対話型質問応答における参照の影響

中西 真央　　小林 哲則　　林 良彦
早稲田大学理工学術院
nakanishi@pcl.cs.waseda.ac.jp

1 はじめに

最近，機械によるテキストの読解能力の開発や測定を目的とする Machine Reading Comprehension (MRC) が盛んとなっている．この質問応答部分を対話形式とした対話型 MRC のデータセットが複数公開され [1, 5]，対話型質問応答の研究が活発化している．対話型 MRC における質問は，従来の MRC とは異なり対話中に位置付けられるため，各質問は対話の文脈に大きく依存する．このため，対話履歴を参照しないと理解することが難しい (と想定される) 質問が多く存在する．

本研究では，対話型 MRC において特有な，文脈に依存する質問を対象とし，特にそこで用いられる指示語の参照先を解消する，あるいは参照しないことの回答正解率への影響を調査した．

具体的には，公開されている参照解消器 (Stanford coreNLP [2]) を数種類の設定において利用し，参照解消の結果による置換を行って質問を変形した．これを，標準的と考えられる 読解モデルで学習した際の，対話中の質問に対する回答の正解率を比較した．その結果，正解率は参照解消の方法によらず, 置換を行わない場合より劣っていた．また，恣意的に意味のない語に置き換えた場合においても正解率に大幅な減少は見られなかった．以上より，利用した MRC モデルは指示語に着目することなく回答を求めていたことが示唆される．他の可能性としては，参照解決の失敗が原因であることが考えられるが，正しい参照解消の結果を反映した大規模な学習は現実的ではなく，その検証は困難である．

2 背景

2.1 MRC およびその近年の傾向

MRC は近年大きな注目を集めており，多数のデータセットが公開されている．その多くのデータセットは，読解の対象となる文章，この文章の理解を試すための質問，および，その答えから構成される．最近公開されたデータセットの多くは, 機械の言語理解能力をより良く調べる目的で設計されている．具体的には，質問と文章中に登場する単語の重複を避けたり，複数文の読解を必要とする質問を増やすことで，単純な単語一致や言い換えによって回答できる質問の数を減らす傾向がある．

2.2 対話型 MRC

従来の MRC データセットに含まれる質問は一問一答形式ばかりであり，質問間に依存関係はないが，最近，情報源となる文章についての質問応答対話から質問と答えを整理したデータセットが公開された．このようなデータセットに含まれる質問は，直前までの質問応答対話の内容に大きく依存する．これは，質問応答対話の進行に伴い，知識の獲得や興味の展開が行われ，これが質問に反映されるためである．

この質問間の依存性によって，対話型 MRC は従来の MRC では見られなかった，以下にあげる 2 つの特徴を持つ [7].

1. 質問中に直前までの質問に登場した内容を指し示す指示語が登場する．
2. 文章中の答えの出現位置は，質問が対話中の何番目に位置するかなどの対話履歴に影響を受ける．

3 データセット

本研究では，Question Answering in Context (QuAC) [1] と呼ばれる対話型 MRC のデータセットを利用する．このデータセットは，質問側，応答側の 2 人組のクラウドワーカーによって行われた質問応答対話から作成されたデータセットである．文章と質問，答えを構成要素として持ち，質問に対する答えを文章中からの抜き出しにより求めるところは，MRC の代表的なデータセットである Stanford Question Answering Dataset (SQuAD) [4] と共通している．

文章は，Wikipedia の人物を主題とする記事から選択された 100 記事の 1 セクションである．質問側は，セクションの題目と文章の最初の 1 パラグラフを与えられ，これに基づき自由に質問する． 応答側は，基本的には与えられた文章中から回答区間を抜き出すことにより答えを返す．1 つの質問応答対話は，12 の質問が問われる，または，答えられない質問が 2 度以上発せられるまで継続する．QuAC における質問応答対話の例を図 1 に示す．

評価指標は，SQuAD で用いられた単語レベルの F1 スコアの平均が用いられる． 本研究でもこの評価指標を用いる．

また，応答側は抜き出した答えを回答する際には，対話を円滑に進めるための以下の 3 種類の対話行為 (dialog acts) も質問側に提供する．

- continuation (follow up/may be follow up/don't follow up)
- affirmation (yes/no/neither)
- answerability (answerable/no answer)

文章と質問から上記の対話行為を推測することもタスクに含まれているが，本研究では答えの抜き出し部分のみに注目する．

4 指示語の置き換え

対話型 MRC の質問には，対話履歴の内容を指し示す指示語が多く登場する．この質問中に含まれる指示語の参照を解決した質問の学習した結果は，指示語のままで学習した結果を上回るという予想ができる．本研究では，指示語を他の語 (参照解消の結果の単語または無意味な文字列) に置き換えた質問を学習，評価することで，質問中の指示語がシステムに与える影響について調査する．

4.1 指示語の参照解決方法

質問中の指示語とそれが指し示す語の探索，置換する方法として以下の 4 つを用いた．方法 1，2，3 は，指示語をそれが指し示す語に正しく置き換えるための置換方法である．方法 4 は，指示語そのものが質問への回答に影響するかを調査するための置換方法である．

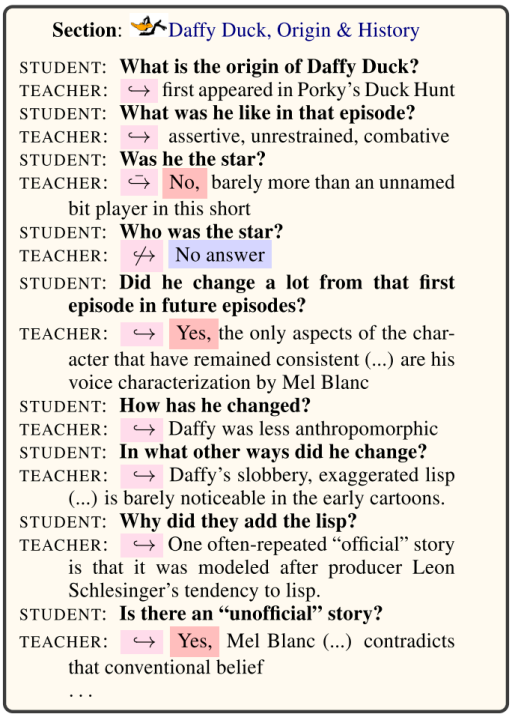
Section: Daffy Duck, Origin & History
STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .

図 1: QuAC 質問応答対話例 ([1] から引用)

方法 1 直前まで (1〜 n-1 番目) の質問と n 番目の質問を番号の順に並べた文章を参照解消器へ入力する．n 番目の質問に含まれる語が他の語を参照していた場合，システムの出力にしたがって，参照元の語を参照先の単語に置き換える．

方法 2 対話中で最初 (1 番目) の質問を，答えを含む平叙文に変換する．平叙文の形にした 1 番目の質問とそれ以降 (2〜n-1) 番目の質問，および n 番目の質問を番号の順に並べた文章を参照解消器へ入力する．n 番目の質問に含まれる語が他の語を参照していた場合，システムの出力にしたがって，参照元の語を参照先の単語に置き換える

方法 3 直前まで (1〜n-1 番目) の質問を全て答えと合わせて平叙文に変換する．平叙文の形である 1〜n-1 番目の質問と，n 番目の質問を質問を番号の順に並べた文章を参照解消器へ入力する．n 番目の質問に含まれる語が他の語を参照していた場合，システムの出力にしたがって，参照元の語を参照先の単語に置き換える

方法 4 直前まで (1〜 n-1 番目) の質問と n 番目の質問を番号の順に並べた文章を参照解消器へ入力する．

n 番目の質問に含まれる語が他の語を参照していた場合，参照元の語を ”XXX” へ置き換える

方法 1 は答えの情報を含まないため，指示語が答えの内容を示す場合には正しく変換できない．一方で，方法 3 では指示語の指し示す語の探索範囲が広がるため，誤って変換する場合が増えることが予測される．

予備実験として，データセット作成時に質問作成者に提供されるセクションの題目および文章の最初の 1 パラグラフを，参照解消器への入力に加える方法で参照を解決した質問で，5 節に示すのと同様の設定で予備実験を行った．しかし方法 1，2，3 で作成した質問で行った実験結果の精度が上回ったため，参照解決方法として採用しなかった．

4.2 参照解決結果の分析

質問をランダムに約 100 問取り出し，方法 1〜3 による変形を行った．その際の参照解決や変形の成否の状況を以下の表 1 に示す．

表 1: 参照解決結果

	方法 1	方法 2	方法 3	合計
評価した質問数	100	100	100	300
参照未解決質問数	49	47	36	132
変形した質問数	18	21	28	67
誤って変形した質問数	5	5	7	17

表 1 から，3 つの方法全てにおいて，質問中の指示語が未解決のままである場合が全体の 1/3 程度と多いことがわかる．この場合，5 節で述べる学習段階において，指示語が解消された質問と，されていない質問を同時に学習することになり，学習結果の悪化につながる．

指示語が未解決のままである質問が多い原因として，本来連続する文章ではない対話中の質問文と答えを，文章として参照解消器へ入力したことや，指示語が指し示す語に相応しい語が現れない場合があることが挙げられる．

4.3 具体例

以下に変換前と変換後の質問の具体例を示す．

共参照が正しく解決された例
変換前: What else was he known for?
変換後: What else was Leonardo known for?

共参照が誤って変換された例
変換前: What year was Romeo and Juliet produced?
変換後: What year was broadway romeo and juliet produced?
(指示語でない語を参照と見なして変換)

変換前: He was supposed to write the lyrics–what happened?
変換後: The first show sondheim was supposed to write the lyrics – what happened?
(参照先の語を誤って変換)

5 実験

4 節において 4 つの方法で変換した質問を使用し，それぞれについて 2 つの読解モデルで回答の抜き出しタスクを学習，評価した．

5.1 モデル

[1] で使用されたベースラインモデルのうち，BiDAF++ モデルおよび BiDAF++ w/ k-ctx モデルを使用した．

BiDAF++ Bi-Directional Attention Flow (BiDAF)[6] は，文章と質問間で双方向の attention を使用した読解モデルであり，MRC データセットのベースラインとして用いられることの多いモデルである．本実験で使用する BiDAF++は，BiDAF の改良モデルであり，BiDAF と同様に Stanford Question Answering Dataset (SQuAD) [4, 3] において良い精度のモデルである．

BiDAF++ w/ k-ctx BiDAF++モデルを対話型 MRC タスクに対応させるため，対話履歴を考慮するよう変更を加えたモデルである．変更点は以下の 2 点である．1 点目は文章のベクトル化の際，k 個前までの

質問の答えとなる単語にマーカーを word embedding に合体させた点である．もう 1 点は，質問のベクトル化の際，対話中の何番目の質問であるかも入力に加えた点である．本実験では [1] で最も良い結果と報告されている k=2 を使用した．

5.2 結果

各モデルの validation 結果を表 2 に示す.

表 2 から, 置換なしと置換ありの学習，評価結果は全て ±1.2 内に収まり，大きな差は見られなかった．最も良い結果は，置換をしない場合であり，次いで良い結果だったのは方法 4 の 参照元の語を”XXX”に置換した場合であった． 参照元を参照先の語へ置換した 1，2，3 の方法は全てそれらを下回る結果となった．

表 2: 実験 1 各方法で置換した質問の validation 結果 (F1 score)

	BiDAF++	BiDAF++ w/ k-ctx
置換なし	49.50	58.91
方法 1	48.98	58.14
方法 2	48.82	58.48
方法 3	48.49	57.77
方法 4	49.35	58.21

5.3 考察

置換なしの質問，指示語を参照先の語へ置換した質問，および指示語を”XXX”へ変換した質問各々の学習結果に，大きな精度の差は見られなかった．これは，代表的な MRC 読解モデルである BiDAF++や BiDAF w/ k-ctx において，質問中の指示語に注目せずとも質問に回答できる可能性があることを示唆している．

指示語を，それが指し示す語に置換するための方法である，方法 1，2，3 の結果が，置換なしおよび方法 4 の結果に優らなかった原因として，表 1 からわかるように，今回検証した方法 1，2，3 では参照すべき指示語を全て解決できず，置換後のデータセット中は指示語が解決された質問と未解決の質問が混ざった状態であり，学習が困難であったことが考えられる．

本来ならば，手動で指示語の共参照が正しく解決された質問を作成し，読解モデルを評価することで，参照解消器による原因は排除するべきであるが，学習に使用できるほどのデータ量を注釈付けすることは難しい．

6 おわりに

本研究では，対話型 MRC において特有な，文脈に依存する質問を対象とし，特にそこで用いられる指示語の参照先を解消する，あるいは参照しないことの回答正解率への影響を調査した．その結果，正解率は参照解消の方法によらず, 置換を行わない場合より劣っていた．また，恣意的に意味のない語に置き換えた場合においても正解率に大幅な減少は見られなかった．

以上の結果から，今回使用した代表的な MRC 読解モデルにおいて，指示語は質問回答の性能に大きな影響を与えることはなく，質問中の指示語に注目せずとも質問へ回答できる可能性があることを示唆している．

今後の展開として，対話型 MRC のもう 1 つの特徴である，回答の出現位置が対話履歴に影響を受けることに注目し正解率の向上を目指すことが考えられる．

参考文献

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

[2] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014.

[3] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 784–789, 2018.

[4] 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. Association for Computational Linguistics, 2016.

[5] Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge. *CoRR*, Vol. abs/1808.07042, , 2018.

[6] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *CoRR*, Vol. abs/1611.01603, , 2016.

[7] Mark Yatskar. A qualitative comparison of coqa, squad 2.0 and quac. *CoRR*, Vol. abs/1809.10735, , 2018.