問い返し質問文生成によって曖昧性解消を行う質問応答システム

中野佑哉 [1,3]　河野誠也[1,2]　吉野幸一郎[2,3,1]　須藤克仁[1,3]　中村哲[1,3]
[1]奈良先端科学技術大学院大学
[2]理化学研究所ロボティクスプロジェクト
[3]理化学研究所革新知能統合センター AIP
{nakano.yuya.nw9, kawano.seiya.kj0, koichiro, sudoh, s-nakamura}@is.naist.jp

1　はじめに

質問応答は自然言語の質問を入力としてその回答を出力するタスクである．特に近年の質問応答ベンチマーク [1][2] で定義されているような，回答候補が含まれる文書群を与えた上でそこから回答を導き出すような質問応答は機械読解とも呼ばれ [3]，機械による自然言語の理解がどの程度進んでいるかを測る重要な指標となっている．

こうした既存の質問応答タスクでは，ユーザが発する質問が応答に十分な情報を含んでいることを仮定している．しかし実際の質問応答タスクを考えた場合，システムに与えられるユーザ発話はしばしば曖昧で，正しい回答を導き出すために追加の情報を必要とする．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．特にシステムからの問い返しを行う枠組みは対話的情報検索 [6] と呼ばれ，ユーザとの言語的なやり取りを通じて必要な情報を取得するためにどのような問い返しが必要かということについて検討が行われている [7]．

しかし，こうした問題に対して大規模ベンチマークを構築することは難しい．一番の問題点は，実際の曖昧な質問とその真の意図を大規模に収集することが困難な点にある．また，どのような曖昧さが質問応答システムの回答を困難にするかも明らかではない．そこで本研究では，質問文の文構造に基づいた変換によって，意図が曖昧となる質問を疑似生成する手法を提案する．文構造を用いて質問文中に含まれる情報を選択的に除くことにより，曖昧さを制御した疑似質問文を生成することができる．本論文では特に句に着目した情報選択手法を提案した．

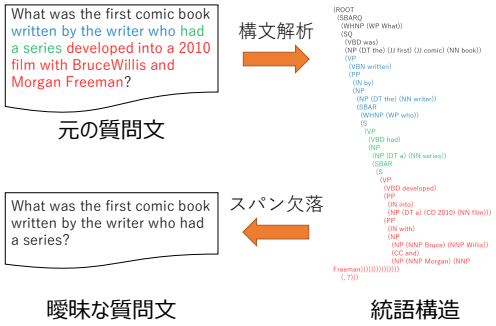

図 1　曖昧な質問文生成の流れ

また，このような疑似生成された曖昧な質問文を与えた場合に，必要な情報を復元する機能が必要となる．そこで，自動生成された問い返し質問文によって必要な情報を回復する枠組みを構築した．この問い返し質問文生成には，格フレームを用いた．

評価では，まず変換に基づく質問文に対する曖昧性付与によって，実際にどの程度既存の質問応答モデルからの回答が難しくなるかについて評価した．その上で，システムからの問い返しに対してユーザが正しく回答した場合，元の質問文への曖昧性付与によって低下した質問応答の精度がどの程度復元されるかについて評価を行った．

2　文構造を用いた曖昧な質問文生成

まず，本論文では質問応答のベンチマークとして広く利用されている HotpotQA データセット [2] を用いる．この既存の質問応答タスク向けデータセットでは，データセット中の質問文が与えられた場合，データセット内の回答候補文書内から回答が一意に定められるような質問文が定義されている．このデータに対して，これまで著者らが検討してきた通り，文構造，特に句に着目した変換によって曖昧な質問文を疑似生成する [8]．本節では，この変換手法の概要と生成質問文の品質評価について述べる．

	EM	F1
Original	55.92	70.15
VP	10.88	28.70
PP	13.73	34.41
Mixed	13.69	33.73

表 1　質問応答モデルの回答精度比較

	VP	PP	Mixed	質問文数
Total	1.928	2.351	2.371	200
Normal Form	2.008	2.492	2.479	71
Irregular Form	1.901	2.265	2.292	129

表 2　主観評価の結果

2.1　質問文変換手法

変換手法の流れを図 1 に示す．元の完全な質問文から曖昧な質問文への変換を行うため，まずデータセット中に定義されている質問文の統語構造を得る．統語構造解析には Stanford Parser[1)] [9][10] を用いる．統語構造として得られた木に対して，VP，PP となる部分木を欠落対象として選択する．選択された部分木を欠落させた文を，該当箇所の情報が欠落した曖昧な質問文として生成して用いる．欠落候補が複数存在する場合は，欠落対象の範囲がもっとも短い部分木を選択する．

2.2　疑似生成された曖昧な質問文の評価

生成された曖昧な質問文に対して既存の質問応答モデルの回答精度を求めることで，実際にどの程度質問応答のために必要な情報が欠落できているかを確認する．また，生成した質問文が文法的におかしい場合は変換が適切でないことが予想されるため，疑似生成結果の質を担保するため英文としての自然さを確かめる主観評価をあわせて行った．

回答精度比較実験（表 1）では，変換後の曖昧な質問文が質問応答モデルの回答精度にどのような影響を与えるか確認する．VP，PP，その両方（Mixed）をそれぞれ欠落対象として生成された質問文を用いた実験結果を示している．表中にある EM は同一の BERT を用いた質問応答モデル [11] が出力する回答の正解率，F1 はそのモデルによって出力される回答の再現率と適合率の調和平均を表す．質問応答モデルの学習には HotpotQA データセットの学習セット部分を，評価には開発セット部分を用いた．Original は変換前の質問文を入力とした場合のスコアを表す．Original のスコアと比較して，全ての変換において質問応答モデルの回答精度が低下しており，提案する変換によって生成した曖昧な質問文がモデルの回答を困難にしていることが確認できた．

1）　https://nlp.stanford.edu/software/lex-parser.shtml

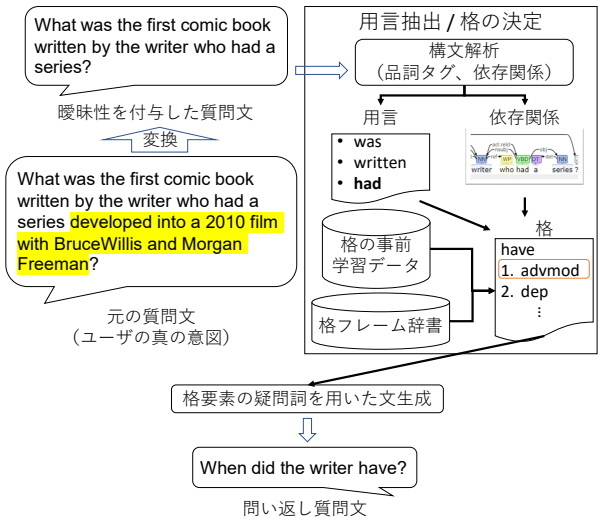

図 2　問い返し質問文生成の流れ

主観評価では，生成した曖昧な質問文が英文として自然かどうかについて被験者 3 名が評価した．この実験では，変換によってモデルの回答が変換前と異なるような質問文を評価の対象とした．元の質問文の統語構造から VP，PP，その両方（Mixed）を欠落の起点とした場合で，それぞれランダムに 200 文ずつサンプリングし，英文としての自然さを 3 段階で評価をしてもらった．各条件に対する評価の平均を表 2 に示す．ここで， Normal Form は元の質問文が疑問詞から始まる一般的な質問文の形であるものを指し， Irregular Form は疑問文以外の形で質問を行っている文を指す．VP のみを用いた変換では，欠落範囲が大きくなりすぎてしまう場合があり，PP のみを用いた変換と比較して自然性が低くなる結果となった．今回の実験結果からは， PP のみを用いる場合と， VP と PP の両方を用いた場合の間に有意な差 ($p < 0.05$) は認められなかった．これは，PP を用いた場合の欠落候補箇所が VP を用いる場合と比較して短く，長さの観点から優先されて多く出現したことが原因として考えられる．また，設定した変換の全てで Irregular Form の質問文に対する評価の平均は Normal Form より低く，自然な文が生成されにくいことが明らかとなった．これは，問いたい内容を表す疑問詞が文中から欠落してしまっていることが主な原因と考えられる．

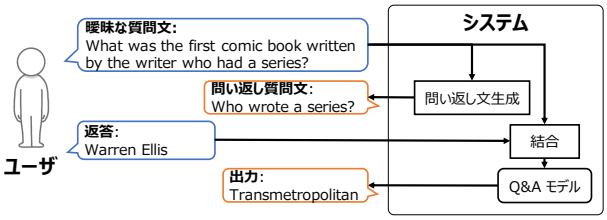

図 3　問い返し対話例

3　ユーザへの問い返し質問文生成

前節では文構造を用いることで，実際に既存の質問応答システムが回答できない形に質問文に曖昧性を付与できることがわかった．このような曖昧な質問文が持つ曖昧性を解消するには，何らかの方法で欠落した情報を補完する必要がある．こうした曖昧性解消を試みるため，大塚ら [4] はユーザの真の意図と推定される改訂質問をユーザに提示するモデルを提案した．このモデルでは，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，所望の回答を得ることができる．しかしこうした手法ではあくまでシステムが改訂質問を生成するため，ユーザから得られる情報の量に限界がある．

これに対して本研究では，ユーザへ問い返し質問を生成することで，対話的に不足情報を補うことができるシステムの構築を目指す．具体的には，用言の依存関係に基づいた問い返し箇所推定と，格フレーム辞書を基にした適切な疑問詞の決定・問い返し質問文生成を行う．

3.1　格フレームを用いた問い返し質問文生成

格フレームは，用言とそれに係る名詞の役割を用言の各用法ごとに整理したものである [12][13]．その中でも河原ら [13] は，大規模汎用英語コーパスと Web から収集した文から格フレーム辞書を教師なしで構築する手法を提案した．本研究では，この研究で構築された英語格フレームを用いる．古川ら [5] は，格フレーム中の格要素，格の種類と疑問詞対応規則を定義して，問い返し文候補を生成する手法を提案した．しかし，この手法を用いて問い返し質問を行うためには，問い返しの対象となる格を推定する必要がある．そこで本研究では，問い返しの対象となる格を適切に推定するために，文中の動詞との共起情報を用いる．

問い返し質問文生成の流れを図 2 に示す．まずはじめに，入力された曖昧な質問文に対して構文解析を行い，文中に出現する用言と用言の係る依存関係

元の質問文	What major truck road is located in Backford Cross?
曖昧な質問文	What major truck road is located?
問い返し文	where was the road located?
問い返しへの回答	Backford Cross

表 3　有効な問い返し例

	EM	F1
結合後の質問文	87.80	89.87

表 4　タスク全体の回答精度

タグ (格) を抽出する．用言として判別する品詞タグは，VB・VBD・VBG・VBN・VBP・VBZ の 6 種類の動詞タグとした．解析には Stanford Parser を用いた．次に，抽出された格と用言のペアおよび格フレーム辞書を参照しながら，文中に出現しない格を辞書中の頻度順に抽出する．用言の選択には，抽出された中でもっとも文末に出現するものを優先した．また，質問文によく出現する格の頻度情報を反映させるため，質問応答の学習データ中に出現する用言（動詞）に係る格役割の頻度情報を用いる．学習データには HotpotQA 学習用データを用いた．格フレームから抽出された格の中から，学習データ中でもっとも頻度が高い格役割を問い返す格として決定する[2)]．これは，質問文に頻出の格を問い返すことが質問応答の情報補完に有効と考えられるためである．ここで再度格フレーム中の抽出された格役割を参照し，その中で最も出現頻度の高い疑問詞を問い返す疑問詞として決定する．最後に，選択された用言と疑問詞を用いて問い返し文を生成する．例えば，図 2 の例では曖昧性を付与したユーザ質問文のうち最後尾の用言が “had” であり，格フレームから推定された必要な格役割のうち，質問文学習データ中で最も頻度が高いものが “advmo” であった．そこでこの格で最も頻度の高い “when” を用いて，“when did the writer have?” という質問を生成している．

4　問い返し質問応答による評価

ここまでに質問文に対する曖昧性付与と，この曖昧性を解消するための問い返し質問文生成手法を説明した．本節ではこの問い返しによって，実際に対話的な曖昧性解消が可能なのかについて評価を行う．まず問い返しタスクの設定を説明し，精度評価結果について述べる．

2)　データセット中の用言（動詞）とそれに係る格の頻度を付録に示す．

元の質問文	Which author dedicated a 1985 romance novel to the author who did in 2009 and wrote under the pen name Gwyneth Moore?
曖昧な質問文	Which author dedicated a 1985 romance novel to the author who?
問い返し文	when did the author dedicate?
問い返しへの回答	2009
元の質問文を入力した場合の出力	Eva Ibbotson
結合後の質問文を入力した場合の出力	Patricia Veryan

表 5 問い返しによって正解できなかった例

4.1 タスク設定

タスクの概要を図 3 に示す．本タスクではまず，2 節の変換手法を用いて生成した曖昧な質問文をシステムの入力として与える．システムは受け取った質問文を基に問い返し質問文を生成し，ユーザの返答から追加情報を得る．こうして得られた追加情報と入力された曖昧な質問文とを，既存の質問応答モデルの入力とした場合の回答をシステムの出力とする．システムに用いる質問応答モデルは，2 節で用いたものと同様に曖昧性を含まない完全な質問文のみから学習した．問い返し質問文に対するユーザの返答は，元の質問文を知っている場合に答えることができる適切な返答を人手で用意した．曖昧な質問文とユーザからの返答を結合しモデルに入力した場合の出力回答が，曖昧な質問文のみを入力とした場合から程度改善されるかについて評価する．

4.2 質問応答モデルの精度評価

問い返し質問による質問応答の改善度合いを測るため，問い返す対象の曖昧な質問文を以下のような制約で限定して実験を実施した．この条件に該当するものは，HotpotQA 開発用データセット中の質問文約 7,400 文に対して 312 文存在した．

1. 質問応答モデルの回答が変換による曖昧性付与によって正解→不正解となる質問文
2. 変換による欠落箇所の長さが最小である質問文
3. 疑問詞から始まる質問文 (Normal Form)

上記の制約によって抽出された対象となる質問文 312 文に対してそれぞれ問い返し質問文生成を行ったところ，欠落箇所への問い返し質問が生成されている例が 123 文存在した（39.4%）．有効な問い返しと回答の例を表 3 に示す．問い返し文として有効な 123 文を用いた場合のタスク全体での回答精度は表 4 の通りである．表 4 中のスコアは，曖昧な質問文と問い返しによって得られた回答とを結合し，質問応答モデルに入力した場合の出力の正解率 EM と 調和平均 F1 を表す．

これらの結果から，曖昧な質問文のみでは回答できない場合に，問い返しで得た追加情報を用いて質問応答の精度を向上することができることを確認できた．しかし，変換前はモデルが回答できるものだけを対象とした実験であっても，単純な結合だけでは完全な回答精度復元には至らなかった．これは，得られた追加情報だけではモデルが正確な回答を出力できないようなものが含まれていたか，あるいは追加で得られた情報の結合方法などに課題があったと考えられる．有効な問い返しによって正解を導き出せなかった例を表 5 に示す．表 5 では変換によって関係代名詞 “who” 以降の従属節の内容が大きく欠落しており，ユーザから追加で得られた情報だけでは答えを正しく出力することができなかったと考えられる．

5 まとめと今後の課題

本研究では，文構造に着目した質問文変換により情報を欠落させた曖昧な質問文を疑似生成し，これに対して適切な問い返し質問文を生成することで，対話的に曖昧性を解消できる質問応答システムを提案した．この結果，問い返しによって必要な情報を追加情報を得て質問応答精度を向上することができたものの，この情報の復元は限定的な範囲となった．今後，変換によって生成した曖昧な質問文の品質向上や，より適切な用言や疑問詞を選択するための手法の提案などに取り組む必要がある．

謝辞

英語版格フレームを提供して頂いた早稲田大学の河原大輔先生と京都大学の黒橋禎夫先生に感謝いたします．

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, Austin, Texas, November 2016. Association for Computational Linguistics.

[2]Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[3]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.

[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, 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]Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, p. 177–186, New York, NY, USA, 2018. Association for Computing Machinery.

[7]Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). 2020.

[8]中野佑哉, 河野誠也, 吉野幸一郎, 須藤克仁, 中村哲. 文構造に基づく質問文への曖昧性付与と質問生成. 人工知能学会 言語・音声理解と対話処理研究会（SLUD）第 90 回研究会, 2020.

[9]Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 455–465, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[10]Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[11]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.

[12]Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, p. 176–183, USA, 2006. Association for Computational Linguistics.

[13]Daisuke Kawahara, Daniel Peterson, Octavian Popescu, and Martha Palmer. Inducing example-based semantic frames from a massive amount of verb uses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 58–67, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

付録. A　学習データから集計された格と頻度

nmod:	81442	xcomp:	4521	acl:relcl:	444	neg:	83
nsubj:	60702	ccomp:	4461	acl:	285	mwe:	62
dobj:	49679	compound:	1740	cc:preconj:	282	appos:	37
nsubjpass:	23910	cop:	1554	csubjpass:	218	nmod:npmod:	27
advmod:	17991	case:	1529	nmod:poss:	177	discourse:	6
dep:	6817	compound:prt:	1344	nummod:	175		
conj:	5335	nmod:tmod:	1132	csubj:	143		
cc:	5152	amod:	951	expl:	108		
advcl:	4943	parataxis:	452	iobj:	100		

表 A.1　学習データ中に出現する用言の格と頻度