5P-03

博物館展示のための自然言語処理による質問文生成手法

松田意仁* 赤嶺有平† 根路銘もえ子‡

琉球大学理工学研究科情報工学専攻* 琉球大学工学部† 沖縄国際大学経済学部‡

1 はじめに

博物館では来館者が自発的に展示物のことを考えるよう促し，学びを促進させるマインズ・オンと呼ばれる展示手法があり，これは展示物に関するクイズを利用した手法などで効果が確認されている[1]．しかし，この効果はクイズに出題された展示物に限られると考えられ，展示物ごとの学びに偏りが生じる．理想的にはすべての展示物に対してクイズを用意することが望ましいが，現状では人手により問題文を作成する必要があるためコスト等の要因により現実的ではない．そこで本研究では人手ではなくニューラルネットワークによる既存の質問生成モデルを利用しクイズ作成を行う．

一般的にニューラルネットワークの学習には大量のデータが求められるが，日本語における質問生成の学習データセットは入手が困難である．一方で，英語においては The Stanford Question Answering Dataset(SQuAd)[2]のようなWikipediaの記事に対する，クラウドソーシングによる10万を超える質疑応答データセットが存在する．これは数文の文章，質問文，回答からなるデータセットであり，文章中に回答が含まれているという性質は，博物館展示の説明文に来館者を誘導させることと相性が良い．

本研究ではSQuAdを学習データとして利用したLopezらの質問生成モデル[3]を活用することで，クイズの導入における博物館側の負担を減らす手法の実用化を目指す．

2 提案手法

クイズを作成する際にはテーマに関わる内容が出題されることが望ましいが，既存の学習データセットはその点が考慮されていない．そこで質問生成モデルへ入力する文章にはテーマと関係する部分が質問されやすくなるように前処理を行なう．係受け関係がある文節は内容に関連性が強いと考えられる．そこでテーマと関係するキーワードを含んだ文節との係受け関係が離れている文節を元の文章から取り除くことでテーマと関連性の強い文に置き換えられると考えられる．質問文を生成する過程において構文解析にはKNP，翻訳にはgoogle翻訳，質問生成にはLopezらの質問生成モデルを利用し，元の文章に対して以下の手順で前処理を行い，クイズに用いる質問文と回答を生成する．

1. キーワードを含んだ文のみを抜き出す．キーワードはテーマ名とそれに強く関連する語句である．
2. KNPを用いて文節間の係受け関係を調べる．
3. 以下の文節を連結しあらたな文を生成する．
   - キーワードを含む文節(主文節)
   - 主文節が係っている文節(親文節)
   - 主文節から係っている文節を辿り文末までに出現する全ての文節
   - 主文節の親文節に係る文節(兄弟文節)
   - 兄弟文節に係っている文節

例:「当時は現代のエイサーと形式が異なり、門付歌と念仏歌だけで踊っていた。」→「現代のエイサーと形式が異なり、踊っていた。」

4. 前処理後の文章にgoogle翻訳で英語に直す．
5. 質問生成モデルにより質問文と回答を生成する．
6. 生成された質問文と回答にgoogle翻訳を用いて日本語に直す．

3 実験

Wikipediaから沖縄と関わり深い「エイサー」,「首里城」の記事からそれぞれ16文，15文の文章を抜き出し，抜き出した文章をもとに質問生成を行った．表1，2，3は前処理を行った上でのクイズとして適している例，回答が誤っている例，テーマに対するクイズとして成り立たない例である．また，表4は前処理を行わなかった場合の例である．

「エイサー」に対する質問生成について．前処理を行なった場合と行わなかった場合のそれぞれの生成された質問数，適切と判断した質問数，不適切と判断した質問数は表5のようになった．

表6は「首里城」に対して「首里城，尚」をキーワード

Method of Question Generation by Natural Language Processing in Museum Exhibition

*Okuto Matsuda, Ryukyu University, Graduate School of Engineeringa and Science, Information Engineering

†Yuhei Akamine, University of the Ryukyus, Faculty of Engineering

‡ Moeko Nerome, Okinawa International University, Department of Economics

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

として質問生成を行なった結果である. 尚は王家の名前であり首里城の歴史的背景と強く関連する語句である.

質問文	回答
エイサーが一般の人々の間で人気になったのはいつですか？	明治時代
明治時代以降、エイサーはどこでスタイルを変えましたか？	本島の中央部
エイサーを初めて本土に紹介したのは誰ですか？	赤野青年会

表 1:クイズに利用可能な例
「エイサー」, 前処理あり

質問文	回答
うるま市の伝統的なエイサーの創設者は誰ですか？	やけいめいエイサー
エイサーの発展に大きな影響を与えたのはどのイベントですか？	エイサ

表 2:回答が誤っているためクイズに適さない例
「エイサー」, 前処理あり

質問文	回答
エイサー文化とともに歩んだ街は？	沖縄

表 3:難易度が低くクイズとして成り立たない例
「エイサー」, 前処理あり

質問文	回答
やけめい青年会、ひらしきや青年会、赤野青年会の伝統はどのくらいありますか？	100年以上
大中は首里にどれくらい滞在しましたか？	3年
沖縄で最初の競争は何でしたか？	ランキングの競争

表 4:「エイサー」, 前処理なしの例

	質問数	適切	不適切
前処理あり	10	6	4
前処理なし	20	6	14

表 5:「エイサー」に対する, 前処理ありとなしのそれぞれの適切, 不適切な質問数

	質問数	適切	不適切
前処理あり	11	3	8
前処理なし	16	5	11

表 6:「首里城」に対する, 前処理ありとなしのそれぞれの適切, 不適切な質問数

4 考察

表5, 6から前処理を行なった場合の方が, 生成される質問数が少なくなっていることがわかる. これは前処理によっていくつかの文節が除外され, 元の文章よりも短くなった内容を質問生成モデルに入力していることが原因だと考えられる. しかしこの前処理によって生成された質問にはテーマ名が含まれやすくなっており, 表4に示すような事前知識が無ければ質問の意図が伝わりづらいものは見られなかった.

また, 前処理を行なった際に生成された「エイサー」についてのクイズでは質問生成モデルの読解能力における問題から, 誰ですか？という問いに対して人以外を答えるような結果も見られた. しかし表5に示すように, 不適切だと判断した質問の割合が減ったことから, テーマと関わりが薄い質問の除外する効果があったと考えられる.

「首里城」に対する結果も同様に, 不適切な質問の数が減っていることが表6からわかる. しかしそれと同時に適切な質問数も減少しており, 実験で行なった前処理のみではクイズとして適切な質問を損なう可能性があるといえる.

これらのことから, 文節間の係受け関係のみを考慮した前処理は, 不適切な質問を減らすという一定の効果はあるものの改善の余地が残っていると考えられる.

5 おわりに

本研究では既存のニューラルネットワークによるクイズのための質問生成を行なった. クイズの質を高めるため, キーワードを含む文節と係受け関係が離れている文節を取り除く前処理を行うことで, テーマから離れた内容の質問の生成数が減少した. しかし前処理を行うことで良い質問を損なう結果も見られ, 文節を取り除く行為は慎重に行う必要がある. 今後は質の高いクイズを損なわないような前処理が課題となる.

謝辞

本研究は JSPS 科研費 JP19K01142, JP19K01145 の助成によるものである.

参考文献

[1] 神保 英 安斉 賢三 齋藤 佑樹 中村 雅子, 博物館での学習における拡張現実(AR)技術の可能性. 東京都市大学横浜キャンパス情報メディアジャーナル = Journal of information studies (15), 16-22, 2014-04

[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing
, 2383-2392, 2016

[3] Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christrian Blaise Crus, Chariveth Cheng, Transformer-based End-to-End Question Generation, arXiv:2005.01107, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2020

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.