Vol.2020-NL-244 No.3
2020/7/3

対話によって曖昧性解消を行う質問応答

中野 佑哉[1,2,a)]　河野 誠也[1,b)]　吉野 幸一郎[1,2,c)]　中村 哲[1,2,d)]

概要：質問応答とは，与えられた質問に対し適切な答えを見つけて提示するタスクであり，機械読解や対話システムなど様々な応用を構成する重要な基本タスクの一つである．これまでの質問応答システムの研究は様々な問題を解決し，いくつかのベンチマークで高い精度を実現してきた．しかしながら，質問応答システムを実際に利用する場合，様々な課題が残されている．その中の一つに，質問応答システムに対するユーザ発話の曖昧性がある．本論文では，解答が一意に定まらない曖昧な質問文に対し，問い返しを行うことによって質問文の意味を一意に定めることを目的とする新たな質問応答タスクを設定した．その上で，この問題のベンチマークとなる質問応答データセットを既存の質問応答タスク向け大規模データセットから変換することにより作成した．また，作成したデータセットに対して既存モデルを用いた際の精度評価実験を行い，どのような問題が存在するか議論した．

1. はじめに

自然言語で書かれた文章の意味を計算機に理解させる機械読解技術が注目を集めている [1]．その中でも昨今質問応答は，機械読解が実現されているかを測る重要なタスクとして，SQuAD[2] や HotpotQA[3] をはじめとした様々なベンチマークが公開されている．これらの研究の進展に伴い，こうした機能を実社会へ応用する機運が高まりつつあり，スマートフォンやスマートスピーカー上で動作する音声アシストシステムの一部として実用化が検討されている．

しかし，ユーザが発する文や質問は時として曖昧であり，既存のシステムが常に正しく動作できるわけではない．こうした曖昧なユーザ発話にどう対応するかについては，発話外の情報を用いる [4]，ユーザに問い返しを行う [5] など，様々なアプローチが検討されている．このように，質問応答技術の実社会への応用に関しては，未だ課題が残されている．

本稿では，特にユーザ発話が曖昧であるような状況を想定し，様々な曖昧性のパターンについて検討する．また，こうした問題を解決する糸口とすることを目指して，既存のデータセットを変換して大規模なデータセットを作成する．この変換に際し，構文構造に着目した質問文の自動変換手法を提案する．また，作成したデータセットを用いて質問応答実験を行い，これまでの手法でこうした問題が十分解かれていなかったことを明らかにする．

[1] 奈良先端科学技術大学院大学
NAIST, Takayama-cho, Ikoma, Nara 630–0192, Japan
[2] 理化学研究所 革新知能統合センター AIP
[a)] nakano.yuya.nw9@is.naist.jp
[b)] kawano.seiya.kj0@is.naist.jp
[c)] koichiro@is.naist.jp
[d)] s-nakamura@is.naist.jp

2. 質問応答と問い返し

まず，本研究の問題意識とそれに基づく先行研究，タスク設定について述べる．

2.1 本研究の問題意識

通常，人同士の対話において話者が相手に質問を投げかけるとき，それまでの対話履歴中に言及された事柄や，一般的に広く知られた事実などを省略すること，また，一度の質問で質問内容が明確でない発話を行うことがしばしばある．こうした場合，回答者は対話的に曖昧性を解消したり，様々な追加情報から正しい質問の意図を読み取ろうとする．

これに対し SQuAD などのこれまでに扱われてきた質問応答のための大多数のデータセットでは，予め一意に定められた回答を得られるような質問文が人手で作成されている．このため，こうしたデータから構築された質問応答システムは，データセット中に出現する形式以外の質問文をうまく受け付けることができない．つまり，データセット中に存在しない曖昧な質問文や，より深い推論が求められる対話中において，既存のデータセットのみを用いて学習されたモデルそのまま適用することは適切ではないと考えられる．例えば，街中の観光案内システムやチャット型の顧客対応システムなど，不特定多数ユーザが使用するシステムの場合，データセット中に出現しない形式での質問文

に対して十分な回答を行うことができない可能性がある．

よって，多くのユーザが使用する，より柔軟な質問応答技術を実現するためには，質問文中の曖昧な語句や表現に対して適切な問い返しをするようなモデルや，その学習に適したデータセットが必要となる．

2.2 先行研究

質問文に対する問い返しの研究として，大塚ら [4] はユーザの真の意図と想定される改定質問をユーザに提示し，質問意図を明確にするモデルを提案した．このモデルにより，質問者は自身の質問意図に近い改訂質問を選択することで，所望の解答を得ることができる．しかし，このモデルはユーザから追加の情報を得る質問を生成するわけではなく，質問者との対話によるインタラクティブな問い返しが十分できているわけではない．また，日本語 Wikipedia からランダムに抽出した 600 記事を基にデータセットを構築しているが，統計的に質問応答モデルを学習するためには十分な量ではない．

また，古川ら [5] はカテゴリと疑問詞を一対一で対応付けした辞書を用いて，格フレームに基づいた問い返し候補文の生成を提案した．古川らの行った実験は，対話中の述語項構造に着目して格フレームを基に問い返しを行うものであるが，曖昧な質問文の質問意図を明確にするために，どのような格と項に対して問い返しを行えばよいかは考察がなされていない．

曖昧な質問文を含むコーパス作成に関する研究として，Sewon[6] らのものが挙げられる．この研究では，曖昧な質問文から考えられうる解答を出力する Multiple Answer Prediction と，曖昧な質問文と複数の解答から各解答が一意に定まるような質問文を生成する Question Disambiguation のタスクを設定した上で，NQ-OPEN[7] を基にクラウドワーカーにより作成された AmbigQA データセットを作成している．本課題は回答に必要な文書の検索もタスクの一部であり，得られた文書と各タスクに応じた入力とを合わせて解答，及び，質問文を出力する．既存の質問応答用データセットを基にクラウドワーカーによる高品質なデータセットを構築したとしているが，その品質担保には多大なコストを要している．

これらに対し，我々は既存のデータセット大規模質問応答ベンチマークを用い，構文構造を考慮したルールによって質問文を変換することにより，曖昧な質問文を含む質問応答データセットの作成を試みる．この手法によって，構文解析や格フレームを用いた問い返しに焦点を絞り，インタラクティブに曖昧性解消を行うことができる質問応答タスクを提案する．

2.3 タスク設定

これまで述べていた問題意識と既存研究の問題点を踏まえ，本研究では曖昧な質問文とその解答ペアから問い返しを行うことによって，オリジナルの質問文を用いた場合と同等の解答精度を実現できる質問応答システムの構築を目指す．

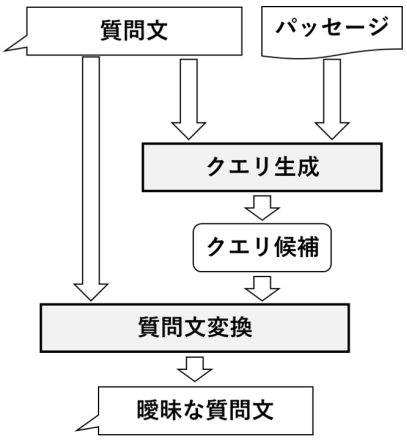

図 1 曖昧な質問文作成の流れ

本研究の構想は以下のとおりである．まず，曖昧性を含む質問文とその解答からなるデータセットを構築する．このデータセットを用いて学習された QA モデルから，適切な解答を得るために必要となる情報をユーザに問い返しを行うエージェントを作成する．問い返しによってユーザから得られた追加情報と最初に得られた質問文とを合わせて質問応答モデルの入力とし，既存の完全な質問文を与えた場合の解答精度達成を目指す．

質問文の曖昧性を解消する既存の試みとして，質問の意図がより明確になるよう言い換えた質問文をユーザに提示するなどの方法がある [4]．これに対し本研究では，曖昧性が項構造解析や格フレームを利用した知識フレームで表現できることを仮定する．これらを用いることで，解答に必要な情報をパッセージを参照することなくユーザから直接得ることができることを期待する．

3. 曖昧なユーザ質問生成

質問応答タスクでは通常，ユーザからの質問とその答えがペアとなったデータセットを用いる．今回我々が設定するタスクにおいても，曖昧な質問文とその答えがペアとなったデータセットが必要となるが，クラウドソーシングを活用した人手による作成方法では，大規模なデータセット作成に多大なコストがかかるという問題がある．そこで我々は，既存の曖昧性を含まない質問文とその答えからなるデータセットから，構文構造を考慮したルールによる質問文の変換によって，今回設定するタスクに相当するデータセット作成を試みた．既存の質問応答タスク向けデータセットとして，HotpotQA データセットを用いた．本節ではこのデータセット作成手順について説明する．

表 1　変換前後の質問文と対応するエンティティの例

エンティティタイプ		質問文
PERSON	元の質問文	Were Scott Derrickson and Ed Wood of the same nationality?
	変換後	Were the person and Ed Wood of the same nationality?
LOCATION		When did the English local newspaper, featuring the sculpture and war memorial in the Forbury gardens, change names?
		When did the English local newspaper, featuring the sculpture and war memorial in the place, change names?
CITY		During Siege of Bharatpur, which Commander-in-Chief of Ireland and Commander-in-Chief of India contributed to the storming of the castle?
		During Siege of the city, which Commander-in-Chief of Ireland and Commander-in-Chief of India contributed to the storming of the castle?
COUNTRY		Where was the world cup hosted that Algeria qualified for the first time into the round of 16?
		Where was the world cup hosted that the country qualified for the time into the round of 16?
NATIONALITY		Who held the record for the longest service in the Australian Parliament for a woman, and was surpassed by a former Australian politician who was the 29th Speaker of the House of Representatives?
		Who held the record for the longest service in the Parliament for a woman, and was surpassed by a former politician who was the 29th Speaker of the House of Representatives?
ORDINAL		Which American film director hosted the 18th Independent Spirit Awards in 2002?
		Which American film director hosted the Independent Spirit Awards in 2002?
OTHERS		What science fantasy young adult series, told in first person, has a set of companion books narrating the stories of enslaved worlds and alien species?
		What science fantasy young adult series, told in first person, has?

表 2　曖昧な質問文への変換によって回答を誤った例

エンティティタイプ		質問文	予測解答
PERSON	元の質問文	Martël Llodra of France, called ”the best volleyer on tour”, defeated Juan Martën del Potro a professional of what nationality?	Argentinian
	変換後の質問文	The person of France, called ”the best volleyer on tour”, defeated Juan Martën del Potro a professional of what nationality?	French
CITY		Which country did the Falkland Islands beat in the 2010 Commonwealth games that has eleven administrative divisions?	Samoa
		Which country did the the place beat city in the 2010 Commonwealth games that has eleven administrative divisions?	New Zealand
NATIONALITY		What’s the name for the character also known as Holger Danske, from medieval French chansons de geste?	Ogier the Dane
		What’s the name for the character also known as the person, from medieval chansons de geste?	Holger Danske
OTHERS		What year was the song that some critics compared to ”Fading” released?	2008
		What year was?	2010

3.1　クエリの作成

曖昧な質問文作成までの流れを図 1 に示す．まず本研究では，質問応答データセット向けのクエリ作成手法として，Qi ら [8] の手法を用いる．この手法は，longest common subsequence (LCS), longest common substring (LCSubStr)，及び，それらを混合させたもので構成されている．この手法を用いて，生成されたクエリを HotpotQA におけるパッセージ検索用クエリに用いる．ここで，パッセージとは，質問文の回答根拠となる文書のことを指す．

図 1 中のクエリ作成を行うため，Qi らの手法を用いて得られた質問文中の一部をクエリ候補とし，これを質問文から欠落させる．ここで，クエリとは，回答が含まれるパッセージ候補を検索するための検索クエリを指す．しかしながら，クエリ候補をそのままの形で質問文から欠落させるだけでは文法規則に沿わない不完全な文が数多く生成されてしまう．そこで，図 1 中の質問文変換部分では，作成されたクエリに含まれる固有表現から，適切な代名詞や言い換え単語を元の質問文中の一致箇所へ置き換えることで曖昧な質問文を作成する．一連の手続きを図 2 に示す．

質問文を変換するにあたり，まず，入力として与えるクエリ中に特定のエンティティが含まれるかどうかを判別する．今回判別する種類のエンティティタグは，”PERSON”，”LOCATION”，”CITY”，”COUNTRY”，”NATIONALITY”，”ORDINAL”とした．まず，”NATIONALITY”，”ORDINAL”と判別できた単語については，それらが純粋な修飾語として表れていれば，質問文中からそのまま欠落させる．それ以外の”PERSON”，”LOCATION”，”CITY”，”COUNTRY”と判別された単語についてはそれぞれ”the per-

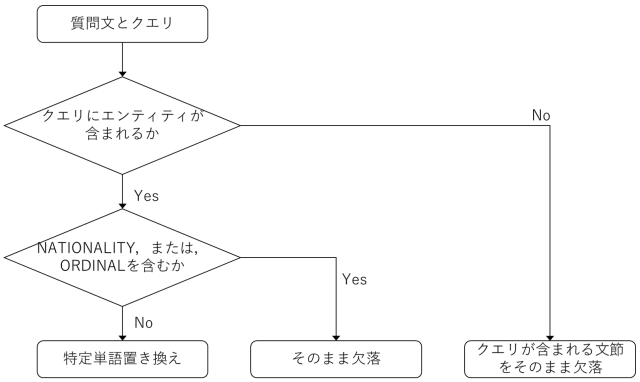

図 2 質問文変換手続き

son”，”the place”，”the city”，”the country”と置き換えて用いる．エンティティが判別されなかった場合は，元の質問文中のうち，作成されたクエリが含まれる文節を削除することにより質問文の変換を行う．質問文の文節抽出，及び，固有表現抽出には，Stanford CoreNLP[9] を用いて得られた構文解析結果を利用した．

また，単純な置き換えのみでは対応できないようなクエリについては，質問文を構文解析した結果に基づく変換処理を行った．構文解析の結果を基に，クエリ中の固有表現や特徴的な語が含まれる文節をそのまま欠落させ，be 動詞や指示代名詞等の周辺単語の調整を含む質問文の変換処理をルールベースで行った．

今回，曖昧な質問文へ変換するために使用した既存の質問応答データセットに含まれる質問文の内，約 11% が”NATIONALITY”，または，”ORDINAL”のエンティティを含むクエリ，約 55% がその他のエンティティを含むクエリ，残りの約 34% がどのエンティティも含まないクエリとして判別された．

3.2 質問文変換の事例分析

変換前後の質問文の事例を表 1 に示す．データセット全体を通して，エンティティが判別できた際の単語置き換えにより変換された質問文は概ね文法に則り，かつ，曖昧性を含む文が生成される結果となった．ここで判別に用いる固有表現認識の精度は Stanford CoreNLP に依存し，分類精度は 90% 程度である[*1][9]．しかし，こうした変換が質問応答モデルの解答予測にどれほどの影響を与えるかについては検証が必要である．そこで次節で，実際にこの変換質問を質問応答システムの入力として用いる実験を行った．

また，構文解析を利用したルールベースの処理では，元の質問文に占めるクエリの割合が大きい場合などは削除対象が大きくなりすぎてしまうなどの問題も存在し，その取り扱いには課題が残った．

[*1] https://nlp.stanford.edu/projects/project-ner.shtml

表 3 実験結果

	EM	F1
HotpotQA-dev (Gold only)	55.92	70.15
Ambiguous HotpotQA (Gold only)	48.56	61.64

表 4 HotpotQA (Distractor Setting) における結果

model	Ans	
	EM	F1
Baseline (Yang, Qi, Zhang, et al. 2018)[3]	45.60	59.02
KGNN (Ye et al., 2019)[11]	50.81	65.75
QFE (Nishida et al., ACL’19)[12]	53.86	68.06
SAE (Tu, Huang et al., AAAI 2020)[13]	60.36	73.58
C2F (Shao, Cui et al. 2020)[14]	67.98	81.24
HGN (Fang et al., 2019)[15]	69.22	82.19

4. 実験

本研究では質問応答タスクにおいて重要な情報が欠落した質問文を完全な質問文から変換して曖昧な質問文を含む質問応答データセットを構築した．この変換が質問応答タスクをどの程度困難にするか，既存の完全な質問文を含むデータセットを用いて学習させたモデルを用いて質問応答評価実験を行った．

4.1 実験設定

本実験では HotpotQA のトレーニング用データセットを用いて質問応答モデルの学習を行い，HotpotQA 開発用データセットから正解箇所が含まれるパッセージのみを抽出した約 7400 ペアと，これらの質問文を基に曖昧な質問文へと変換したデータセットに対して解答精度評価を行った．質問応答モデルは Devlin ら [10] のものを利用した．学習には，事前学習モデルとして BERT-Base-Uncased[10] を用い，バッチサイズ 12，学習率 $3e^{-5}$，エポック数 2.0 とした．

4.2 スコア比較

曖昧な質問文を質問応答モデルに入力として与えた場合の予測解答と正解ラベルとの一致率（EM，F1）に基づいて評価する．ここで EM，F1 はそれぞれ，正解ラベルにおける予測解答の正解率と調和平均を指す．実験結果と HotpotQA Distractor Setting における各既存モデルの結果をそれぞれ表 3，表 4 に示す．表 3 の HotpotQA-dev は HotpotQA オリジナルの質問文を用いた場合のスコアで，Ambiguous HotpotQA は提案手法によって曖昧に変換された質問文を入力として用いた場合のスコアである．この結果から，曖昧な質問文を入力した場合の解答精度は，既存の完全な質問文を用いた際の解答精度に大きく劣る結果となり，曖昧性解消の必要性が改めて確認できる結果となった．

本手法を用いて曖昧な質問文へと変換することによって

誤った回答を行った事例を表 2 に示す．表 2 の OTHERS に着目すると，曖昧な質問文を入力として与えた場合，質問応答モデルはパッセージ中にある複数の年代表記を同定することができず，誤った解答を行っている．このような解答誤りの例が他のエンティティが判別された例でも同様に確認できた．特に”PERSON”などのエンティティが判別され，単語の置き換えのみを行った場合においても質問応答システムの解答予測に誤りが生じることが確認できた．今後は，どのような現象が生じた場合に質問応答モデルの解答精度低下が生じるか，更なる分析を行う必要がある．

5. まとめ

本研究ではユーザからの曖昧な質問文を対話によって解消できる質問応答システムを構築するため，既存の質問応答タスク向けデータセットから構文構造を考慮したルールを用いた変換によって，曖昧性を含む質問応答データセットを構築する手法を提案した．また，変換結果に対して既存の質問応答モデルを用いた評価実験を行った．実験の結果，提案する変換手法によって曖昧な質問が生成され，既存の質問応答モデルの解答精度が低下することが確認された．しかしながら，今回作成したデータセットには，文法的に不自然な質問文が含まれていたり，各質問文の解答難易度が不明瞭であったりと問題点も多く，改善の余地がある．

今後，これら問題を解決し，データセットとしての信頼性を高めつつ，格フレームなどを用いた問い返しにより，柔軟に曖昧性を解消するモデルを構築する予定である．

参考文献

[1] 西田京介，斉藤いつみ，大塚淳史，西田光甫，野本済央，浅野久子：機械読解による自然言語理解への挑戦，NTT 技術ジャーナル (2019).

[2] Pranav, R., Jian, Z., Konstantin, L. and Percy, L.: SQuAD: 100,000+ Questions for Machine Comprehension of Text, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392 (2016).

[3] Zhilin, Y., Peng, Q., Saizheng, Z., Yoshua, B., William, C., Ruslan, S. and Christopher, M.: HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380 (2018).

[4] 大塚淳史，西田京介，斉藤いつみ，西田光甫，浅野久子，富田準二：問い返し可能な質問応答：読解と質問生成の同時学習モデル，日本データベース学会和文論文誌，Vol. 18-J, No. 16 (2020).

[5] 古川智雅，吉野幸一郎，須藤克仁，中村　哲：曖昧性を持ったユーザ発話に対する格フレームを用いた聞き返し発話候補の生成，言語処理学会第 24 回年次大会 発表論文集，pp. 905–908 (2019).

[6] Sewon, M., Julian, M., Hannaneh, H. and Luke, Z.: AMBIGQA: Answering Ambiguous Open-domain Questions (2020).

[7] Tom, K., Jennimaria, P., Olivia, R., Michael, C., Ankur, P., Chris, A., Danielle, E., Illia, P., Matthew, K., Jacob, D., Kenton, L., Kristina, N. T., Llion, J., Ming-Wei, C., Andrew, D. J., Uszkoreit, Q. L. and Slav, P.: Natural Questions: a Benchmark for Question Answering Research, *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 453–466 (2019).

[8] Peng, Q., Xiaowen, L., Leo, M., Zijian, W. and Christopher, D. M.: Answering Complex Open-domain Questions Through Iterative Query Generation, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2590–2602 (2019).

[9] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit, *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60 (2014).

[10] Jacob, D., Ming-Wei, C., Kenton, L. and Kristina, T.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of NAACL-HLT 2019*, pp. 4171–4186 (2019).

[11] Deming, Y., Yankai, L., Zhenghao, L., Zhiyuan, L. and Maosong, S.: Multi-Paragraph Reasoning with Knowledge-enhanced Graph Neural Network, *ArXiv*, Vol. abs/1911.02170 (2019).

[12] Nishida, K., Nishida, K., Nagata, M., Otsuka, A., Saito, I., Asano, H. and Tomita, J.: Answering while Summarizing: Multi-task Learning for Multi-hop QA with Evidence Extraction, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2335–2345 (2019).

[13] Ming, T., Kevin, H., Guangtao, W., Jing, H., Xiaodong, H. and Bowen, Z.: Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents, *AAAI*, Vol. abs/1911.02170 (2019).

[14] Shao, N., Cui, Y., Liu, T., Wang, S. and Hu, G.: Is Graph Structure Necessary for Multi-hop Reasoning? (2020).

[15] Yuwei, F., Siqi, S., Zhe, G., Rohit, P., Shuohang, W. and Jingjing, L.: Hierarchical Graph Network for Multi-hop Question Answering, *ArXiv*, Vol. abs/1911.03631 (2019).