答えを用いない対話型質問の生成

中西 真央　　小林 哲則　　林 良彦
早稲田大学 理工学術院
nmao@asagi.waseda.ac.jp

1 はじめに

対話型質問の生成とは，質問応答対話における質問を生成するタスクであり，自然言語処理研究の新しい分野の 1 つである．従来の質問生成研究は一問一答形式における質問を生成することを目的にしていた．また，従来の質問生成研究の多くが答えを用いる質問生成手法であるが，これは特に対話システムなどの応用領域への適用を制限する可能性がある．そこで本研究は，対応する回答を用いない対話型質問生成のフレームワークを初めて提案する．

CoQA データセット [7] を使用した実験の結果は，質問焦点と質問のパターンが正しく推定されると，生成される質問の質が大幅に向上することを示した．さらに，質問の焦点は妥当な精度で推定可能であり，質問の焦点の推定が質問の質の向上に貢献できることが示された．これらの結果は本研究の方向性が有望である見込み示している．一方で，質問パターンの特定は非常に困難な問題であり，生成する質問の質の向上に貢献するためには大幅な改良が必要である事を明らかにした．

2 背景

質問生成は，対話システムや質問応答システム（QA システム）など，様々な応用分野の発展に伴い，自然言語処理分野で大きな注目を集めている．質問生成の主に用いられる手法は，ニューラルネットワークの発展をきっかけに，テンプレートを使用する手法からニューラルネットワークを用いた end-to-end の手法に移行した．

近年，質問応答システムの研究対象は，より自然な状況下であると言える対話型質問応答に拡大され，CoQA[7], QuAC[2] などのデータセットが対話型質問応答のためのデータセットとして開発された．

このような対話型質問応答システムの発展を受け，昨年，初めて対話における質問生成（対話型質問生成）

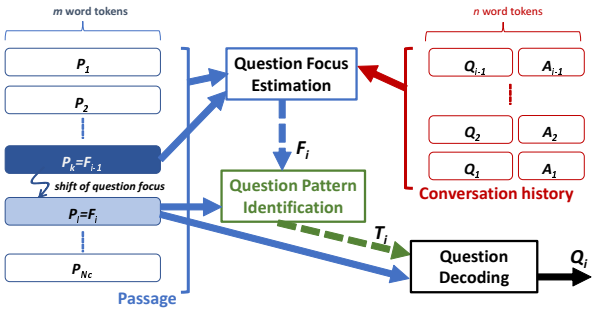

図 1: 提案手法フレームワークの概要図

に取り組む手法が提案された [3]．この提案手法は，質問間の相互参照と対話の流れを考慮する手法である．

しかし，この提案は回答を使用する手法であり，特に対話システムなどの応用領域を，何らかの形で制限する可能性がある．したがって，本研究では対話型質問応答において答えを用いない質問生成手法を初めて提案する．

3 対話型質問生成

3.1 概要

図 1 に提案するフレームワークの概要を示す．提案フレームワークは 3 つの構成要素 (1) 質問焦点推定，(2)　質問パターン同定，(3) 質問表層生成から成る．

ここでは以下の項目を仮定している．

- 生成する質問に対応する答えを与えなくても，質問の焦点を与えることで，現在の対話の文脈に沿った質問を生成可能である．ここで，質問の焦点は，対話履歴を利用することでテキストの特定の領域として推定可能であるとする．
- 質問パターンが特定されていれば，質問の質はさらに向上する．質問のタイプを表現する質問パターンは，対話履歴と推定された質問の焦点を使用することで推定可能であるとする．

3.2 提案モデル

以下の説明では，現在の QA ターン t が i であるとする．質問生成システム全体への入力は，対象のテキスト P と，現時刻の対話履歴 H_i である．

テキスト P は、一連の N_c チャンク $(P_1, \cdots, P_{N_C})$ に分割される．i 番目の QA ターン対話履歴 H_i は， $H_i = ((Q_1, A_1), \cdots, (Q_{i-1}, A_{i-1}))$ として定義され，セパレータで区切られた質問と答えの単語列として以下のように実装する．ただし，以後 $H_i = (w_1^{H_i} \cdots w_n^{H_i})$ と略して使用する．$H_i = (\cdots, w_{q_1}^t \cdots w_{q_{|Q|}}^t, \langle sep \rangle, w_{a_1}^t, \cdots, w_{a_{|A|}}^t, \cdots)$.

i 番目の QA ターンの質問焦点 F_i は 1 チャンクとして推定される．

生成される質問に対応する質問パターン T_i は、事前に定義した質問パターンセット T_Q から選択される．

3.2.1 質問焦点推定

質問焦点推定は文章と対話履歴から質問焦点を推定する．質問焦点とは，対話の進行に応じた文章中の注目部分である．

質問焦点推定ネットワークは，embedding layer, contextual layers, attention layer, modeling layer, output layer から構成される．

embedding layer は，テキストに含まれるチャンク P_c をベクトル列に変換する．contextual layer は，2 つの Bi-GRU を使用して実装され，1 つはテキストチャンク用，もう 1 つは対話履歴用である．attention layer では，対話履歴から テキストチャンク に対する attention weights を獲得する． この層の出力 $\tilde{H^{c_i}} (c = 1, \cdots, N_C)$ は対話履歴に関連する単語が強調された文脈表現である．ここで，W_e および W_e は学習パラメータである．

$$e_{t,j}^f = tanh(\boldsymbol{W}_e^f [h_t^{c_i}; h_j^{H_i}]) \tag{1}$$

$$\alpha_{t,j}^f = \frac{exp(e_{t,j}^f)}{\sum_{k=1}^n exp(e_{t,k}^f)} \tag{2}$$

$$c_t^f = \sum_j \alpha_{t,j}^f h_j^{c_i} \tag{3}$$

$$\tilde{h_t^{c_i}} = tanh(\boldsymbol{W}_h^f [c_t^f; h_t^{c_i}]) \tag{4}$$

modeling layer は 1 層の Bi-GRU から構成され，対話履歴への注意機構を経由した 各チャンクの文脈表現間の相互作用を捕捉する．

output layer は 2 つの線形層で構成され、最も可能性の高いチャンク番号 y_{F_i} を推測する．推測されたチャンクは，以降の工程で現時刻の質問焦点 F_i として扱われる．

3.2.2 質問パターン同定

質問パターン同定 は推定した質問焦点と対話履歴から質問のパターンを推定する．この質問パターンとは，質問の内容を表す何らかの抽象表現である．5 つの質問パターンのいずれかを使用してこのタスクに取り組み，実験的に比較する．

質問パターン同定モデルは，embedding layer, contextual layers, attention layer, modeling layer, output layer から構成され，全体的な構造は質問焦点推定モデルとほとんど統一であるが，推定された質問焦点のみを考慮する点が異なる．具体的には，質問焦点は $[\boldsymbol{E}^{F_i}; \boldsymbol{f}_{F_i}^{NE}]$ と表現される．これは，元の質問焦点 $\boldsymbol{E}^{F_i}$ が 固有名詞タグ $\boldsymbol{f}_{F_i}^{NE} \in \mathbb{R}^{m \times 18}$ によって拡張された表現である．固有名詞として， spaCy [1]が推定する 18 種類を用いた．

attention layer と modeling layer によって得られた，質問焦点と対話履歴の文脈表現は output layer に入力され，最も可能性の高い質問パターンのインデックス $y_{T_i} \in \mathbb{R}^{N_P}$ が最終的に取得される．N_P は、事前定義された質問パターンの数を表す．

3.2.3 質問表層生成

最後の質問表層生成 では，推定した質問焦点と質問パターンから質問を構成する．質問デコードモデル として従来のアテンション付きエンコーダーデコーダーモデルを用いる．このモデルは推測された質問パターンを用いるか用いないかで異なる．つまり，質問パターンを使用しない場合，エンコーダへの入力は質問焦点 F_i のみである．一方質問焦点を使用する場合，エンコーダへの入力は，推測された質問焦点 F_i と，推測された質問パターン $T_i = (w_1^{T_i}, \cdots, w_l^{T_i})$ とを $\langle sep \rangle$ で区切った連結したものである．

4 実験

4.1 データセット

本研究では，Conversational Question Answering (CoQA) [7] と呼ばれる対話型質問応答 のデータセット

[1] https://spacy.io

を使用し実験を行った．CoQA は，質問側，回答側の 2 人組のクラウドワーカー (Amazon Mechanical Turk) 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り，テキストに関する 8k の対話から得られた，127k の質問応答を含む．回答はフリーテキストで提供されるが，回答の根拠として適当なテキスト領域には明示的に *Rationale* として注釈がつけられている．生成する質問に対応する真の質問焦点は，この *Rationale* と重複するテキストチャンクとして識別する．

4.2 質問パターン

本実験では，以下に示す 5 つの質問パターンのいずれかを使用してこのタスクに取り組み，質問パターンの推定精度および生成された質問の精度を比較する．

Question Word Pattern 10/200 (クラス数: 11/201) 質問に頻出する単語で分類した質問パターンクラス．使用するデータセットの学習データに含まれる質問を使用し，先頭から 1-gram，上位 10 個の頻出パターンを質問パターンセットと，先頭から 1,2,3-gram，上位 200 個のパターンの粒度の異なる 2 つのセットを作成し，これを質問パターンセットとして用いた．なお，質問パターン 学習時には一部のデータがサンプルに偏らないよう，サンプルの最大数がそれぞれ 5000, 30 となるように調節して学習を行った．

Li's Answer Class (クラス数: 6) Li and Rath [5] によって定義された答えの分類．質問に対する答えのクラスを，階層的に上位 6 クラスと下位 40 クラスに分類している．本実験では上位 6 クラス (ABBREY, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC) を質問パターンセットとして用い，以降 Li's Answer Class と呼ぶ．CoQA に対する Li's Answer Class ラベルは存在しないため，[5] のデータを用いて識別器を学習し，CoQA データセットに適用した．CoQA に対する識別精度は 70.6%であった．なお，質問パターン 学習時には各クラスのサンプルの最大数を 5000 として学習を行った．

Question Conceptual Class (クラス数: 7) Lehnert [1] によって定義された質問の内容に注目した概念的な分類．図 2 に具体的なクラスとその説明を示す．Krishna and Iyyer [4] が CoQA を含む 3 つの質問応答データセットに含まれる質問に対して Question Conceptual Class の分類を行った．分類は一部の質問を手動で分類したのち，分類した質問を用いて学習した識別器によって残りを自動的に分類する手法が取られ，CoQA に対する識別精度は 80%だった．

Conceptual class	Question asks for...	Sample templates
Specific concept completion	fill-in-the-blank information	Where / When / Who ...
General concept completion		
Verification	Yes-No answers	first word is VERB
Causal	the reason for occurrence of an event and the consequences of it	Why ..., What happened after / before ..., What was the cause / reason / purpose ...,
Instrumental	a procedure / mechanism	How question with VERB
Judgemental	a listener's opinion	Words like you, your present
Quantification	an amount	How many / long ...

図 2: Question Conceptural Class

表 1: 質問焦点のみの場合の BLEU スコア

入力テキスト	精度	B1	B2	B3	B4
文章全体	-	30.19	12.85	0.32	0.13
ランダム	20	33.83	16.08	0.59	0.13
推定した焦点	59.78	34.64	16.65	0.70	0.18
正しい焦点	100	34.19	16.30	0.71	0.21

4.3 実験設定

単語ベクトルとして，GloVe [6] ($d = 300$) を採用した．質問焦点推定時のテキスト分割数 N_c は，5 分割，10 分割の試行の結果，5 分割とした．

5 結果

5.1 質問焦点のみによる質問生成結果

質問焦点のみを使用した質問生成結果を表 1 に示す．B1, ... , B4 はそれぞれ BLEU1 スコア，..., BELU4 スコアを表す．また，2 列目の精度とは質問焦点に正しいチャンクが定めらている割合を表す．

表 1 は，質問焦点の推定精度が約 60%であること，文章全体を入力した場合より，推定した質問焦点を入力した場合の方がより良い精度であることを示している．これは，質問の焦点が対話履歴を利用することで推定可能であること，および対話の文脈に一貫した質問は，答えを与えずとも，質問の焦点を与えることで生成可能であるという仮定が正しいことを示す．

5.2 質問パターンも使用した質問生成結果

正しい質問焦点，および正しい質問パターンを入力した場合における質問生成生成精度を表 2 に，推定した質問焦点および推定した質問パターンを入力した場

表 2: 真値の質問焦点，質問パターンを使用した質問生成結果 BLEU スコア.

パターン	B1	B2	B3	B4
Question Word Pattern 200	56.22	38.84	18.69	7.10
Question Word Pattern 10	34.49	7.82	2.53	1.00
Li's Answer Class	25.33	2.73	0.736	0.301
Question Conceptual Class	24.03	3.13	0.750	0.318

表 3: 推定した質問焦点，質問パターンを使用した質問生成結果 BLEU スコア.

パターン	B1	B2	B3	B4
Question Word Pattern 200	32.36	16.06	0.372	0.040
Question Word Pattern 10	17.322	0.355	0.028	0.004
Li's Answer Class	24.863	2.354	0.505	0.178
Question Conceptual Class	23.107	2.736	0.624	0.250

合の質問生成結果を表 3 に示す．また，表 4 に質問パターン推定結果を示す．表 4 における 3, 4 列目の真値，推定とは，入力した質問焦点が正しいものか推定したものであるかを表している．

表 2 から，Question Word Pattern 200 を使用した場合，文章全体から質問を生成した場合に比べ，すべての BLEU スコアが勝っていることがわかる．また，4 つのパターンすべてにおいて，文章全体から質問を生成した場合に比べ，BELU3, BLEU4 では勝っていることがわかる．通常，質問生成の生成結果の精度は，より長期な連続を考慮した BLEU3，BLEU4 に注目して判断される．このことから，正しい質問パターンを入力することで，生成される質問の精度が向上すること，すなわち提案手法が有望であることが示唆されたと言える．

一方で，表 3 から，推定した質問パターンを入力することで質問生成の精度が大きく低下することがわかる．表 4 の結果も考慮すると，質問パターンの推定が非常に困難であり，提案手法に大きな悪影響を与えていることがわかる．また，質問パターン 4 種のうち，Question Word Pattern は質問生成に非常に有効であるが，対話履歴とテキストからでは推定が困難であるのに対し，他 2 つのパターンは，比較的推定は容易であるが質問生成良い効果を発揮しないことがわかる．これは，質問生成に有効な情報は具体的で粒度が細かい情報であること，一方で具体的で粒度の細かい指標は困難が難しくこの二つはトレードオフの関係であることを説明する結果と言える．結果的に，Question Conceptual Class が，提案手法における質問パターンとして最適であることが明らかになった．

表 4: 質問パターン推定結果

パターン	クラス数	真値 Focus	推定 Focus
Question Word Pattern 200	201	0.73	0.45
Question Word Pattern 10	11	10.988	8.984
Li's Answer Class	6	25.312	22.203
Question Conceptual Class	7	47.422	45.266

6 おわりに

本研究は，対話型質問生成分野において，回答を用いず質問を生成する手法を初めて提案した．実験結果から，質問の焦点の推定は生成される質問の質の向上に貢献することがわかった．一方，質問パターンの使用は，推定が容易な質問パターンは質問生成に効果がなく，生成に効果的な具体的なパターンは推定が難しいことがわかった．推定が困難な原因の一つに，質問間に依存関係のない質問であっても推定可能であるとした仮定が誤りであることがあげられる．今後，依存関係のある質問のみに注目し本提案手法の効果について検証する．

参考文献

[1] The process of question answering - a computer simulation of cognition. *American Journal of Computational Linguistics*, Vol. 6, No. 3-4, 1980.

[2] 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.

[3] Yifan Gao, Piji Li, Irwin King, and Michael R. Lyu. Interconnected question generation with coreference alignment and conversation flow modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[4] Kalpesh Krishna and Mohit Iyyer. In *Association for Computational Linguistics, Year = "2019", Title = Generating Question-Answer Hierarchies*.

[5] Xin Li and Dan Roth. Learning question classifiers. In Tsuei-Er Chen and Yi-Fen Liu, editors, *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), August 24 - September 1, 2002, Taipei, Taiwan*, pp. 556–562. Morgan-Kaufman Publishers, San Francisco, CA, USA, 2002.

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

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