国際会議における質疑応答練習を目的とした ChatGPT による質問生成とその評価

相場真由子 [1]、齋藤大輔 [2]、峯松信明 [2]
[1] 東京大学工学部、[2] 東京大学大学院工学系研究科
{aiba, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

概要

本研究では、国際会議における質疑応答の練習を支援する目的で、大規模言語モデルである ChatGPT に当該論文に関する質問を生成させ、論文を執筆した学生、及び、英語プレゼンを指導する語学教師に評価させた。いくつかの方法でテキストベースで専門的な質問を生成させ、評価尺度としては妥当性や有用性を考え複数の尺度を用意した。その結果、論文のキーワードや引用文献を特定せず、「論文 PDF を入力して質問を生成させる」簡易的な方法による質問が最も高い妥当性・有用性を示した。国際会議に向けた質疑応答の練習は、語学教師側に専門知識がない場合でも、本手法を用いることで、より現実的な練習が可能となると考えられる。

1 はじめに

英語は国際的な学術交流において不可欠な言語である。特に、学生・研究者は英語で論文を読み書きする能力、英語で発表する能力、英語で他の研究者と議論（質疑応答）する能力が求められる。

初めての国際会議発表を前にした学生は、主に発表と質疑応答の二つを練習することになる。発表練習については、タイムキープやスライドの調整、発表の流れ等を自分で確認できる。一方で、質疑応答の練習では、論文内容を理解している英語の質問者が必要になる。指導教員が対応できれば良いが、時間的に困難な場合も多い。また、英語プレゼン指導を授業として実施する英語教師は、一般の発表・質疑応答は指導できるが、内容の専門性が高くなると、質問することが難しくなる。

このような課題に対処するために、近年注目されている、ChatGTP[1] に代表される大規模言語モデル (Large Language Model: LLM) と音声処理技術を統合し、任意分野の学術論文に対して質問を（音声で）生成し、論文執筆者がそれに答える形の質疑応答支援システムの構築を進めている。大規模言語モデル＋音声インタフェースによるオンライン英会話教材は既に市場に登場しているが [2]、それを国際会議の質疑応答という場面に適用させることが本プロジェクトの目的である。そこで本研究では、ChatGPT に対して質問するのではなく、学会論文に対する専門的な質問をいくつかの方法で ChatGPT にテキストベースで生成させ、論文を執筆した学生と語学教師に、複数の観点から評価させた。

2 先行研究

2.1 プロンプトエンジニアリング

ChatGPT のような AI チャットボットでの対話において、ユーザは自然言語で指示することによって対話を進める。この指示をプロンプトと呼び、AI チャットボットとの対話の中で AI チャットボットからより良い返答を得られるように指示文章を構造化することをプロンプトエンジニアリングと呼ぶ。

プロンプトには、人工言語であるプログラミング言語のように正確に定義された文法があるわけではないが、プロンプトの僅かな変化によって返答が変わることが知られており、様々な効率的なプロンプトが提案されている [3][4]。文献 [3] では、全 16 種類のプロンプトが紹介されており、メタ言語作成パターン (”When I say X, I mean Y.”) やペルソナパターン (”Act as persona X.”) といったプロンプトが有効であることが示されている。

2.2 大規模言語モデルによる質問生成

対話に限らず、質問・問題を生成させるプロンプトについても研究が行われている。出題してほしい範囲の説明文とテーマを与え、例えば国語の試験で用いることのできる質の良い問題を生成させるプ

表 1　質問生成モードの組み合わせ

	参考文献あり	参考文献なし
節ごと	A	D
キーワードごと	B	E
指定なし	C	F

表 2　論文一本当たりの平均質問数

	参考文献あり	参考文献なし
節ごと	13.6	12.1
キーワードごと	9.8	7.5
指定なし	9.3	7.5

ロンプトが検討された [5]。更に [6] のような、問題形式 (4 択、記述式) や教科、出題範囲を選び、問題を生成するサービスが提供されている。また、文献 [7] では、データサイエンスの入門分野において、テキストベースの学習教材から問題を生成し、評価するためのパイプラインを検討している。

3　発表論文に対する質問生成実験

3.1　用いた国際会議論文

国際会議で発表した経験が 2 回以下の修士学生 8 名に会議論文を提出して頂いた。学生本人は第一或いは第二著者である。2 本が音声情報処理、3 本が英語教育の技術支援、3 本が画像処理、ネットワーク、電磁波の論文となっている。

3.2　質問の生成方法

今回の研究では、ChatGPT PLUS の GPTs を利用した。GPTs とは、自分でコンフィギュレーションをカスタマイズして、ある特定の目的に特化した GPT を作成できる機能である。GPTs では、PDF や画像ファイル、プログラミングコード等を Knowledge としてアップロードしておくことができる。

今回は GPTs を利用して、テキストベースの会話の中で学術論文が与えられ、それに対する質問を生成することを目的とした GPT、Thesis Explorer を作成した。ここで、Thesis Explorer は質問の難易度や分野、一度に生成する質問数には制限を設けず、専門的な質問も一般的な質問も生成するよう指示した。プロンプトの具体例を appendix に示す。

論文に対し、1) 論文のどの部分から質問生成を行うのか、及び 2) 参考文献を Thesis Explorer に明示的に読ませるか否か、を考慮して以下の 6 通りの生成モードで質問を生成させた。具体的には、質問の生成範囲は、以下の 3 通りの方法で指定した。

- 節毎に質問を生成。
- 列挙されているキーワード毎に質問を生成。
- 節やキーワードを指定せずに質問を生成。

表 3　生成質問に対する学生の評価基準

Criteria	Description
Relevance	Is the question relevant enough to the topic of your paper?
Clarity	Is the question clear enough to understand?
Specificity	Is the question specific enough to answer? If the question is too general or too vague, you may have trouble in answering the question.
Inspiration	Did the question inspire you? Did the question open your eyes?
Expected	You may have made presentation rehearsals in advance. Is the question expected enough?

表 4　生成質問に対する語学教師の評価基準

Criteria	Description
Likeliness	Is the question likely to be raised from you or general language teachers?

参考文献の有無に関しては、前述した GPTs の機能として利用できる Knowledge として、参考文献 10 件をアップロードしたものとアップロードしていないものを用意した。表 1 に各生成モードを示す。表 2 には、各モードにおいて生成された論文一本当たりの平均質問数を示す。節ごとに個別に生成させると、最終的な質問数は多くなる。

3.3　評価方法

生成された質問と 6 つの生成モード（A〜F）について、評価者として、当該論文の第一或いは第二著者の学生、及び大学で英語を教えている語学教師に評価させた。スプレッドシート 1 枚の中に、当該論文に対して 6 つの生成モードで生成された 6 つの質問群が記載されている。学生には当該論文に対する質問群のみを（平均 58.9 個）、教師には全論文の全質問（471 個）を評価させた。尚、教師には論文の概要と発表スライドを提供し、これらを読んだ後、評価してもらった。なお評価者は、質問が A〜F の生成モード毎に示されていることは知らされているが、質問群と生成モードの対応は知らされていない。以下、学生と教師に示した評価項目を説明する。

質問に対する妥当性・有用性の評価

学生には、各質問に対して、Relevance, Clarity, Specificity, Inspiration, Expected の 5 項目（表 3）に対して 1〜5 の 5 段階で評価させ、更に、質問に対する回答が論文の中にあるかどうかを回答させた。

教師には、Likeliness（一般的な教師、或いは、自身が当該質問を投げかける可能性があるか）という 2 項目に対して likely, rather likely, cannot judge, rather unlikely, unlikely の 5 段階で評価させた（表 4）。(rather) unlikely の場合には、理由も回答させた。

ChatGPT は任意の分野について高い専門性を有した質問ができると仮定すれば、学生が十分 relevant, clear, specific, inspiring と評価した質問（例えば評価値が 3 点以上）に対して、教師が unlikely, rather unlikely と評価した割合が高くなると予想される。

生成モードに対する選好度の評価

評価者に個々の質問に対して評価させることで、評価者は、6 つの生成モードの「質問傾向」を把握することになると推測される。そこで、質問を評価させた後、各モードの選好度合いについても評価させた。学生には「練習で使いたいもの」はどれか、教師には「授業で使いたいもの」はどれか、という観点から、合計 100 点を割り振る形で評価させた。

3.4 アンケート結果

全論文、全生成モードで生成された質問に対する学生の評価値、教師の評価値を、生成モードごとに集計した。学生の結果を表 5、教師の結果を表 7 に示す。表 7 には 2 項目の和（L+RL、U+RU）も示している。質問に対する回答が論文中にはない（つまり鋭い質問であった）と学生が判定した割合を表 6 に示す。全質問を対象とした場合と、Relevance を 3 以上と判定した質問を対象とした場合を示している。各質問への回答後に 6 つの生成モードを相対評価させた結果を表 8 に示す。

3.5 考察

本研究の趣旨は、どの生成モードが質疑応答練習に相応しいのかであるため、まず表 8 に着目する。学生、教師何れも節もキーワードも特定せず、参考文献も明示的に読み込ませない、F を最も高く選好した。表 1 に示されるように、6 つのモードを参考文献の有無或いは質問生成範囲別で分けると、参考文献の有無で選好度は大きく異なり、参考文献を読み込ませない方が選好度が高い。以下、参考文献の有無に着目して考察する。

表 5 より、何れの評価項目においても学生が最も高く評価したのが F であった。また、何れの項目も参考文献を明示的に読み込ませると、学生の評価を落とす傾向にある。表 6 より、答えが論文の中にない質問（鋭い質問）の割合は、質問が十分に関連性の高いものに限定すると F の評価が高い。全ての質問を考慮すると F より高いものがあるが、これは表 5 から分かるように、参考文献を読み込ませたモードには関連性の低い質問が含まれるためである。

表 7 より、一般教師、評価者自身のどちらを想定する場合でも、F が最も L+RL が高く、最も U+RU が低い。参考文献の有無を踏まえれば、参考文献を読み込まない方が L+RL が高く、U+RU が低くなる。これは、参考文献を読み込ませると質問の専門性が高くなるためだと考えられる。語学教師は専門性が高い質疑応答練習を授業で実施することが難しいと考え ChatGPT を導入したが、期待に反して、学生は教師が対応できる質問の割合が高いモードを高く評価している。

表 5 各モードに対する学生の評価（平均、分散）

	Relevance	Clarity	Specificity	Inspiration	Expected
A	3.90, 1.89	3.96, 1.59	3.69, 1.72	2.29, 1.19	2.61, 2.19
B	3.97, 1.84	4.01, 1.31	3.69, 1.57	3.18, 1.55	2.87, 2.04
C	3.29, 2.30	4.03, 1.38	3.97, 1.11	2.58, 1.88	2.43, 1.78
D	4.45, 0.77	4.37, 0.86	3.88, 1.44	3.00, 1.65	3.03, 2.09
E	4.30, 1.50	4.27, 1.25	3.85, 1.59	3.28, 2.04	3.10, 2.29
F	4.68, 0.36	4.70, 0.38	4.35, 0.91	3.45, 1.98	3.60, 1.84

表 6 答えが論文の中にないと学生が判定した割合 (%)

	A	B	C	D	E	F
全て	46.8	46.2	54.5	29.9	46.7	40.0
R≥3	35.2	35.4	31.7	28.0	39.6	40.0

表 7 各モードに対する教師の評価（項目の選択割合）
L:Likely, RL:Rather Likely, C:Cannot judge,
RU:Rater Unlikely, U:Unlikely
一般的な語学教師がその質問を呈するかどうか？ (%)

	L	RL	L+RL	C	U+RU	RU	U
A	18.2	27.9	46.2	4.0	49.9	29.9	20.0
B	19.5	24.2	43.7	7.3	49.0	28.1	20.8
C	21.5	23.1	44.6	4.2	51.2	20.7	30.6
D	22.0	30.2	52.2	7.0	40.8	26.3	14.5
E	17.6	30.5	48.1	7.2	44.7	24.3	20.4
F	25.6	30.4	56.0	7.79	36.2	22.1	14.1

評価者自身がその質問を呈するかどうか？ (%)

	L	RL	L+RL	C	U+RU	RU	U
A	20.2	32.2	52.4	4.3	43.4	22.9	20.5
B	22.1	29.6	51.8	7.3	40.9	20.6	20.3
C	27.0	21.1	48.1	5.1	46.8	18.2	28.6
D	24.7	36.8	61.4	7.5	31.1	18.5	12.6
E	17.6	38.0	55.6	7.2	37.3	17.1	20.2
F	32.9	34.6	67.4	6.57	26.1	13.9	12.2

表 8 各生成モードへの平均配点

	A	B	C	D	E	F
学生	10.0	15.0	9.1	18.4	21.0	26.5
教師	10.6	13.7	14.1	17.7	17.4	26.5

表 9 評価値が 3 以上の質問の割合（%）

	Relevance	Clarity	Specificity	Inspiration
A	80.7	83.5	77.1	41.3
B	83.3	83.3	79.5	67.9
C	62.1	84.8	89.4	51.5
D	95.9	94.8	81.4	60.8
E	88.3	91.7	83.3	68.3
F	100.0	98.3	93.3	68.3

表 10 有効質問に対して U/RU が選択された割合（%）

一般的な語学教師を想定

	Relevance	Clarity	Specificity	Inspiration
A	47.2	48.5	50.6	51.5
B	47.4	48.5	49.2	49.1
C	41.9	49.4	50.0	47.1
D	41.0	40.8	42.0	42.7
E	42.1	41.8	42.7	42.3
F	32.2	32.5	31.8	40.9

評価者自身を想定

	Relevance	Clarity	Specificity	Inspiration
A	40.9	42.8	44.8	45.6
B	40.0	41.8	42.7	41.8
C	37.4	46.7	46.9	43.1
D	34.4	34.6	35.2	35.6
E	36.8	36.4	36.7	36.6
F	24.7	24.9	24.1	20.3

そこで、教師による U+RU 判定に対して、学生が各評価項目に対して 3 点以上と評価した質問に限定して集計した。ここで、3 点以上の質問を有効質問と定義する。有効質問の割合をモード毎に表 9 に示す。どの指標においても、有効質問割合は F が最も多かった。次に、有効質問に限定して教師が U+RU と判定した割合を、モード別、項目別に表 10 に示す。有効質問に限定した場合でも、F が最も U+RU が低く、参考文献を読み込ませない方が U+RU が低い。これは表 7 と同じ傾向である。つまり、全質問、有効質問のどちらを対象としても、学生は語学教師が対応できる質問を多く含むモードを選好し、ChatGPT が有すると考えられる深い専門性が有効に寄与できていないと考察できる。

しかし、別解釈として、評価者である学生の知識不足により、ChatGPT が生成した質問の意図、意義を十分把握できていない可能性がある。実際の質疑応答でも質問者と発表者の会話が噛み合わないことがしばしば起きるが、これは多くの場合、発表者側に十分な背景知識、関連知識がないために起こる。類似したことが今回の質問評価で起きている可能性は否めない。例えば、指導教官や博士号を取得した研究者に評価してもらったり、更には、雑誌論文への査読の質問として ChatGPT に生成させると、参考文献を明示的に読ませたモードの質問が選好される可能性がある。今後、知識レベルを変数として評価者を募り、再実験してみたい。

次に、表 6 より、Thesis Explorer が生成する質問のうち、生成モードにより差はあれど、4〜7 割程度の質問が論文の中に答えがあるという結果であった。ここで、Thesis Explorer に渡しているのは発表スライドではなく論文であることに注意して考察を加える。発表内容や、聴衆のバックグラウンドによって、実際の質疑応答で論文中に記載のある事柄について質問されることも十分考えられる。更に、英語での質疑応答練習にこのシステムを利用することを考えれば、論文内で議論されている事柄について質問した後で発展的な質問をするという手順は教育的観点から考えて妥当であるから、質問に対しての答えが論文の中にある質問も十分有意義である。但し、この割合がどの程度であるべきかについては議論の余地があり、今回の実験においてはその配分の選好度について調査を行なっていないため、被験者に追加の調査をする必要がある。

また、学生にこのシステムを自分の質疑応答練習の前に使いたいかというアンケートにおいて全員から前向きな回答を得た。語学教師からは、プレゼン練習の授業の前に論文に対する質問を生成させ、それらの質問を読んでから質疑応答の練習に臨む場合に効果的であるという意見が複数寄せられた。

4 まとめ

本研究において、学術論文に対する質問をテキストベースで生成し、それらを当該論文の著者と語学教師から評価してもらった結果、ほぼ全員から質疑応答練習の為に利用したいという前向きな評価を得た。国際会議での質疑応答を模擬する目的で評価が行われ、特定の範囲や参考文献を指定せずに質問を生成するアプローチが最も高く評価された。しかし、この結果は、評価者の専門知識の有無や利用目的によって結果が変わることが予想される。

そのため、今後の研究では、評価者や利用目的を変更して再度実験を行うことが期待される。さらに、今回の実験の被験者であった学生や語学教師からの評価の理由を更に詳しく調査することで、実際の利用者が重視する要素についての理解を深め、更なる改良を重ねていく必要がある。

参考文献

[1] Chatgpt. https://openai.com/chatgpt. (Accessed on 09/18/2023).

[2] Ai 英会話アプリ「スピーク」. https://www.usespeak.com/jp. (Accessed on 01/11/2024).

[3] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. **arXiv preprint arXiv:2302.11382**, 2023.

[4] Maciej P Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering–example of chatgpt. **arXiv preprint arXiv:2303.05352**, 2023.

[5] Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G Baraniuk. Towards human-like educational question generation with large language models. In **International conference on artificial intelligence in education**, pp. 153–166. Springer, 2022.

[6] 【教員必見！】chatgpt 活用でテスト問題を 10 秒で作成！ — sakubun. https://sakubun.ai/template/test. (Accessed on 01/02/2024).

[7] Huy A. Nguyen, Shravya Bhat, Steven Moore, Norman Bier, and John Stamper. Towards generalized methods for automatic question generation in educational domains. In **Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption**, pp. 272–284. Springer, 2022.

A　本研究に用いた Thesis Explorer のプロンプトについて

本研究で用いた Thesis Explorer は以下のリンクより利用できる。（尚、ChatGPT PLUS への登録が必要になる。）https://chat.openai.com/g/g-OpWRbiwL7-thesis-explorer

以下に、本研究で使用した Thesis Explorer に与えた Instructions を示す。

Role and Goal: Thesis Explorer is tailored for analyzing engineering theses, formulating diverse questions on algorithm effectiveness, data analysis, experimental design, and results relevance. It handles technical and conceptual queries adeptly.

Constraints: The GPT will generate a broad spectrum of questions, covering both intricate technicalities and wider conceptual aspects, without limitations on the complexity or nature of inquiries.

Guidelines: Questions are to be prefaced with relevant context, providing a foundation for each inquiry. The GPT can also offer insights or suggestions related to the question. This approach aids in deepening the user's understanding of the thesis content.

Clarification: If the thesis content is unclear or requires additional details for effective questioning, the GPT will request clarification.

Personalization: Maintaining a scholarly and academic tone, Thesis Explorer adapts its questioning style to the user's responses and preferences, ensuring a tailored and engaging dialogue.