

記述式解答群に対する観点を指定した概要把握 —観点の指定法に関する検討—

平澤賢人*1・高瀬治彦*1
Email: 425M237@mic-u.ac.jp

*1: 三重大学大学院工学研究科電気電子工学専攻

◎Key Words

演習支援, 解答要約, 大規模言語モデル

1. はじめに

教育現場では、学習者に対して質の高い講義を行うことが、講師に期待される。そのため、講師は講義における学習者の理解状況に応じて、フィードバックを行う。^(1, 2) このため、講師は小テストや課題などの問題演習を活用し、学習者の理解状況の把握を行う。しかし、講義は、多人数講義として多くの場合実施される。多人数講義では、講義内に学習者の理解状況を把握するのは時間的にも労力的にも困難である。

講義の演習の解答形式には、穴埋め式や選択式、記述式などさまざまな種類がある。これらの演習のうち、学習者の理解状況をより詳細に把握することができるのは記述式の演習である。これは、学習者が自身の言葉で解答を記述するため、解答に学習者の理解状況が現れやすいためである。しかし、記述式の解答形式では、学習者の解答が多様であるため、要約等の支援がされてきた。^(3, 4) 文献(3)の手法では、解答を計算機によって採点を行い、支援している。しかし、結果が点数の形で要約されるため、フィードバックに必要な情報が欠落してしまう。また、文献(4)の手法では、解答群から自動抽出された重要語の前後を分類し提供している。しかし、講師が解答に対して着目すべき観点は複数存在するため、それを考慮した要約が必要とされている。

我々の研究グループでは、講師が解答に対して着目したい観点を、クエリ(観点を質問文の形式で表したもの)として提示することで、クエリに沿った学習者の解答の概要を講師が素早く把握できるよう提示する方法について検討してきた。しかし、クエリの記述方法によって抽出される記述内容が変化してしまう問題について検討されていない。⁽⁵⁾ 本稿では、投げかけるクエリの表現ゆれとそれに対する記述抽出の変化について調査を行い、適切な観点指定法について検討する。

2. 記述式解答群における解答概要把握支援システム

本節では、従来研究である二村らの解答の概要把握支援システムについて説明を行う。

2.1 概要把握支援システムの概略

従来法である二村らのシステムは、解答中に複数存在する着目したい観点を「クエリ」(質問形式の文章)により指定することで、クエリに関連する記述を学習者の解答から抽出し提供するものである。短文記述式解答の課題の解答に対して講師が自分の知りたい情報に関連する

クエリを投げかけることで、クエリに沿った内容を抽出し、クラスタリングして提供する。⁽⁵⁾

具体的な例を使用して説明する。まず、講師は「コンピュータの5大装置それぞれの働きを100~200文字で説明しなさい」という課題を課したとする。学習者はこれに解答する。そして、講師は解答群閲覧のための観点を決定する。例えば、演算装置(5大装置の一つ)の働きに対して着目したとする。講師はこの観点に対するクエリを、「演算装置は何を行いますか?」のように指定する。システムは、このクエリに従い、各解答から該当する箇所を抽出する。例えば「演算装置がデータの算術演算や論理演算などの演算、制御装置がすべての装置のコントロール、記憶装置がデータやプログラムの記憶、入力装置がプログラムやデータを主記憶装置への入力、出力装置がコンピュータが処理したデータの出力という働きをしている。」という解答から、「データの算術演算や論理演算などの演算」という部分を抽出する。すべての解答からの抽出結果を、その意味に基づいてクラスタリングし、講師に提供する。最後に、講師はこの結果から、指定した観点に関する記述の概略を把握し、これに基づいて学習者の理解状況を把握する。

2.2 クエリの表現ゆれについて

二村らのシステムでは、講師が観点をクエリという文章(の断片)により指定する。そのため、同じ観点を指定するために、さまざまな表現のクエリが使用されてしまう。例えば、「演算装置」に関する記述を得たい場合に、単純に「演算装置は何を行いますか?」というクエリがある。また、丁寧なクエリとして、「コンピュータの5大装置には演算装置、制御装置、記憶装置、入力装置、出力装置がある。この5大装置のうち、演算装置は何を行いますか?」というものもある。さらに、体言止め「演算装置が行うものは何?」や、言い換え「演算装置の働きは何ですか?」などの表現のクエリが講師により指定される可能性がある。

クエリに従った解答からの抽出には、大規模言語モデル(LLM: Large Language Model)を使用しており、クエリの変化による抽出結果の変化には不明な点が多い。実際、2.1節の例で、クエリを「演算装置が行うものは何?」とした場合には、抽出結果は「データの算術演算や論理演算などの演算、制御装置がすべての装置のコントロール、記憶装置がデータやプログラムの記憶、入力装置がプログラムやデータを主記憶装置への入力、出力装置がコンピュータが処理したデータの出力という」となる。

したがって、クエリの表現ゆれに対する記述内容の変化を具体的なデータを用いて比較検討することが必要である。次節では、被験者に詳細なクエリを提示し、抜き出してもらった記述と被験者が作成したクエリを LLM に投げかけて抽出した記述の比較を行う。これにより、最適な記述を抜き出すクエリを調査し、講師が安定して投げかけられるよう支援する方法の一助とする。

3. 実験

本節では、講師のクエリがゆれることでどの程度抽出結果に影響を及ぼすかを、実験を通じて調査をした。

3.1 実験方法と条件

本実験では、複数のクエリを用いて、LLM または被験者が、解答群からの抽出を行い、それらを比較する。

分析対象の解答群は、工学系の大学の講義「計算機工学」(計算機のハードウェアに関する講義, 受講者 77 人)の各回の最後に、講義内容に関する課題を課すことで収集した。実験には、2 回分の解答群をもちいた。加えて、各回の課題に対して、講師の観点を詳細なクエリの形で記述したものをあらかじめ用意した。以下に、具体的な課題とクエリを記す。

課題 1「フラッシュメモリの記憶方式を簡単に説明しなさい (10~30 文字)」回答数 (71 人)

元クエリ 1

「フラッシュメモリは、電源を切ってもデータが消えない記憶媒体です。データを記録する際、その中核となるフローティングゲートと呼ばれる部分に電子を蓄えます。このフローティングゲートは、周囲が電気を通さない絶縁体の酸化膜で囲まれています。では、フラッシュメモリは記憶をするために、酸化膜でどのような働きをしますか？」

元クエリ 2

「フラッシュメモリは、電源を切ってもデータが消えないように、内部に電子を閉じ込めて記録を保持しています。この閉じ込めた電子が勝手に逃げ出さないよう、あるものが壁となって移動を管理しています。では、フラッシュメモリはデータを記憶するために、何で電荷を制御していますか？」

課題 2「仮想メモリの目的・実現のための基本的な考え方を答えなさい」回答数 (64 人)

元クエリ 1

「仮想メモリとは、主記憶装置の物理的な容量を超えるメモリ空間をプログラムに提供するために、補助記憶装置の一部を主記憶装置の延長として利用する仕組みです。データを主記憶と補助記憶の間で動的に移動させることで、限られた物理メモリを効率よく管理します。この仕組みの中で、主記憶装置はデータの一時的な保持場所として中心的な役割を担っています。このような仮想メモリの仕組みにおいて、主記憶装置にはどのような特徴 (ポジティブなものとネガティブなものどちらでも可) がありますか？」

元クエリ 2

「仮想メモリとは、プログラムが必要とするメモリ容量が主記憶装置の物理的な容量を超える場合に、補助記憶装置を活用することで不足を補う技術です。現代の OS では、複数のプログラムが同時に動作するため、主記憶装置だけでは容量が不足するケースが生じます。そこで補助記憶装置を組み合わせることで、より大きなメモリ空間を実現しています。このような仮想メモリの仕組みを実現するための考え方として、補助記憶装置をどのように使いますか？」

さまざまなクエリおよびそれに対する抽出結果は、以下の手順により、被験者から収集した。ここで被験者は、当該講義の受講経験者である大学生 6 人とした。

1. あらかじめ定めた詳細なクエリと同義のクエリを被験者が 3 つ生成
(各クエリに対する指示, 端的, 体言止め, 言い換えの 3 つ)
2. あらかじめ定めた詳細なクエリに従い、被験者が解答群から記述を抽出
3. 手順 2 の後、被験者が手順 2 の抽出結果を得るための適切なクエリを生成
4. 手順 1 と 3 で生成したクエリに従い、LLM が記述を抽出

抽出結果には、解答中の一部を抽出する場合だけでなく、該当箇所がないことを意味する「抽出なし」の場合がある。これをふまえて、2 つの抽出結果群の、同じ解答からの抽出結果が、ともに存在するものと、一方もしくは両方で「抽出なし」の場合に分けて、抽出結果の類似度を評価する。

抽出結果がともに存在する場合は、抽出結果の類似率により評価した。これは、同じ解答から抽出した対となる抽出結果に対して、LLM により意味を表すベクトルに各記述を変換し、それらのコサイン類似度により評価した。抽出結果群どうしの類似度は、そこに含まれる抽出結果うち該当する全ペアの類似度の平均とした。なお、ベクトル化には paraphrase-multilingual-mpnet-base-v2⁶⁾を使用した。また、クエリ同士の類似率も同様に求めた。

抽出結果がないものを含む場合は、抽出漏れと抽出過剰により評価した。これは、解答ごとに抽出結果の有無により、抽出漏れ (人 : 抽出あり, LLM : 抽出なし), 抽出過剰 (人 : 抽出なし, LLM : 抽出あり) を計数した。

元クエリから被験者が抜き出した記述と複数のクエリから LLM が抽出した記述の比較結果を以下に示す。

3.2 人と LLM の抽出結果に関する考察

まず、人の抽出と LLM の抽出にどの程度差があるのかを明らかにするために、元クエリから被験者が抽出した記述と複数のクエリから LLM が抽出した記述を比較する。

課題 1 元クエリ 1 (1-1) ~ 課題 2 元クエリ 2 (2-2) までの被験者の抽出結果に対するクエリの抽出結果の類似率・抽出漏れ・抽出過剰を次の表 1~4 に示す。

表 1 1-1 (元クエリ-人 VS LLM) の結果 (6 人)

クエリ/ 評価指標	類似率	抽出漏れ	抽出過剰
元クエリ	0.90	3.8	7.5
端的	0.87	0	21.7
体言止め	0.87	0	21.2
言い換え	0.88	0	21.4
適切	0.88	0	21.7

表 2 1-2 (元クエリ-人 VS LLM) の結果 (6 人)

クエリ/ 評価指標	類似率	抽出漏れ	抽出過剰
元クエリ	0.91	0.83	25.7
端的	0.92	0.17	32.8
体言止め	0.85	0	32.1
言い換え	0.88	0.2	32.6
適切	0.89	0	31.1

表 3 2-1 (元クエリ-人 VS LLM) の結果 (6 人)

クエリ/ 評価指標	類似率	抽出漏れ	抽出過剰
元クエリ	0.59	20.5	6.3
端的	0.53	5	31.8
体言止め	0.51	3.6	37.7
言い換え	0.55	2.4	37.1
適切	0.58	3.7	34.9

表 4 2-2 (元クエリ-人 VS LLM) の結果 (5 人)

クエリ/ 評価指標	類似率	抽出漏れ	抽出過剰
元クエリ	0.80	7.6	20.4
端的	0.82	0.44	28.6
体言止め	0.77	1.32	27.9
言い換え	0.80	0	28.6
適切	0.81	0.68	27.9

まず、類似率に着目すると、課題によって異なるが、クエリ間での大幅な数値の変動は見られなかった。このことから、クエリの表現がゆれたとしても、抽出結果は意味的な変化がほとんど発生しなかったことがわかる。

次に、抽出漏れ、抽出過剰に着目すると、元クエリと他のクエリとで異なる傾向になった。他のクエリと比べて、元クエリでは抽出漏れが多く、抽出過剰が減っている。抽出漏れと抽出過剰の合計を見ると、他のクエリより少ない。ただし、表 3 のみ他と傾向が異なる。これは、被験者による抽出結果に含まれる「抽出無し」の場合が、他と比べ多かったため、異なる傾向になったと考える。そのため、以下ではこの場合を除外して考察する。

以上より、元クエリのように詳細にクエリを記述することで、人に最も近い抽出結果を得ることができる。また、他のクエリ間では大きな違いはないが、わずかに「体言止め」の場合に、最も人の抽出結果から外れる。

3.3 LLM によるクエリの抽出結果に関する考察

3.2 節より、元クエリ（詳細な記述）により最も高精度

な結果を、体言止めのクエリにより最も低精度な結果をもたらすことが示された。この 2 つのクエリによる抽出結果を比較することで、クエリの表現ゆれによる抽出結果の影響を確かめる。

まず、元クエリと体言止めのクエリの、クエリ自体の類似率を表 5 に示す。すべての類似度は 0.8～0.9 の範囲にあり、多少の意味的な違いが確認できる。具体的なクエリの例を以下に示す。

体言止め

- ・課題 1-1
「フラッシュメモリが記憶を保持するための酸化膜の働き」
- ・課題 1-2
「フラッシュメモリの電荷制御方法」
- ・課題 2-2
「仮想メモリの仕組みを実現するための考え方として、補助記憶装置を使い方は何か？」

表 5 元クエリ VS 体言止めの類似率

課題	1-1	1-2	2-2
平均	0.84	0.87	0.82

また、課題 1-1, 1-2, 2-2 における抽出結果は以下のとおりである。これらの表は、元クエリによる抽出結果に対する体言止めの抽出結果の類似率・抽出漏れ・抽出過剰を示している。

表 6 1-1 (元クエリ-LLM VS 体言止め-LLM) の結果

評価指標	類似率	抽出漏れ	抽出過剰
平均	0.96	0	17.5

表 7 1-2 (元クエリ-LLM VS 体言止め-LLM) の結果

評価指標	類似率	抽出漏れ	抽出過剰
平均	0.87	0.7	8

表 8 2-2 (元クエリ-LLM VS 体言止め-LLM) の結果

評価指標	類似率	抽出漏れ	抽出過剰
平均	0.86	1.8	15.6

表 6～8 より、類似率は表 1～4 と同様に高い値となった。それに対し、抽出漏れはほとんど無く、抽出過剰が多い結果となった。これは、体言止めのクエリを用いることで、余分な抽出が発生していることを意味する。

最も類似率が低かった課題 2-2 の場合について、具体的に、同じ解答からのそれぞれのクエリによる抽出結果の例を示す。

Case1 (類似率 0.86) :

- ・元クエリによる抽出結果
「酸化膜で電荷を閉じ込め」
- ・体言止めによる抽出結果
「電荷を閉じ込めることで」

Case2 (類似率 0.86) :

- 元クエリによる抽出結果
「補助記憶装置の一部を主記憶に見せかけて使用する
方法主」
- 体言止め
「主記憶に見せかけて使用する方法」

多くの場合は、上に示した例のように、意味的には同義であるが、主語や目的語の有無による違いが見受けられた。したがって、講師が 2 つの記述を見たときに記述間の差は少ないと考える可能性が高い。

以上の結果から、最も人間の抽出に近いクエリは元クエリであり、一方で、体言止めのクエリが人間の抽出から遠いクエリであることが分かった。また、元クエリによる抽出結果と比べ、体言止めのクエリでは抽出過剰が多く、講師が閲覧する際の障害となりかねない。そのため、抽出結果の可読性を高めるためにクエリ情報を追加してもらうような機能やクエリのサジェスト機能を追加することで支援していきたい。

4. おわりに

本稿では、クエリに沿った学習者の解答の概要を講師が素早く把握することを目指した。そのために、現在構築を目指している支援システムのクエリの表現ゆれによる抽出結果の変化を調査した。調査の結果、最も人間の抽出結果に近い結果をもたらすのは詳細なクエリであることがわかった。また、クエリが簡略化されることにより、抽出結果の意味は変わらないものの抽出結果が過剰に抜き出されてしまうこともわかった。そのため、抽出結果の可読性を高めるために、今後はクエリを投げかける際のサジェスト機能やクエリに関する情報を追加してもらえようような機能を追加するなど、適切なクエリを安定して投げかけられるようシステムを改良していきたい。

参考文献

- (1) 中島英博：“多人数講義で学生の深い学習を促す教員の特質”，名古屋高等教育研究 Vol.15, pp.161-177 (2015)。
- (2) 西森敏之：“大学生の授業における態度と数学教師の対策—日本数学会のある調査より—”，高等教育ジャーナル—高等教育と生涯学習—, vol.6, pp.1-31 (1999)。
- (3) 田中博，斎藤健：“ニューラルネットワークを用いた記述式問題の自動採点”，言語処理学会 第 22 回年次大会 発表論文集, pp.123-126 (2016)。
- (4) 伊藤慎治，大庭知也，高瀬治彦，川中普晴，鶴岡信治：“記述式解答群の主要な内容の把握 支援 —三段階表示の効果の検証—”，2016 年度 PC カンファレンス論文集, pp.41-44 (2016)。
- (5) 二村駿輔：“記述式演習の解答群の概要把握の支援—大規模言語モデルを利用したインタラクティブな要約—”，三重大学大学院工学研究科 修士論文 (2023) 。
- (6) Nils Reimers, Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)