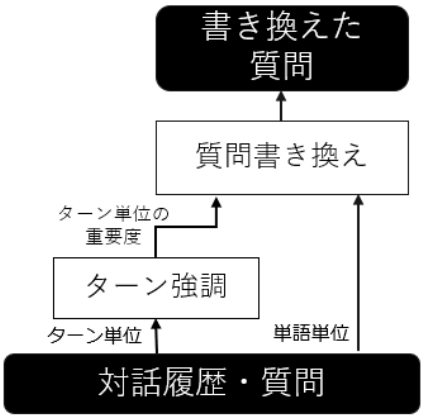

図 2 TurnRewrite モデル構成図

3.1 人工的なデータセットの構築

TurnRewrite モデルでは，書き換えられた質問を生成するために必要なターンをすべてのターンから選択し強調することが必要であるが，既存のデータセットではターン毎に質問を書き換える上での重要度でのラベル付けが行われていない．そこで本研究では，事前に CANARD データセットに含まれる対話履歴と書き換え前後の質問を用いるルールベースにより，対話履歴の各ターンに対してそのターンが書き換え質問の生成において必要か不要かのラベル付け (必要:1，不要:0) を行った．

手順としては，まず，CQA データセットに含まれる書き換え前と書き換えた質問を比較し，書き換えに使われた単語の集合を求め，事前に決められたストップワードを取り除くことで対話履歴から導入された単語を求めた．次に，これらの単語が対話履歴のどのターンから取り出されたものなのかを割り当てる．対話履歴を最も質問に近いターンを調べ，上記の手段により求めた単語が 1 つでもそのターンに含まれていれば，ターンと単語を関連付けてターンに必要とラベル付けした．続くターンを調べる際は，それまで調べたターンと関連付けられた単語を除く，書き換えで導入された単語がターンに含まれるかを調べた．この手順を具体例で示したものを付録 A に示す．

3.2 ターン強調モデル

ターン強調モデルの構成を図 3 に示す． ターン強調モデルは対話履歴と書き換え前の質問を入力として対話ターンごとに書き換える上での重要度を出力する．まず，対話履歴のそれぞれのターンの

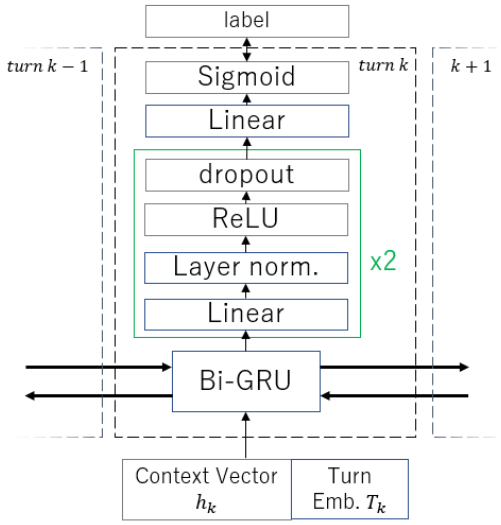

図 3　ターン強調モデル構成図

文 $h_1, ..., h_k$ と，書き換え前の質問 Q_{raw} について，文エンコーダを利用して文表現を取得する．各対話ターンの表現を h'_i と表す．このベクトル表現に，履歴を扱う上で重要と考えられるターンの番号を表す学習可能ベクトルを連結し，履歴ターンのベクトルを生成する．続いて，対話履歴の前後に質問を付け足した $[Q'_{raw}, h'_1, \ldots, h'_k, Q'_{raw}]$ を bi-GRU を経由させることでターン間の関係性を取り込んだ内部表現を作る．最後にその表現を多層パーセプトロンに入力することで，それぞれの履歴ターンに対して $[0, 1]$ 範囲の重要度の予測を出力する．

3.3　質問書き換えモデル

質問書き換えモデルの構成を図 4 に示す．本モデルは，書き換え前の質問，対話履歴，履歴ターン毎の必要度ラベルを入力として，対話履歴の文脈なしに質問意図が理解できる質問を出力する．

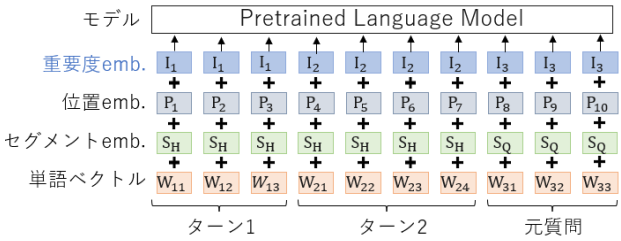

図 4　質問書き換えモデル構成図

本モデルは，書き換え前の質問と対話履歴を並べて 1 つの文章として PLM に入力してファインチューニングするが，モデルの入力となる単語ベクトルに対話履歴の重要度に応じたベクトルを加算している点が通常の PLM を利用するモデルとは異なる．この手法は，入力のうち重要な部分にマーカーを引く様子をモデル化した [15] の手法を参照している．加算するベクトル $turn\_importance\_emb$ はターン強調モデルが単語の属するターンに対して出力する重要度である $importance(h_i)$ と学習可能ベクトル I, U を用いて式 1 で計算される．すなわち，同じターンに属する単語のベクトルそれぞれに対して常に一定のベクトルを足し合わせている．なお，I と U はそれぞれ最も重要・非重要なターンであるを示す学習可能なベクトルである．

$$\begin{aligned} &turn\_importance\_emb(h_i) \\ &\quad = importance(h_i) \times I + (1 - importance(h_i)) \times U \end{aligned} \tag{1}$$

3.4　学習過程

TurnRewrite モデルの学習は 2 ステップで行った．まず 3.1 節で作成したデータセットを利用してターン強調モデルがルールで割り当てられたラベルを出力するように Binary Cross Entropy ロスで学習させた．この際，ラベルに不均衡があったため，バッチ内のラベルの出現回数の逆数の重みをロスに掛け正規化した．続いて，ターン強調モデルと質問書き換えモデルを接続し，質問書き換えモデルを teacher forcing しながら次トークンを予測させる Cross Entropy ロスを最小化することで提案の TurnRewrite モデル全体を学習させた．

4　実験

4.1　実験設定

CANARD データセット [3] (統計情報を付録 B に示す) を用いて質問書き換えの評価実験を行った．対話トピック，書き換えるべき質問，先行する対話履歴を入力とし，対話履歴を見ないで解釈が定まる質問を出力することがタスクとなる．実験においては 文エンコーダ として SimCSE [16] を用いた．また質問書き換えモデルとして関連研究 [12] で最も良い精度を達成した PLM である T5 (T5-base) を用いた．SimCSE および T5 は Hugging Face Transformers ライブラリによる実装を利用し，bi-GRU や実験環境は Pytorch を用いて実装した．学習時の学習率，ターン強調モデルにおける層の深さや幅はグリッドサーチでパラメタチューニングを行い，オプティマイザは AdamW を利用した．

4.2 実験結果

表 1 BLEU スコアによるモデル性能評価

model	BLEU
Human	59.9
T5 (baseline)	55.3
TurnRewrite	56.1
upperbound	58.7

BLEU スコアによる各モデルの書き換え精度の評価結果を表 1 に示す．ここで，Human は [3] にて報告されたクラウドワーカーが書き換えを行った際の BLEU スコアである．upperbound は，ターン強調モデルの gold データを利用した場合を示す．すなわち，ターン強調モデルが正解するという設定であり，TurnRewrite モデルの書き換え精度の上限を与える．なお，Human 以外のそれぞれのスコアは seed を変えて 3 回平均を取っている．

表 1 の T5[1] と TurnRewrite モデルの BLEU 値の比較からターン強調の効果が確認できる．TurnRewrite モデルと upperbound の比較からは，性能向上の余地があることが示唆される．基本的にはターン強調モデルの改善により性能向上が可能であるが，ターンの必要性に関するラベルの品質も問題となる．すなわち今回利用したデータセットである CANARD では，書き換えられた質問として正解が 1 例しか与えられておらず，クラウドワーカーによる誤りや解釈の違いが存在すると考えられる．例えば，[3] によると 10%の書き換え質問は履歴参照が残っている．

4.3 質問形式ごとの効果

CANARD データセットにおける質問を大きく「列挙型」と「置換型」に分類し [2]，それぞれの質問種別のデータをもとにモデルを個別に学習させ，複雑な履歴参照が必要な列挙型質問の書き換えにおけるターン強調モデルの有効性について評価した．

それぞれの質問形式の例を表 2 に示す．列挙型は，質問書き換えにおいて対話履歴におけるエンティティの列挙が必要な質問群である．本研究では，”any other”などの特定の語句を含むものを列挙型として扱った．また，列挙型意外の質問を置換型に分類した．これらの質問の多くは代名詞の参照解決を必要とする．4.2 節と同じモデルで学習させた結果を表 3 に示す．結果から，より履歴参照が必要である列挙型ではベースラインである T5 を利用した場合に比べてターン強調モデルを利用した場合の方が性能が向上することがわかった．これは複雑な参照を含むデータについては履歴ターンの強調が有効であることを示唆している．一方，置換型の T5 と upperbound を比較すると，置換型の質問種別に対しては T5 のみで高い精度を出しており，ターン強調モデルによる履歴ターンの強調があまり効果を持たないことが示された．これは置換型は列挙型に比べ 1 ターンに含まれる人名などを参照することで質問を書き換えることができ，対話履歴に含まれるターンの関係性を考慮したターン強調を行う有効性が低いためであると考えられる．また履歴強調モデルを学習させた TurnRewrite と弱教師データセットを利用した upperbound を比較すると，ターンの誤選択による性能劣化が存在することがわかった．これは，履歴ターンの強調よりもターンの誤選択による性能劣化が大きいことを示唆している．

表 2 列挙型，置換型質問の例

質問形式	元質問 → 書き換え質問
列挙型	Any others? → Any other {songs were featured on films besides ”Dear Lover”, ”Walking After You” for the ...}
置換型	Who did they play in the super bowl? → Who did {the Colts} play in the super bowl?

表 3 質問種別ごとのモデル性能 (BLEU スコア)

model	列挙型	置換型
T5 (baseline)	34.3	62.4
TurnRewrite	41.7	59.8
upperbound	42.5	63.0

5 まとめ

本研究では CQA タスクのサブタスクである QR タスクに対してターン単位で履歴から重要なターンを強調することによる書き換え精度向上を目指した．ターン強調モデルの学習を行うため，履歴のそれぞれに対して重要であるか否かのラベル付けをルールで行ったデータセットを作成した．評価実験の結果から，ターン強調をする機構がないモデルに比較して書き換え精度の向上が見られたことから，ターン強調の有効性が確認できた．

1) [12] の報告では BLEU が 58.1 と報告されているが，実験設定が明記されておらず再現することができなかった．
2) Train/Dev/Test のデータ数はデータ数が最も小さい列挙型に揃えた．結果，Train, Dev, Test はそれぞれ 5595, 631, 1060 質問からなるデータセットになった．

参考文献

[1] Hariom A. Pandya and Brijesh S. Bhatt. Question answering survey: Directions, challenges, datasets, evaluation matrices. **CoRR**, Vol. abs/2112.03572, , 2021.

[2] Gupta Somil, Rawat Bhanu Pratap Singh, and Yu Hong. Conversational machine comprehension: a literature review. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 2739–2753, 2020.

[3] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. Can you unpack that? learning to rewrite questions-in-context. In **Empirical Methods in Natural Language Processing**, 2019.

[4] Siva Reddy, Danqi Chen, and Christopher Manning. Coqa: A conversational question answering challenge. **Transactions of the Association for Computational Linguistics**, Vol. 7, No. 0, pp. 249–266, 2019.

[5] Choi Eunsol, He He, Iyyer Mohit, Yatskar Mark, Yih Wen-tau, Choi Yejin, Liang Percy, and Zettlemoyer Luke. QuAC: Question answering in context. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2174–2184, 2018.

[6] Zhao Jing, Bao Junwei, Wang Yifan, Zhou Yongwei, Wu Youzheng, He Xiaodong, and Zhou Bowen. RoR: Read-over-read for long document machine reading comprehension. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 1862–1872, 2021.

[7] Chenguang Zhu, Michael Zeng, and Xuedong Huang. Sdnet: Contextualized attention-based deep network for conversational question answering. **ArXiv**, Vol. abs/1812.03593, , 2018.

[8] Hsin-Yuan Huang, Eunsol Choi, and Wen tau Yih. Flowqa: Grasping flow in history for conversational machine comprehension. **ArXiv**, Vol. abs/1810.06683, , 2018.

[9] Yeh Yi-Ting and Chen Yun-Nung. FlowDelta: Modeling flow information gain in reasoning for conversational machine comprehension. In **Proceedings of the 2nd Workshop on Machine Reading for Question Answering**, pp. 86–90 url =, 2019.

[10] Yu Chen, Lingfei Wu, and Mohammed J. Zaki. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. In **Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20**, pp. 1230–1236, 2020.

[11] Vakulenko Svitlana, Longpre Shayne, Tu Zhucheng, and Anantha Raviteja. Question rewriting for conversational question answering. p. 355–363. Association for Computing Machinery, 2021.

[12] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. **CoRR**, Vol. abs/2004.01909, , 2020.

[13] Tseng Bo-Hsiang, Bhargava Shruti, Lu Jiarui, Moniz Joel Ruben Antony, Piraviperumal Dhivya, Li Lin, and Yu Hong. CREAD: Combined resolution of ellipses and anaphora in dialogues. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3390–3406. Association for Computational Linguistics, 2021.

[14] Kim Gangwoo, Kim Hyunjae, Park Jungsoo, and Kang Jaewoo. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 6130–6141, 2021.

[15] Qu Chen, Yang Liu, Qiu Minghui, Croft W. Bruce, Zhang Yongfeng, and Iyyer Mohit. Bert with history answer embedding for conversational question answering. p. 1133–1136, 2019.

[16] Gao Tianyu, Yao Xingcheng, and Chen Danqi. SimCSE: Simple contrastive learning of sentence embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910. Association for Computational Linguistics, 2021.

A　対話履歴のラベル付けルール

対話履歴のラベル付けの手法を図 5 を用いて説明する．

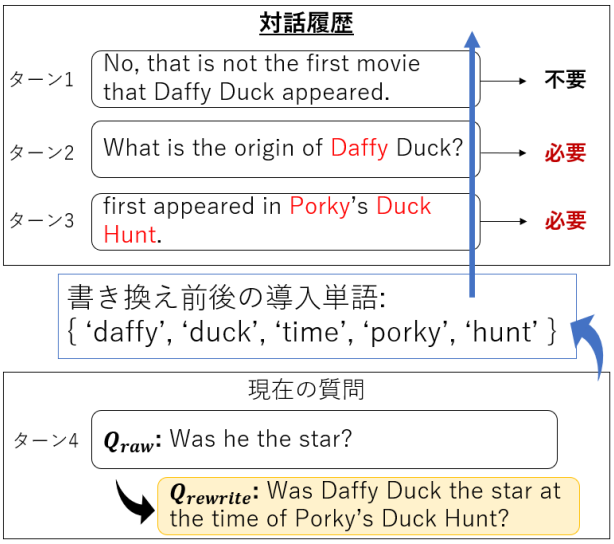

図 5　ラベル付けの手法

この例では，図上部に示すようにターン 1 からターン 3 までの 3 ターンが会話履歴として与えられ，図下部に示すようにターン 4 の質問を書き換える場面を想定している．まず，ラベル付けのためにデータセットから 4 ターン目を人手で書き換えられた質問を用意する (図黄色背景)．次に，4 ターン目の書き換え前後の単語を比べ，書き換えられた質問に増えた単語を求める (図の例では”daffy”など，中央青枠に示した)．ここで，”at” や ”the” といった単語も増えているが，これらはストップワードのため除外されている．続いて，これらの単語を含む対話履歴を「書き換えに必要」とラベル付けしていくが，このとき最も新しいターン（例では 3 ターン目の”first appeared ...”）から調べる．例では最も新しいターン 3 をまず調べ，{”Porky”, ”Duck”, ”Hunt”}が含まれるため，必要とラベルされる．続いて 2 ターン目も同様に調べるが，この際には 3 ターン目に含まれていた{”Porky”, ”Duck”, ”Hunt”}以外の単語とマッチングを行う．ここでも{”Daffy”}が含まれるため必要とラベルされる．1 ターン目も同様に調べると，確かに”Daffy”や”Duck”が含まれるが，これらはターン 2, 3 でマッチングされているため，ここではマッチングが起こらず，ターン 1 は不要とラベルされる．

B　CANARD データの統計

CANARD データセットは QuAC データセットの Dev セットと一部の Train セットを利用して作成されている．CANARD データセットの各種統計を表 4 に示す．

表 4　CANARD データセットの統計

	Train	Dev	Test
[Q]uestions	31,526	3,430	5,571
Dialogs	4,383	490	771
turns / history	8.0	8.6	8.8
words / raw Q	6.5	6.4	6.6
words / rewritten Q	10.0	9.9	9.8

特に，書き換えられた質問の単語数と書き換え前の質問の単語数の比率を図 6 に示す．

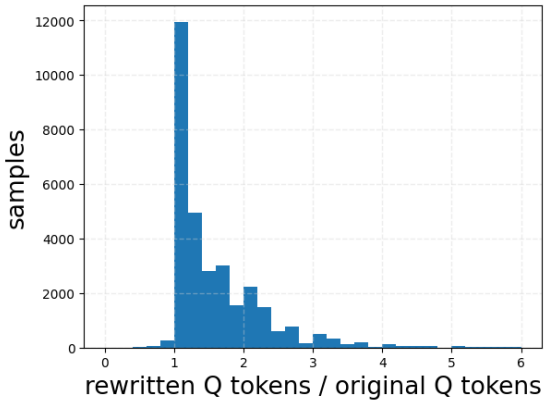

図 6　CANARD Train データセットにおける書き換え前後の単語数の比率

図 6 より，多くの書き換えた質問の単語長は書き換え前の単語長の 2 倍を下回るが，それ以上になる質問書き換えも一定数存在することが確認できる．