対話型質問応答の省略補完

笹沢裕一　高瀬翔　岡崎直観
東京工業大学
{yuichi.sasazawa at nlp.c, sho.takase at nlp.c, okazaki at c}.titech.ac.jp

1 はじめに

質問生成とは、与えられたテキストに対して質問文を生成するタスクである。例えば、「言語処理学会は、わが国の言語処理の研究成果発表の場として、また国際的な研究交流の場として、1994 年 4 月 1 日に設立されました。[*1]」というテキストに対して、「言語処理学会は何年に始まったのか？」や「言語処理学会は今年で何周年を迎えるのか？」といった質問文を生成したい。典型的な応用例としては、国語の読解問題を自動生成するといった教育目的 [4] や、対話ボットが話しかけるきっかけを作り出すこと [6] などが挙げられている。

質問生成の既存研究は、テキストだけから質問文を生成する設定 [2]、テキストと答えから質問文を生成する設定 [8, 9, 11] などがある。これに対し、本研究では省略された質問文を入力に与え、完全な質問を生成するタスクに取り組む。例えば、先ほどの段落で挙げたテキストに対して、「何年？」という省略された質問文（疑問詞）を入力に加え、「言語処理学会は何年に始まったのか？」という省略が補完された質問文を生成する。

このタスク設定は、最近流行の兆しを見せている対話や文脈を考慮した質問応答 [1, 7] に応用できる可能性がある。通常の質問応答タスクはそれぞれの質問文が独立であり、与えられた質問文だけで解答可能であった。これに対し、対話や文脈を考慮した質問応答では、過去の質問文やその解答の履歴を活用することが求められる。例えば、Conversational Question Answering Challenge (CoQA) [7] のデータセットでは、“Did she plan to have any visitors?” という質問に ‘Yes’ で答えた後、次の質問文として “How many” が与えられ、お客さんの数を答えなければならない。

このような対話を考慮した質問応答に向けて、現在の質問文と前の質問文とその答えを連結したものを質問応答モデルに入力する手法 [7, 10]、前の質問文を答える際に生成された中間表現を現在の質問の応答モデルに引き継ぐ手法 [5] などが提案されている。本研究では、省略された質問文から完全な質問文に変換するアプローチが、対話を考慮した質問応答にどのくらい貢献するのか、実験する。

[*1] http://www.anlp.jp/

2 関連研究

これまでに、ニューラルネットワークに基づく質問生成の研究がいくつか発表されている。Du ら [2] は、テキストのみを入力として質問文を生成する研究に取り組んだ。彼らの手法は、アテンション機構を利用した RNN 型のエンコーダ・デコーダモデルであり、テキストを入力として質問を生成する。この手法は、それまでのルールベースの手法よりも優れた性能を示したが、解答をモデルに入力していないため、生成される質問文の制御ができないという欠点がある。

これに対し、モデルに質問の解答を与える研究が発表されている。Kim ら [11] は、テキスト中の解答の位置もモデルの入力に与え、質問の解答の位置を知らせている。これに加えて、テキスト中の品詞情報や固有表現タグを活用している。質問文のデコードでは、コピー機構 [3] を利用している。Zhou ら [8] は、テキストと解答の関連性を質問生成モデルに組み込むため、エンコーダーに多視点マッチングアルゴリズムを利用することを提案している。この手法でも、デコーダーにはコピー機構を採用している。

Kim らと Zhou らの研究では、デコーダにコピー機構を採用したため、テキスト中の解答が質問文の中に含まれやすくなるという弊害が発生した。そこで、Song ら [9] は、解答の情報をテキストから分離し、質問の回答が質問文の中に含まれないようにする方法を提案した。

これらの従来研究に対し、本研究では質問文の疑問詞を質問生成の追加情報としてモデルに与える。これによ

り、目的の疑問詞を含むような質問文を生成できるような制御を行う。

3 提案手法

本研究では、テキストと省略された質問文を入力し、完全な質問文に復元することを目指す。手法の概要を図1に示す。

テキストと質問が含まれるデータとして、質問応答タスクで有名な SQuAD(v1.1) を利用する。SQuAD のデータには段落文と質問文、質問に対する解答が含まれる。このデータセットの解答は全て、段落文中の部分文字列となっている。本研究では段落文をそのまま質問生成の種となるテキストとする。

次に、質問文から疑問詞句を取り出したものを省略された質問とする。具体的には Stanford Core NLP[*2]を用い、SQuAD の質問文の係り受け解析を行い、疑問詞（品詞タグが “WDT”, “WP”, “WP$”, “WRB” のいずれかであるもの）と、その疑問詞から係り受け木の根の単語までのパスに現れる単語全てを疑問詞句と見なす。この処理により、例えば “How many?” や “What countries?” のような省略された質問を得ることができる。ただし、質問文が 5 文字以下のものや、命令文、肯定文にクエスチョンマークを付けて Yes/No 形式の質問文にしたものなど、疑問詞が存在しない質問文は取り除く。以上の手順により、SQuAD のデータセットをテキスト、質問、省略された質問、解答の組に変換し、87,484 件の学習データ、10,559 件の評価データを得た。

省略された質問から完全な質問に復元するタスクを、翻訳問題の一種と捉え、テキストと省略された質問を入力とし、完全な質問を出力としたエンコーダ・デコーダモデルを学習する。モデルの入力には、テキストと省略された質問を区切り文字で連結した単語列を与える。モデルとして、アテンションを使用した RNN 型のエンコーダ・デコーダモデルを採用する。

4 実験

4.1 実験設定

実験に用いる手法は、全て Du ら [2] の手法に基づいている。この手法は torch ベースの OpenNMT に基づいて実装されており[*3]、アテンションを使用した RNN 型 Encoder-Decoder モデルである。単語ベクトルとして glove.840B.300d[*4]を使用し、質問生成モデルの学習中も単語ベクトルを更新した。2 層の LSTM を採用し、隠れベクトルの次元数は 600 とした。モデルパラメータの最適化には確率的勾配降下法（SGD）を使用し、ドロップアウト率を 0.3 に設定した。

モデルに入力する情報は以下の通りである。

テキスト テキストのみを入力する。

テキスト + 解答 テキストと解答を連結したものを入力する。

テキスト + 疑問詞句 テキストと質問文の疑問詞句を連結したものを入力する。

テキスト + 疑問詞句 + 解答 テキストと質問文の疑問詞句と解答を連結したものを入力する。

4.2 実験結果

（元々の）完全な質問文を正解と見なし、モデルにより生成された質問文の質を BLEU-4 で測定したものを表2に掲載した。テキストに加えて疑問詞のみを入力として加えた場合が、最もよい BLEU スコアを示した。また、解答を入力に追加した場合は、疑問詞句を追加で与えたかどうかに関わらず精度が少し低下した。先行研究では、解答を入力する際にモデルを工夫しているが、本研究ではテキストと連結しているだけであるため、解答の情報を活用できなかった可能性がある。ただ、表1から分かるように、テキストに疑問詞句を加えた場合は、BLEU スコアの向上が確認できた。

次に、提案手法（テキスト + 疑問詞句）がどのような質問文を生成したのか、生成された質問文の中から無作為に 100 件を抽出し、その質を以下の基準で分別した。

完全な質問文 正解の質問文と完全に同じ質問文

ほぼ完全な質問文 正解の質問文と比較したとき、単語が似た意味の別の単語と変わっていたり、正解の質問文に含まれていた前置詞句が失われたり、余計な前置詞句が追加されているが、正解とほとんど同じ意味の質問文

同じ解答に至る質問文 正解の質問文と質問の意味が変わっているが、同じ解答にたどり着く質問文

違う解答に至る質問文 正解の質問文と質問の意味が変わっており、テキスト中に含まれる内容を間違った解答としてしまうような質問文

[*2] https://stanfordnlp.github.io/CoreNLP/

[*3] https://github.com/xinyadu/nqg

[*4] https://nlp.stanford.edu/projects/glove/

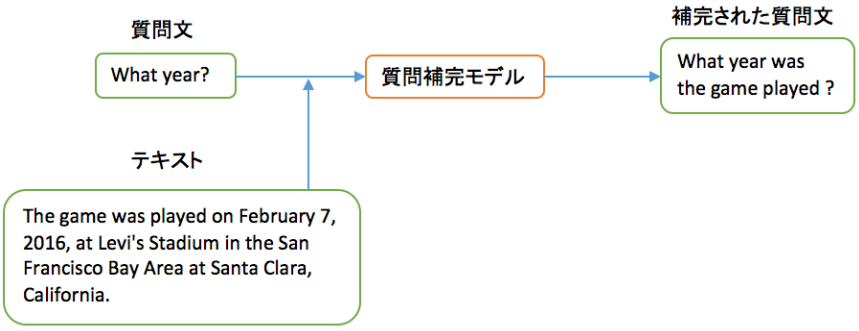

図1: 提案手法の概要

1単語の差で解答にたどり着けない質問文　以上のどれにも当てはまらないが、正解の質問文と比較したとき、1つの単語が別の単語に変わっているため、質問の意味が変わってしまったもの

意味は通じているが、解答にたどり着けない質問文　以上のどれにも当てはまらないが、正解の質問文と意味が変わっており、その解答もテキスト中に現れないもの

意味が通じない質問文　明らかな文法のミスや単語の並びの不自然さの為に、質問文として意味がとれないもの

それぞれの基準に該当する質問文の例と無作為に抽出した100件における分布を表1に示す。「完全な質問文」から「違う解答に至る質問文」までを正しい質問文が生成できたとみなすと、39%が正しい質問文であったことになる。生成例の中で最も多かったのは「意味は通じているが、解答にたどり着けない質問文」であった。「意味は通じているが、解答にたどり着けない質問文」の原因は、同じ単語が複数回生成されたため、文法や意味的におかしな文になっている場合や、「意味は通じているが、解答にたどり着けない質問文」と同様、疑問詞句とそれ以外の部分の不整合から意味が通じない文になっている場合などがあった。

4.3　補完された質問文を用いた質問応答

本研究で開発した質問補完モデルが、対話を考慮した質問応答に適用できるのか検証した。ここでは対話型のデータセットであるCoQAを用いて質問応答のタスクを解く。質問応答のモデルはCoQAのベースラインであるDrQA+PGNet[*5]を使用する。DrQAは3層のLSTMを使用し、隠れベクトルは300次元である。ドロップアウト率はLSTMに対して0.4、単語ベクトルに対して0.5である。最適化にはAdamを用いた。PGNetは2層のLSTMで構成され、隠れ状態ベクトルは500次元である。ドロップアウト率は全ての層に対して0.3とした。最適化にはSGDを利用した。

CoQAの学習データの中で、疑問詞句のみで構成される質問文に対して提案手法で補完を行い、DrQA+PGNetの学習の学習データとした。CoQAのデータの中で補完を行わなかった残りの質問文も合わせて学習データとした。評価では、省略された質問文のみを対象とし、提案手法で質問文の補完を適用した後、DrQA+PGNetの質問応答モデルで解答を得た。

DrQA+PGNetの質問応答モデルに入力として与える情報として、以下の4種類を比較した。

0履歴　現在の質問文のみを入力する。

0履歴＋ルール補完　現在の質問文の疑問詞と、1つ前の質問文の疑問詞を取り除いたものを繋げ、省略された質問を補完するベースライン手法である。以下に例を示す。

(前の質問文):Where did he live?
(質問文):In what state?
(生成された質問文):In what state did he live?

0履歴＋質問補完　質問補完のモデルで省略された質問文を補完し、入力する。このとき、質問補完のモデルに与えるテキストは、質問の答えを含む文とした。

1履歴　現在の質問文と1つ前の質問文とその解答を連結させたものを入力する。

表3に質問応答の解答のF値を示した。0履歴、0履歴＋ルール補完の場合に比べて0履歴＋質問補完は精度が向上し、省略の補完により解答を導出しやすい質問文が生成できていることが分かる。ただし、この設定では

[*5] https://github.com/stanfordnlp/coqa-baselines

表1: 補完された質問の例

完全な質問文:3% テキスト:The country is known chiefly for its dominance in middle-distance and long-distance athletics, having consistently produced ... 質問文:What for? 解答:dominance in middle-distance and long-distance athletics 正解:What is the country known for? 予測:What is the country known for?
ほぼ完全な質問文:19% テキスト:In September 2007, Sky launched a new TV advertising campaign targeting Sky+ at women. 質問文:When? 解答:September 2007 正解:When did Sky launch a TV advertising campaign target towards women ? 予測:When did Sky launch a new TV advertising campaign?
同じ解答に至る質問文:7% テキスト:Early ties were already visible in the “Apologie” of William the Silent , condemning the Spanish Inquisition, which was written by his court minister,the Huguenot Pierre L’Oyseleur, lord of Villiers. 質問文:What Dutch document? 解答:“Apologie” 正解:What dutch document condemned the spanish inquisition? 予測:What dutch document was written by the court minister?
違う解答に至る質問文:10% テキスト:[citation needed] his lab was torn down in 1904, and its contents were sold two years later to satisfy a debt. 質問文:What? 解答:sold 正解:What happened to the things inside the lab after it was torn down? 予測:What happened to bell ’s lab?
1 単語の差で解答にたどり着けない質問文:11% テキスト:Peridinin is not found in any other group of chloroplasts. 質問文:What? 解答:Peridinin 正解:What is only found in peridinin-type chloroplasts? 予測:What is not found in any other group of eukaryotes?
意味は通じているが、解答にたどり着けない質問文:37% テキスト:As well as the proposer, other members normally contribute to the debate. 質問文:Who? 解答:other members 正解:Who contributes to Members Business in addition to the proposer ? 予測:Who is the main proponent of the debate?
意味が通じない質問文:13% テキスト:The area is also known for its early twentieth century homes, many of which have been restored in recent decades. 質問文:How recently? 解答:recent decades 正解:How recently has the homes in Fresno been restored? 予測:How recently is the area of the area?

質問文を補完するときに解答を含む文を入力として与えているため、対話における質問応答の問題設定から、少し逸脱していることに注意が必要である。また、1 履歴と比べると精度は低かったため、質問補完の性能の向上、および補完する際のテキストの与え方に改善の余地があると考えられる。

表2: 質問補完の精度（BLEU スコア）

実験手法	BLEU
テキスト	12.0
テキスト + 解答	11.2
テキスト + 疑問詞句	**24.9**
テキスト + 疑問詞句 + 解答	24.0

表3: 質問応答タスクでの正解率（F 値）

入力	F 値
0 履歴	44.8
0 履歴 + ルール補完	56.6
1 履歴	67.4
0 履歴 + 質問補完	62.4

5 おわりに

本研究では、省略された質問文から完全な質問文を生成する手法を提案し、省略の補完が約 3 割くらいの成功率であることを確認した。今後の課題としては、質問補完の精度そのものを向上させることと、対話や文脈を考慮した質問応答での効果を検証することである。

6 謝辞

本研究は JSPS 科研費 JP18K18119 の助成を受けたものである。

参考文献

[1] Eunsol Choi et al. “QuAC: Question Answering in Context”. In: *EMNLP*. 2018, pp. 2174–2184.
[2] Xinya Du, Junru Shao, and Claire Cardie. “Learning to Ask: Neural Question Generation for Reading Comprehension”. In: *ACL*. 2017, pp. 1342–1352.
[3] Caglar Gulcehre et al. “Pointing the Unknown Words”. In: *ACL*. 2016, pp. 140–149.
[4] Michael Heilman and Noah A. Smith. “Good Question! Statistical Ranking for Question Generation”. In: *HLT-NAACL*. 2010, pp. 609–617.
[5] Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. “FlowQA: Grasping Flow in History for Conversational Machine Comprehension”. In: *arxiv:1810.06683* (2018).
[6] Nasrin Mostafazadeh et al. “Generating Natural Questions About an Image”. In: *ACL*. 2016, pp. 1802–1813.
[7] Siva Reddy, Danqi Chen, and Christopher D. Manning. “CoQA: A Conversational Question Answering Challenge”. In: *arXiv:1808.07042* (2018).
[8] Linfeng Song et al. “Leveraging Context Information for Natural Question Generation”. In: *NAACL*. 2018, pp. 569–574.
[9] Kim Yanghoon et al. “Improving Neural Question Generation using Answer Separation”. In: *arxiv:1809.02393*. 2018.
[10] Mark Yatskar. “A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC”. In: *arxiv:1809.10735* (2018).
[11] Qingyu Zhou et al. “Neural Question Generation from Text: A Preliminary Study”. In: *NLPCC*. 2017, pp. 662–671.