

修士論文

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—

令和5年度修了

三重大学大学院 工学研究科

博士前期課程 電気電子工学専攻

二村 駿輔

目次

第1章	はじめに	1
1.1	講師が行う理解状況把握.....	1
1.2	記述式演習における支援.....	2
1.3	本研究での取り組み	2
第2章	自然言語処理技術	3
2.1	Bag of Words	3
2.2	大規模言語モデル	3
2.2.1	主要な大規模言語モデル	3
2.2.2	達成しているタスク	4
第3章	提案手法.....	8
3.1	インタラクティブな概要把握	8
3.2	質問と解答群からの概要把握.....	8
3.2.1	各解答から該当記述の抽出	9
3.2.2	抽出記述のクラスタリング	9
第4章	実験.....	11
4.1	実験条件と手順	11
4.1.1	使用した解答データ・大規模言語モデル.....	11
4.1.2	実験手順.....	11
4.2	LLM による該当部分の抜き出し精度の確認.....	15
4.3	抜き出し箇所のクラスタリング精度の確認.....	17
第5章	まとめ	19
謝辞	20
参考文献	21

図目次

図 1	GPT の事前学習	4
図 2	BERT の事前学習.....	4
図 3	インタラクティブな概要把握システム.....	8
図 4	ファインチューニングの流れ [10]	10
図 6	質問 1-2 の抜き出し結果の樹形図	13
図 5	質問 1-1 の抜き出し結果の樹形図	13
図 7	質問 2-1 の抜き出し結果の樹形図	14
図 8	質問 2-2 の抜き出し結果の樹形図	14

表目次

表 1	MARC-ja の例 [11]	6
表 2	JSTS・JNLI の例 [11]	6
表 3	JSQuAD の例 [11]	6
表 4	JCommonsenseQA の例 [11]	6
表 5	JGLUE のタスク達成度.....	7
表 6	解答例と抜き出し結果	12
表 7	質問 1-1 の抜き出し結果 (単位:%)	16
表 8	質問 1-2 の抜き出し結果 (単位:%)	16
表 9	質問 2-1 の抜き出し結果 (単位:%)	16
表 10	質問 2-2 の抜き出し結果 (単位:%)	16
表 11	抜き出し結果の平均 (単位:%)	16
表 12	把握に使用した記述.....	18
表 13	確認できた記述の割合 (単位:割).....	18
表 14	クラスタ内の異なる内容の記述数 (単位:個数).....	18

第1章 はじめに

教育において学習の指導を行う講師は学習者に対して質の高い講義を行うことが求められている。そのためには学習者の理解状況を講師が把握する必要がある。これを行うための行動やそれを支援する方法が多く存在する。第 1 節では講師に使用されている理解状況把握方法と本研究で着目した方法について説明する。第 2 節では特に記述式演習における支援について説明する。第 3 節では第 1 節および第 2 節をふまえ、本研究で行った取り組みについて説明する。

1.1 講師が行う理解状況把握

講師は学習者の理解状況を把握してそれに適したフィードバック（解説，改善）を行うことが必要である[1]。フィードバックとは学習者の苦手な箇所やつまづいた箇所に応じて、講義で補足説明を行ったり、課題を実施したりすることである。これらの実施により、講義の改善および学習者の理解を深めることにつながる。宮崎らは学習者の理解度を把握することが学習や教育の向上につながると述べている[2]。また、適切なフィードバックを行うことによって学習者の学習意欲が向上し、講義の質を高めることができる。児島は学習者の学習意欲を高めるためには、不安を減らし、「わかる」という経験を重ね、自信をつけさせることが重要だと述べている[3]。学習者がつまづいている箇所について解説や補足説明があることは学習者にとって不安の解消へとつながる。また、溝上らは知識と情報のズレが大きすぎる場合、学習意識は低くなると述べている[4]。つまり、このフィードバックが学習者の理解状況に適した内容である必要がある。それに加え、説明までの時間がかかってしまう場合、学習者が不安に思っていた箇所やつまづいた箇所があいまいとなり、苦手意識を持ったまま放置されてしまう。松田らは学習する内容には前後のつながりがあることが多いと述べている[5]。苦手意識を持ったまま別の範囲の学習内容を学ぼうとしたとき、以前の知識を活用することができず、苦手意識がほかの分野の学習意欲などに影響し、その内容の理解にも影響が出てしまうことも懸念される。これらのことから、講師は学習者の理解状況をなるべく早く把握し、その状況に応じたフィードバックを行うことが重要である。その際に、講師は講義の内容、演習の意図を把握しているため解答の要点と数を確認するだけで学習者全体の理解状況の把握を行うことができる。

理解状況把握のための代表的な取り組みとして、学習者への問いかけや机間巡視、演習などが挙げられる。それらの特徴として、問いかけは即座に学習者からの反応を得ることができるが、日本では他者の存在や授業の雰囲気を感じて質問を行わない傾向が強い[6]ことから、反応が薄く、理解状況の把握に対する効果が得られ難いケースが存在する。机間巡視では学習者が多くなると把握が難しくなることや個別対応が必要なケースが増え、時間を要してしまうことが考えられる。そこで、講義中に行われる演習に着目する。演習には選択式、穴埋め式や記述式がある。その中でも特に学習者自身の言葉で解答を行うことでより理

解状況が把握できる記述式演習に着目する。記述式問題であれば、問題作成に選択肢の作成などの時間のかかる要素が少なく、ほかの出題形式よりも時間がかかりづらい。ここから、記述式演習に着目することとした。ただし、記述式演習から学習者全体の理解状況を把握するためには多くの解答を読む必要がある。特に多人数クラスとなると読むべき解答が必然的に多くなり、時間がかかってしまう。これはフィードバックを行うまでの時間もかかってしまうことにつながる。加えて選択式や穴埋め式と違い、自由記述であるため解答を見るだけでは学習者全体の理解状況把握は難しく、問題の意図を理解したうえで一つ一つの解答の要点を抑える必要がある。なお、ここでの演習は成績を決定するために用いるのではなく、単に理解状況を把握するために用いられる簡単な演習をさす。また、フィードバックには個人個人にあてたものと全体に向けたものもあるが、今回は全体に向けたフィードバックを対象とする。

1.2 記述式演習における支援

記述解答群に対する講師支援は様々な手法で試みられている。単語の出現頻度を表したノードを関係性に合わせて接続して表現する共起ネットワークを作成して講師に提供することで支援する手法がある[7]。この手法では単語の関係をネットワークで表現しているため単語間での共起性や共通点を講師が視覚的に確認することができる。実際に講習会後のアンケート記述に対して共起ネットワーク分析と対応分析を組み合わせて受講者の問題点の明確化した例もある。ほかにも解答群内の単語の頻度を一般文書群内のものと比較することで重要度を計算して重要度が高い語をキーワードとして自動的に抽出し講師に提供することで主要な内容の把握を行う手法も存在する[8]。しかし、これらの支援方法では講師が解析をしなればいけないことや、講師による注目箇所や確認したい単語を指定してそれについての分類をするといったインタラクティブな分類ができないことが問題になっている。インタラクティブな分類ができなければ講師の観点を含んだフィードバックに必要な分類が行われず、解答すべての内容で分類がされてしまう。

1.3 本研究での取り組み

これらの背景から本研究では多人数の講義における記述式演習において講師がフィードバックを行うのに必要な概要とその概数をインタラクティブに素早く把握できることを目的とする。これを実現するため、解答群について講師が指定した観点に関する記述の概要（主要な内容とその概数）を提供する。これにより講師が一つ一つの解答を確認することなく、必要とする要点と関係する解答の概数を把握することができ、素早いフィードバックを行うことが期待できる。

第 2 章 自然言語処理技術

人間が書いたり話したりする文章（自然言語）をコンピュータによって処理して文脈や意味を分析する技術のことを自然言語処理技術という。自然言語処理技術のなかには古典的な単語ベースの分析や機械学習によって大量のデータを学習してモデルを作成するものもある。今回、文書分類や内容要約に使用されている主要な技術を紹介する。

2.1 Bag of Words

Bag of Words とは単語の出現頻度をもとに文章を数値ベクトル化する自然言語処理の技術のことである。文章を数値ベクトルに変換することでコンピュータによって自然言語を扱える形にする。単語の出現頻度をもとにしているため高頻出の単語を確認することで話題の検出が可能になる。また、単語の有無でベクトル化しているので同じ単語を複数使用している文書同士の分類が可能になる。主な処理の流れは文章を単語に分割して単語の種類数を次元数、それぞれの単語の出現回数を要素の値とするベクトルを作成する。そのベクトルをコサイン類似度により比較したりクラスタリングに用いたりすることで分類を行っている。

2.2 大規模言語モデル

言語モデルとは人が話したり書いたりする言語を単語の出現確率によってモデル化したもののことである。大量のテキストデータを学習することで次にくる単語の確率を学習し、文章の構造や表現を獲得するモデルである。その中でも特に深層学習によって高度で大規模なモデルを構築したものを大規模言語モデルという。

2.2.1 主要な大規模言語モデル

大規模言語モデルはさまざまなアーキテクチャ・学習方法に基づいたモデルが開発されており、自然言語処理分野の多くのタスクに利用され好成績をおさめている。アーキテクチャ・学習法の違いで適しているタスクも異なっており、テキストやプログラムなどを生成する生成系のモデルや入力文章の分類・分析・要約などをする識別系のモデルが存在する。生成系のモデルの主要なものとしては GPT[9] があげられる。GPT は Transformer とよばれるニューラルネットワークのデコーダ部分を利用したモデルであり、図 1 のようにランダムに文章の後半を隠して次にくる単語を推測するという学習を繰り返すことで単語間の出現確率を獲得している。次にくる単語を予測するためそれを連続で行うことで文章やプログラムの生成を行っている。識別系のモデルの主要なものとしては BERT[10]があげられる。BERT は Transformer のエンコーダ部分を利用したモデルであり、図 2 のように文章中の単語をランダムに隠してその部分を前後の文から予測するという学習を繰り返して文章の中での単語の出現確率を獲得している。そうすることで文脈を理解した文章の識別

ができる。

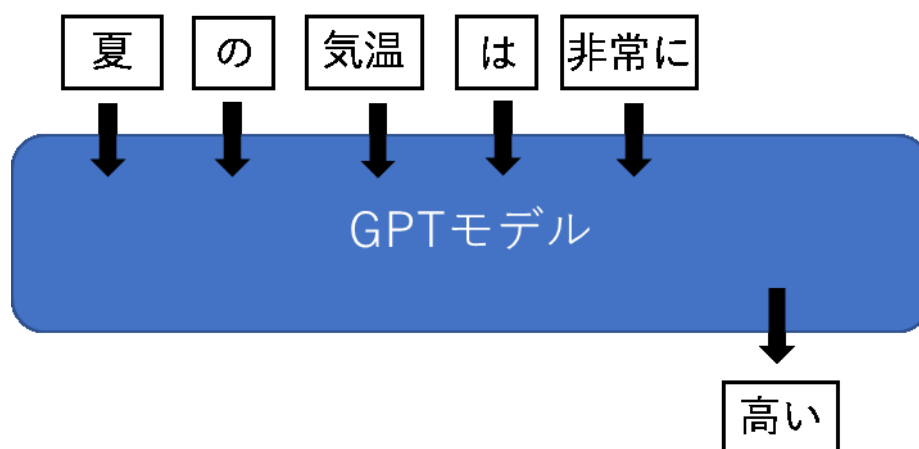


図 1 GPT の事前学習

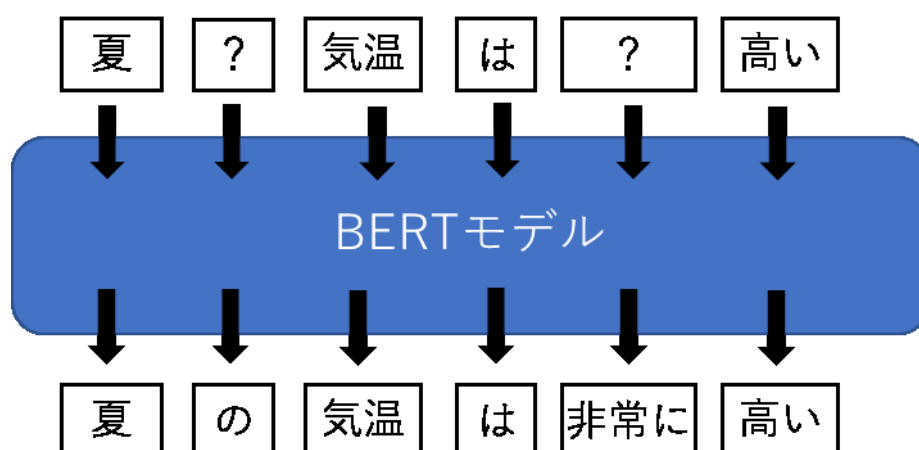


図 2 BERT の事前学習

2.2.2 達成しているタスク

自然言語処理においては文書分類や類似判定、感情分析など多くの達成したいタスクが存在する。言語モデルがそれらのタスクに対してどれだけ言語能力があるのか一律に判断するのは難しい。そこで言語モデルによる言語理解能力の共通の指標としてベンチマークが作成されている。これらより、大規模言語モデルに期待されているタスクが明らかになる。日本語では JGLUE[11]とよばれるベンチマークが作成されている。JGLUE は早稲田大学と Yahoo!JAPAN によって構築、公開された日本語言語理解ベンチマークのことである。文章分類、文ペア分類、QA のタスクから構築されている。文章分類タスクでは MARC-ja という商品レビューのネガポジの二値分類がタスクとされている。文ペア分類タスクでは

JSTS という文のペアの意味的類似度を 0-5 で推定すること, JNLI という文のペアの含意, 矛盾, 中立の推論関係の認識がタスクとされている. QA タスクでは JSQuAD という文書を読み, その文書に関する質問に答えること, JCommonsenseQA という質問に対して 5 つの選択肢がある中から解答を選ぶことがタスクとなっている. タスクそれぞれの例を表 1~5 に示す. このような日本語言語理解のベンチマークに対して表 6 に示すようにベストなモデルは人間のスコアと同等かそれ以上の結果を残している. 内容識別に関するものであるため BERT をベースにしたモデルが利用されている.

2.2.3 タスク達成のための手法

大規模言語モデルの多くはタスクを達成する手法としてファインチューニングを行っている. ファインチューニングとは少量の追加学習をおこなうことでモデルをタスクに合わせたパラメータと出力に調節する手法のことである. 仕組みとしては事前学習した大規模言語モデルに類似判断や分類, ネガポジ判定などの行いたいタスクと同じ形式をした学習データと正解ラベルを用意し, 再度学習させることでタスクにあった出力, パラメータに調節する.

表 1 MARC-ja の例 [11]

レビュー文章	レーティング	ラベル
大切な友人の誕生日にはかならず贈る 1 冊。	5	positive
きれいな商品でした。また安くて良いものをお願いします。	4	positive
組み立ての説明が日本語対応でない付属のモーターでは非力すぎる	2	negative
関節が終わっているこの程度のおもちゃにこの値段はありえない	1	negative
探し物の絵がわかりづらい。ゲームをスタートさせるまでにやや時間がかかる。	5	negative
日本語対応して欲しいところ。その外の面では比類なく優れていると思います。	2	positive

表 2 JSTS・JNLI の例 [11]

文 1 / 前提文	文 2 / 仮説文	類似度	推論関係
街中の道路を大きなバスが走っています。	道路を大きなバスが走っています。	4.4	entailment
テーブルに料理がならべられています。	テーブルに食べかけの料理があります。	3.0	neutral
野球選手がバットをスイングしています。	野球選手がキャッチボールをしています。	2.0	contradiction
フリスビーをくわえた犬がいます。	建物の前にバスが一台停車しています。	0.0	—

表 3 JSQuAD の例 [11]

[タイトル] 東海道新幹線
1987 年（昭和 62 年）4 月 1 日の国鉄分割民営化により、JR 東海が運営を継承した。西日本旅客鉄道（JR 西日本）が継承した 山陽新幹線 とは相互乗り入れが行われており、東海道新幹線区間のみで運転される列車にも JR 西日本所有の車両が使用されることがある。2020 年（令和 2 年）3 月現在、東京駅 - 新大阪駅間の所要時間は最速 2 時間 21 分 、最高速度 285km/h で運行されている
質問: 2020 年、東京～新大阪間の最速の所要時間は
正解: 2 時間 21 分
質問: 東海道新幹線と相互乗り入れがされている路線はどこか？
正解: 山陽新幹線

表 4 JCommonsenseQA の例 [11]

質問: 会社の最高責任者を何というか？
選択肢: 教師, 部長, 社長 , 部下, パイト
質問: スープを飲む時に使う道具は何？
選択肢: スプーン , メニュー, 皿, フォーク, はし

表 5 JGLUE のタスク達成度

モデル / 評価尺度	MARC-ja acc	JSTS Spearman	JNLI acc	JSQuAD EM/F1	JComQA acc
人間	0.990	0.872	0.917	0.873/0.946	0.988
東北大 BERT base	0.957	0.865	0.876	0.879/0.946	0.782
東北大 BERT base (char)	0.957	0.857	0.861	0.864/0.937	0.728
東北大 BERT Large	0.961	0.866	0.878	0.887/0.951	0.822
NICT BERT base	0.960	0.868	0.881	0.904/0.952	0.807
早稲田大 RoBERTa base	0.962	0.867	0.876	0.868/0.926	0.849
早稲田大 RoBERTa Large128	0.954	0.888	0.904	0.881/0.940	0.901
早稲田大 RoBERTa Large512	0.963	0.882	0.904	0.925/0.965	0.873
XLNet-RoBERTa base	0.962	0.836	0.872	-	0.708
XLNet-RoBERTa Large	0.965	0.880	0.902	-	0.842
LUKE Japanese base	0.965	0.877	0.912	-	0.842
LUKE Japanese Large	0.965	0.902	0.927	-	0.893
DeBERTaV2 base	0.970	0.886	0.922	0.899/0.951	0.873
DeBERTaV2 Large	0.968	0.892	0.924	0.912/0.959	0.890

第3章 提案手法

3.1 インタラクティブな概要把握

第1章ではインタラクティブに概要を把握することを目標と述べた。これは多くの記述が含まれる解答の中から講師が必要とする記述についての概要や数の把握を行うということである。講師の指定した着目点、模範解答を入力としてそれに適した解答群の抽出や分類の提示によって概要把握の支援とする。対話形式で講師から何度も入力を受け取り、分類を実行する手法も存在するが、素早いフィードバックをする必要があることと概要が把握できればよい、という点から今回は一度のみ講師からの入力を受け取るものとする。使用する入力は講師の指定した着目点、模範解答などあるが概要把握が目的のため模範解答による分類では正解でない解答内容の把握が困難になってしまう。そこで本研究では講師の指定した着目点として講師の必要とする記述についての質問と解答群の2つを用いて書かれている概要を獲得する。

上記の条件をもとに作成される概要把握システムを図3に示す。システムの入力には解答者一人一人の解答すべてと講師からの質問を使用する。講師からの質問は講師がテキストで作成してシステムに入力する。それをうけてシステムは解答の中の質問に関する記述を内容ごとにグループ化し、代表記述を抽出する。講師には解答群の概要として代表記述とグループ内の解答数を提示する。このシステムを利用することで解答群の概要とその解答数をインタラクティブに素早く把握できるようになる。

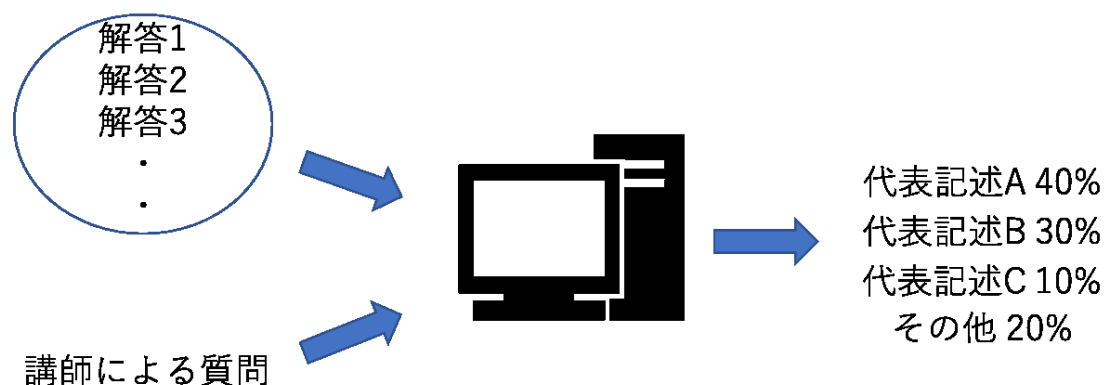


図3 インタラクティブな概要把握システム

3.2 質問と解答群からの概要把握

3.1節で述べたインタラクティブな概要把握は自然言語処理においても困難なタスクになっている。入力が質問と解答であるため類似性による分類は効果がなく、解答をただ分類

するだけでは講師の質問に関する分類ができなくなってしまう。

そこで、代表記述とグループ内の解答数を把握する方法を各解答から質問に対する記述部分の抽出と抽出した記述部分の分類によって実現する。JSQuAD では文書の中から質問に関する内容の抽出ということが高いレベルで行えていることから大規模言語モデルによる記述抽出を行う。文書の分類については前述したように自然言語処理においては多く行われてきた。大規模言語モデルを利用したものから古典的なものまでさまざまある。今回は記述抽出によって解答内の不要な語句は取り除かれていると考えられるため記述のベクトル化とクラスタリングによって計算コストの少ない分類を行った。クラスタリングには多くの手法があるが、クラスタ数を決定する必要があることや記述内容の詳細によってクラスタを細分化や凝集する可能性もあるため階層型クラスタリングによって行う。

3.2.1 各解答から該当記述の抽出

大規模言語モデルによる質問に関する内容の抽出のために事前学習されたモデルを QA タスクに合わせたファインチューニングを行う。QA タスクは識別系のタスクであるため BERT 系列の大規模言語モデルが適している。ファインチューニングの際には一般的な QA タスクの学習データの質問と文書を質問に関する記述を正解データとして学習することで QA タスクにあった出力とパラメータに調節する。ファインチューニングの流れを図 4 に示す。初めに質問と解答の先頭に CLS、間に SEP とよばれるトークンを付与し、質問であるか解答であるかの埋め込み表現を加算してトークナイザーでトークン化した入力文を作成する。それを BERT に入力し、質問に関する記述のはじめと終わりの位置（赤矢印）とともに学習する。

解答から質問に関する記述の抽出には学習したモデルにファインチューニング時と同様にトークン化した解答と質問を入力する。正解データは与えられていないため、ファインチューニングしたモデルによって解答の中で質問に関する記述のはじめと終わりの位置を予測し抽出する。これを解答すべてに行うことで各解答の質問に関する記述の抽出を行うことができる。

3.2.2 抽出記述のクラスタリング

抜き出された記述を形態素に分割して形態素を一単位として Bag of Words によってベクトル化する。この際に、解答の内容でグループ化を行いたいため、自立語の中で感動詞、副詞、接続詞を除いた形態素でベクトル化をする。ベクトル化した記述に対して Jaccard 距離を基準に群平均法で階層併合的クラスタリングを行い、樹形図を作成する。群平均法では、併合後の類似度（非類似度）を併合されたクラスタ内の事例と、ほかのクラスタ内の事例との類似度（非類似度）の平均値を類似度（非類似度）とする[12]。任意の高さを決定することで樹形図からクラスタ化した記述群を得ることができる。高さは講師の確認したい記述の細かさや全体の何割の解答がクラスタになっているかなどによって決定する。

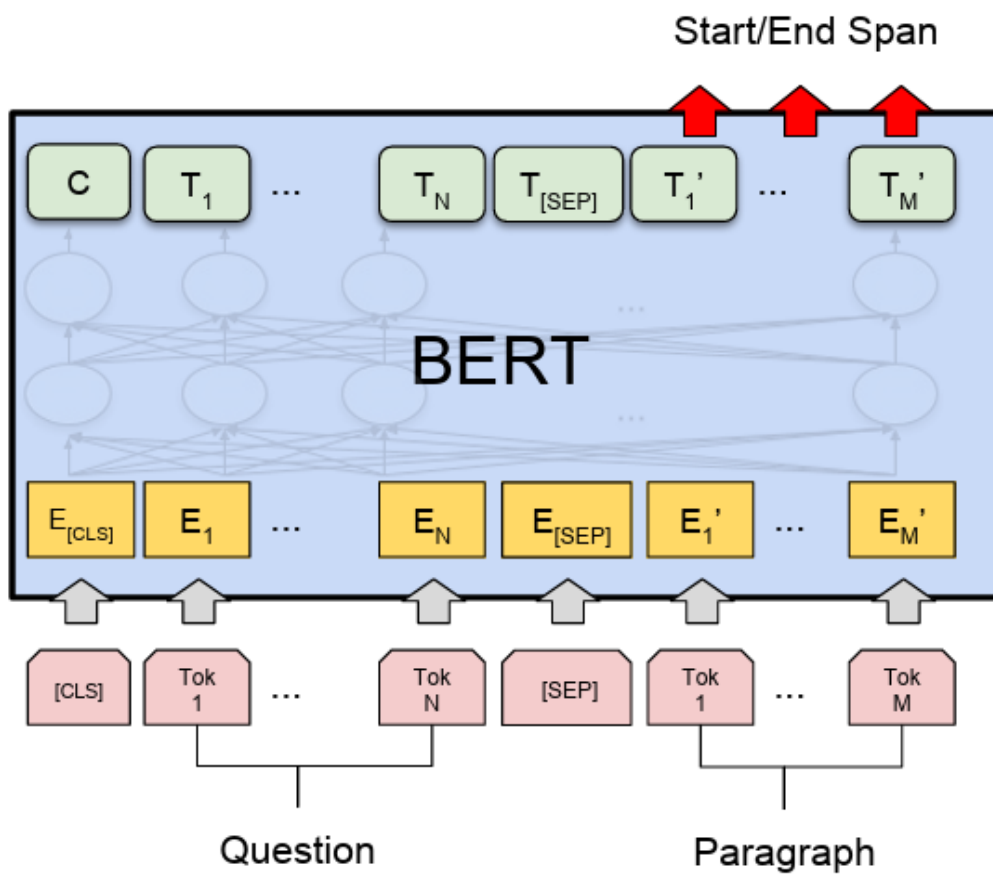


図 4 QA タスクにおける BERT 使用方法 [10]

第 4 章 実験

本章では実際の講義中で行われた演習の記述解答を用いて提案法の有効性の評価をする。そのために大規模言語モデルによる質問に対する該当部分の抜き出し精度の確認と抜き出し箇所のクラスタリング精度の確認をする。

4.1 実験条件と手順

4.1.1 使用した解答データ・大規模言語モデル

実験に使用した記述解答は三重大学工学部で実施された 2022 年度の計算機工学の講義の演習解答である。使用した演習は、問 1「本日の講義をふまえて仮想メモリの目的・実現のための基本的な考え方を答えなさい」問 2「本日の講義をふまえてハーバードアーキテクチャの長所・短所（ノイマン型に比べて）を答えなさい」の二種類である。それぞれ解答数は 64 個と 76 個であった。

大規模言語モデル（事前学習モデル）には文献[13]の DeBERTa の日本語学習モデルを使用し、一般公開されている文章読解 QA データセット（RC-QA データセット）を用いて QA タスク用のファインチューニングを行ったものを使用した。

4.1.2 実験手順

問 1 に対して質問 1「仮想メモリ実現のための考え方は？」質問 2「仮想メモリの目的は？」、問 2 に対して質問 1「メリットは？」質問 2「デメリットは？」という質問を作成し、大規模言語モデルを用いて解答から質問に関する記述の抜き出しを行った。結果の例を表 6 に示す。その後、抜き出した記述をベクトル化し樹形図を作成した。作成した樹形図を図 5～8 に示す。

表 6 解答例と抜き出し結果

問題 1 の解答例	質問 1 に関する抜き出し
仮想メモリの目的は限られた大きさの主記憶装置をあたかも大容量化したように使用することである。また、実現のための基本的な考え方は、2 レベルのマッピング方式で TLB を導入することでより高速な仮想メモリを実現することができる。	2 レベルのマッピング方式で TLB を導入することでより高速な仮想メモリを実現することができる
仮想メモリの目的は、主記憶容量の不足を補うことである。実現のための基本的な考え方は、補助記憶装置の一部も主記憶装置に見せかけて使用することである。	補助記憶装置の一部も主記憶装置に見せかけて使用すること
仮想メモリは主記憶に入りきらないプログラムやデータを格納する目的で、主記憶の容量を補助記憶の領域を用いることで低速ながら大容量化を図るもの。キャッシュメモリと同様、物理メモリのアドレスと仮想メモリのアドレスはマッピングされ、主記憶と仮想メモリ間でデータが交換される。	主記憶に入りきらないプログラムやデータを格納する目的
ノイマン型では主記憶上にプログラムをおいて実行するため、複数のプログラムを同時に実行させる場合、主記憶容量が不足することがある。仮想メモリはこの主記憶の容量不足を解消するために、補助記憶装置の一部を主記憶のように利用する方法として使用される。	補助記憶装置の一部を主記憶のように利用する方法
目的主記憶の増加を促す。基本的な考え方ノイマン型アーキテクチャーでは、多数のプログラムを同時に実行させる場合、主記憶容量が不足することがある。そこで、補助記憶装置の一部も主記憶に見せかけて使用する方法として、仮想メモリ方式が導入された。	*
仮想メモリには主記憶装置の記憶容量を仮想的に増大させるという目的がある。そして、ページング方式、セグメンテーション方式、マッピング方式といった 3 つの分割方式の考え方で構成されている。	3 つの分割方式

* は抜き出し箇所がないと判断されたもの

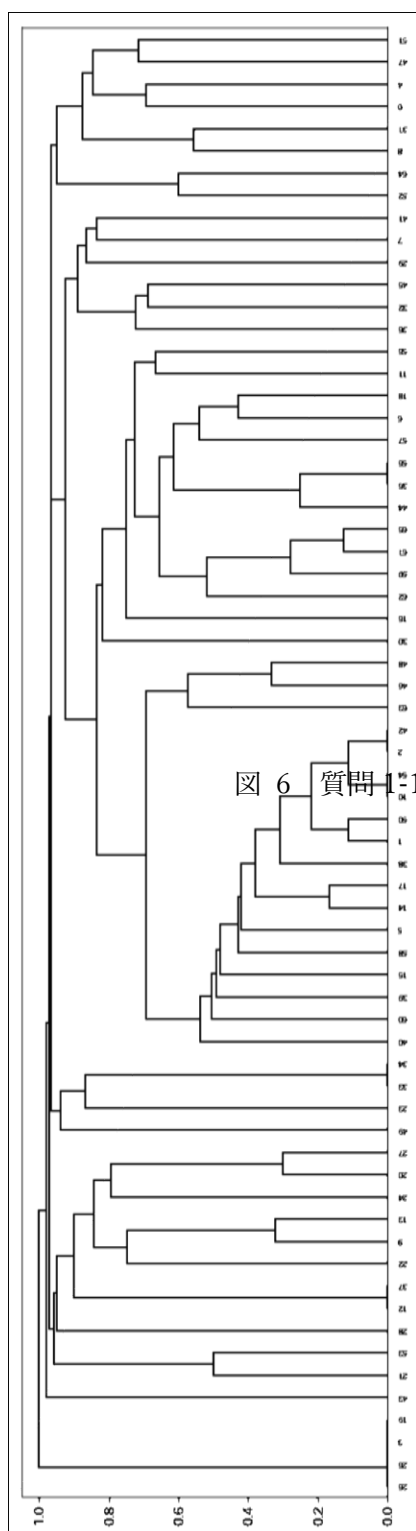


図 6 質問11の抜き出し結果の樹形図

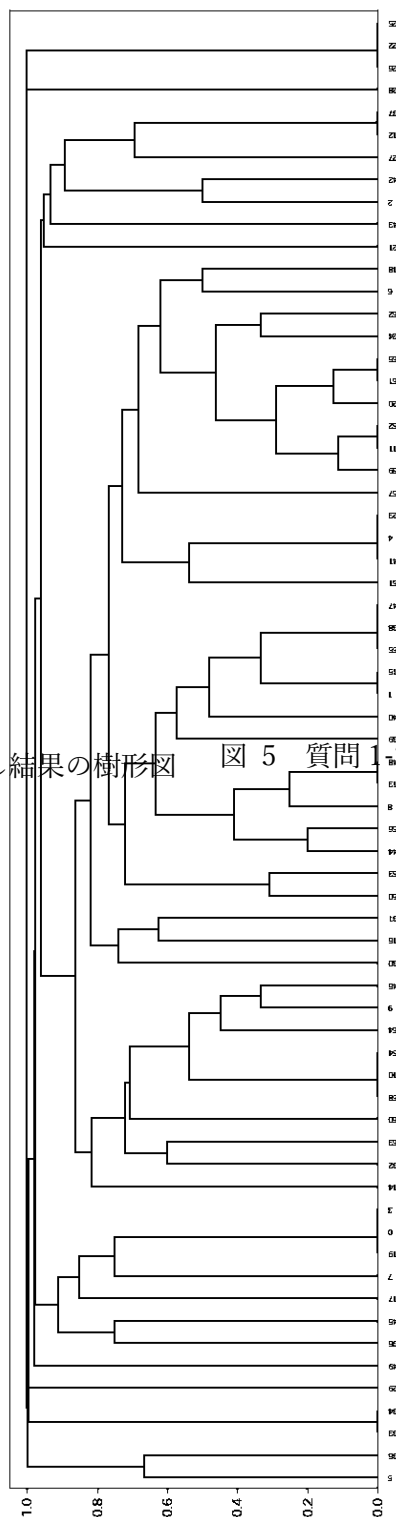


図 5 質問12の抜き出し結果の樹形図

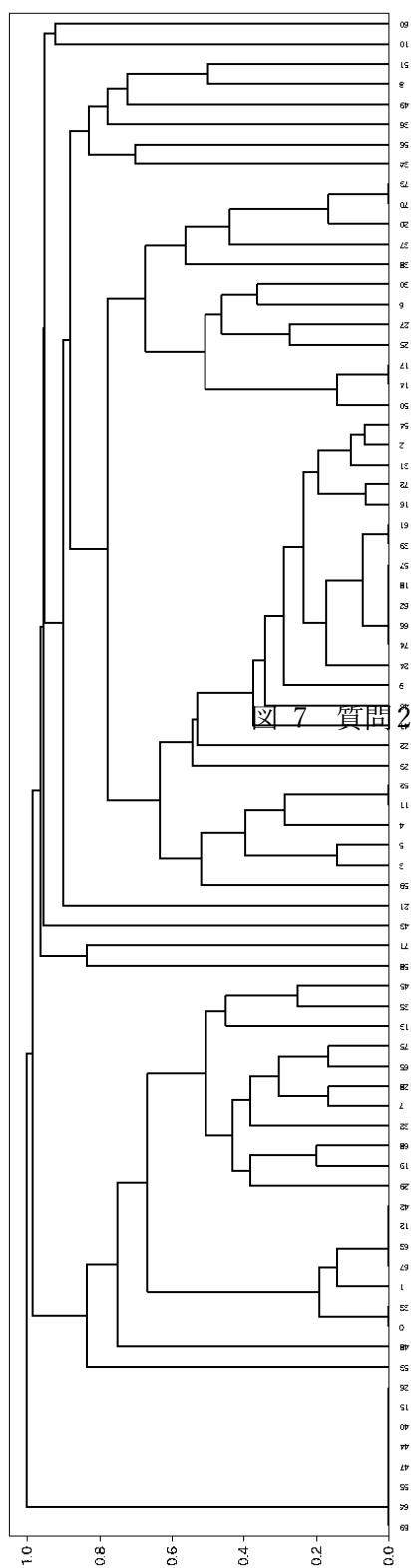


図 7 質問 2-1 の抜き出し結果の樹形図

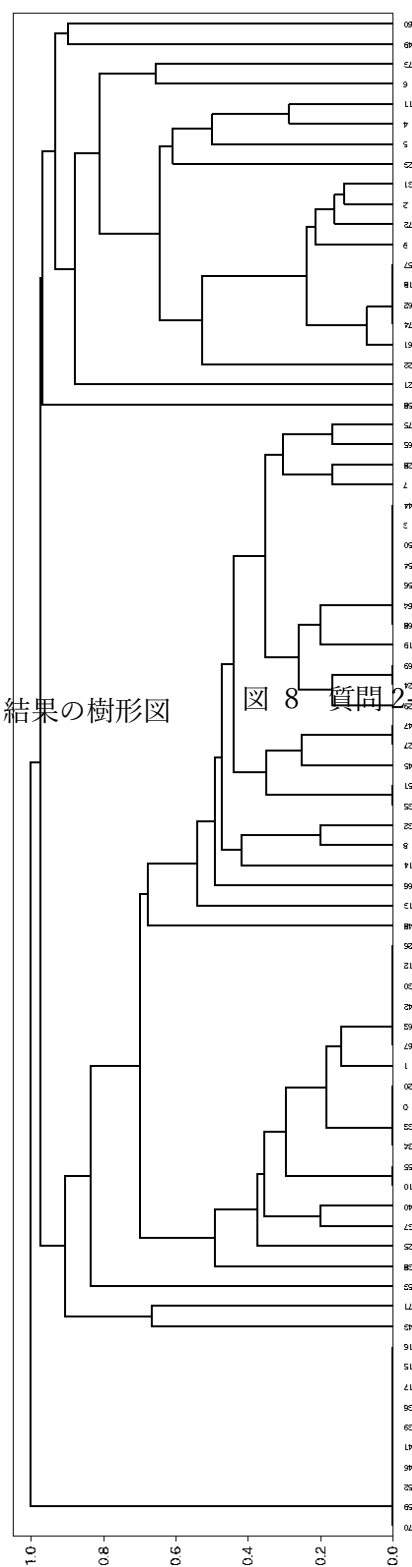


図 8 質問 2-2 の抜き出し結果の樹形図

4.2 LLM による該当部分の抜き出し精度の確認

大規模言語モデルによる質問に対する該当部分の抜き出し精度の確認のために抜き出した記述、解答全文、質問を他年度に同様の講義を受けたことがある被験者 3 名それぞれに提示した。その後、解答全文を見て自身が判断した質問に関わる箇所（質問箇所）と抜き出した記述についての関係を下記の 5 つをもとに選択してもらった。

- ① 質問箇所と抜き出し箇所に 1/3 以上の過剰があれば「抜き出し過剰」
- ② 質問箇所と抜き出し箇所が一致すれば「過不足なく抜き出し」
- ③ 質問箇所と抜き出し箇所に 1/3 以上の不足があれば「抜き出し不足」
- ④ 質問箇所に関わらない抜き出し箇所は「抜き出せていない」
- ⑤ 質問に関わる記述がなければ「質問内容がない」

表 7～10 にそれぞれの質問ごとの抜き出し結果と表 11 にその平均を示す。②の過不足なく抜き出せている列に注目すると 4 割弱から 6 割の抜き出し精度となっている。これはほとんどの問題において割合が最も高くなっている。しかし表 5 に示した JSQuAD における精度では完全一致が 9 割であり 3～5 割ほど下回っている結果になった。これは、実際の解答の中に質問が明示的に入っていないことや、表 7、表 8 の⑤のようにそもそも質問内容が解答に含まれておらず、無関係な箇所を抜き出してしまうことが影響していると考えられる。

①②③を合計した少しでも抜き出しに成功しているものは 5 割から 7 割となっている。①は全問題を通して割合が少なく、③は割合が多い問題と少ない問題に分かれた。表 9 と表 10 を比較すると同じ問でも「メリットは?」「デメリットは?」という質問の違いで②と③に大きな差ができています。これは質問 2-1 に関する記述を二つ書いている生徒が多く、片方のみの抜き出しをしていることが多かったためであった。①には影響がないことも含めて二つ以上の内容を含む記述の抜き出しでは過剰抽出ではなく欠落が生じる可能性が高いことが確認できた。

④⑤の的外れな記述を抜き出している割合は 3 割～5 割となった。④については質問 1-1、1-2 では質問が明示的に解答分の中に入っていないことが多く、異なった記述を抜き出していた。質問 2-1、2-2 では利点とメリット、欠点とデメリットなど類義語を含む抜き出しなどができていたが、メリットとデメリットのような似た用途の意味が異なるものは混同してしまうケースが多く、抜き出しできていないが多くなった。デメリットの質問でメリットを抜き出すよりもメリットの質問でデメリットを抜き出すことが多かったことから文字からすくなくならず影響を受けると考えられる。このことから対義語が同一の解答内で記述されるケースでは抜き出しがうまくいかないことも確認できた。

結果として半数以上は質問に関わる抜き出しとうまくいかない場合の確認ができた。抜

き出し結果のクラスタリングへの影響は次節において検討する。

表 7 質問 1-1 の抜き出し結果 (単位:%)

	①	②	③	④	⑤
被験者 1	4.7	48.4	7.8	25.0	14.1
被験者 2	1.6	35.9	17.2	17.2	28.1
被験者 3	9.4	43.8	20.3	18.8	7.8

表 8 質問 1-2 の抜き出し結果 (単位:%)

	①	②	③	④	⑤
被験者 1	3.1	50.0	4.7	37.5	4.7
被験者 2	3.1	45.3	3.1	21.9	26.6
被験者 3	9.4	43.8	3.1	34.4	9.4

表 9 質問 2-1 の抜き出し結果 (単位:%)

	①	②	③	④	⑤
被験者 1	1.3	40.8	19.7	38.2	0.0
被験者 2	5.3	38.2	15.8	39.5	1.3
被験者 3	3.9	36.8	21.1	38.2	0.0

表 10 質問 2-2 の抜き出し結果 (単位:%)

	①	②	③	④	⑤
被験者 1	3.9	61.8	0.0	34.2	0.0
被験者 2	3.9	59.2	1.3	34.2	1.3
被験者 3	5.3	59.2	0.0	34.2	1.3

表 11 抜き出し結果の平均 (単位:%)

	①	②	③	④	⑤
質問 1-1	5.2	42.7	15.1	20.3	16.7
質問 1-2	5.2	46.4	3.6	31.3	13.5
質問 2-1	3.5	38.6	18.9	38.6	0.4
質問 2-2	4.4	60.1	0.4	34.2	0.9

4.3 抜き出し箇所のクラスタリング精度の確認

抜き出した記述をクラスタリングして作成されたクラスタを確認することで主要な記述を把握できるか確認をするため、同様の被験者 3 名にそれぞれの問題の解答群から特定の記述を含む解答が何割であるかを回答してもらった。特定の記述は筆者が質問に関する解答の中で主要（1, 2 番目に頻出）だと判断したものであり、表 12 に示す。回答してもらった結果を人による概数把握の基準とする。4.1 で作成した樹形図から一定の閾値をもとにクラスタの作成をおこなう。閾値は樹形図をもとに一律 0.6 とした。このクラスタ内の長さが中央値である記述そのクラスタの代表の記述とし、記述内容とその数を人による概数把握と比較する。代表記述が同じ内容のクラスタがあればそれらを足した数を概数とする。また、クラスタ内に異なる内容の記述がどれだけあるかの確認を行った。

表 13 にクラスタから得られた記述の割合と被験者によって確認できた記述の解答の平均割合を示す。表 14 にはクラスタ内の異なる内容の記述数を示す。表 13 をみると最も人に近いもので人手の 6 割、できていないもので 2.5 割であった。平均すると人手の 4.3 割がクラスタリングによって確認できた。その中で表 14 に示すクラスタ内の異なる記述をみると 3, 4 を除くとほとんど同義の記述をクラスタリングできていることが確認できた。クラスタリングから把握できた記述の割合は前節の抜き出し結果とともに表 13 を確認すると、誤って抜き出している④と⑤の記述に関してはクラスタリングにおいても確認したい記述にはならないためその影響によりクラスタリングによって得られる記述の割合が減少したと考えられる。記述 3, 4 に関してはクラスタ内の中央値の長さの記述を代表例としたため、目的の記述を含むクラスタ内にあった「主記憶装置」「主記憶容量の不足」といった抜き出し不足の記述が同じクラスタとなり、確認したい記述と異なる記述が混在してしまった。そこで抜き出し不足の記述のもとの解答を見ると目的の記述が書かれていることが多く、抜き出しが成功していれば同義な記述に数えられることが分かった。また、色のついてある記述 1 を除いたすべての記述ではクラスタによって把握できる最大数の記述とその次の記述であることが確認できた。質問 1-1 では解答では「マッピングを行うこと」という記述が二番目に頻出であったが、クラスタリングでは質問に関係ない「主記憶装置の増大化」という記述に関するものが二番目になっていた。これは④の誤って抜き出した記述に似ているものが多く含まれており、クラスタリングされたものである。これは抜き出し精度の向上によって解決されることが考えられる。上記から、クラスタリングによって記述の正確な割合の計測は抜き出し精度の向上が必要だが、主要な記述の内容とその多寡については把握できるといえる。

表 12 把握に使用した記述

問 1-1	「マッピングを行うこと」 「補助記憶装置の一部を主記憶に見せかける」
問 1-2	「主記憶装置を仮想的に増大させる」 「主記憶容量の不足を補う」
問 2-1	「フォン・ノイマンのボトルネックを回避できる」 「データのビット幅をそろえる必要がない」
問 2-2	「回路が複雑になる」 「ハードウェアが複雑になる」

表 13 確認できた記述の割合 (単位:割)

	クラスタリング	人手
1「マッピングを行うこと」	0.6	2.3
2「補助記憶装置の一部を主記憶に見せかける」	2.3	4.7
3「主記憶装置を仮想的に増大させる」	1.9	5.3
4「主記憶容量の不足を補う」	1.1	3.3
5「フォン・ノイマンのボトルネックを回避できる」	4.0	7.7
6「データのビット幅をそろえる必要がない」	3.2	6.3
7「回路が複雑になる」	3.3	7.0
8「ハードウェアが複雑になる」	2.2	4.3

表 14 クラスタ内の異なる内容の記述数 (単位:個数)

	同義な記述	異なる記述
1「マッピングを行うこと」	4	0
2「補助記憶装置の一部を主記憶に見せかける」	15	0
3「主記憶装置を仮想的に増大させる」	9	3
4「主記憶容量の不足を補う」	4	3
5「フォン・ノイマンのボトルネックを回避できる」	30	0
6「データのビット幅をそろえる必要がない」	24	0
7「回路が複雑になる」	24	1
8「ハードウェアが複雑になる」	15	2

第5章 まとめ

本研究では多人数の講義における記述式演習において講師がフィードバックを行うのに必要な概要とその解答数をインタラクティブに素早く把握できることをめざす。そのための手法として講師の質問をもとにした大規模言語モデルによる記述の抜き出しとその記述の階層型クラスタリングを提案した。

実験では実際の解答を使用して抜き出しの精度判定とクラスタリングの確認を行った。抜き出し精度は4.2節で述べたように一般のQAタスクより低くなっていたが、4.3節の実験のクラスタリング結果から主要な記述の多寡を把握することができることがわかった。少数の記述は把握できないが、主要となる記述に関しては把握できると考えられる。フィードバックに必要である記述が主要なものであればインタラクティブな概要把握が可能となったといえる。

今後の課題としては抜き出し結果の影響により正確な記述の割合の把握ができていないと考えられるため、抜き出し精度の向上があげられる。抜き出し精度の向上のため出題の際の使用語句の指定や出題形式の選定、モデルのファインチューニング方法や学習法の選定などを検討する必要がある。ほかにもクラスタリング結果を講師が見やすいように見せるためのインターフェイスの開発や、自動的なクラスタ数決定の手法の検討があげられる。

謝辞

本論文は、著者が三重大学工学研究科電気電子工学専攻に在学中に行った研究をまとめたものです。本研究を進めるにあたり、懇切丁寧なご指導とご督励を賜った本学大学院工学研究科電気電子工学専攻の高瀬治彦教授，川中普晴教授，北英彦准教授に深く感謝いたします。また，日頃熱心に討論して頂いた計算機工学研究室の皆様方に厚く御礼申し上げます。最後となりましたが，本論文をまとめるにあたり，御助言，御討論していただいた方，その他お世話になりましたすべてのの方々に感謝致します。

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. 宮崎佳典, 相馬あおい, 厨子光政, 法月健：“英単語並べ替え問題における機械学習による学習者の迷い検出の試み”，Computer & Education, vol. 45, pp. 31-36 (2018)
3. 児島千珠代：“学習意欲と授業についての考察”，清泉女子大学紀要, vol. 64, pp.33-48 (2017)
4. 溝上慎一：“大学生の学習意欲”，京都大学高等教育研究, vol. 2, pp. 184-197 (1996)
5. 松田健, 大谷康介, 中野美知子：“学習履歴取得システム開発に関する一考察”，情報処理学会, vol. 79, pp. 469-470 (2017)
6. 藤井利江, 山口裕幸：“大学生の授業中の質問行動に関する研究-学生はなぜ授業中に質問しないのか？-”，九州大学心理学研究, vol. 4, pp. 135-148 (2003)
7. 西村奏咲, 清水忠：“テキストマイニングを用いたアンケート解析”，薬学教育 5, (2021)
8. 高瀬治彦, 川中普晴, 鶴岡信二, 森田直樹：“記述式小テストの解答群の分析手法-解答群からのキーワード自動抽出-”，コンピュータ & エデュケーション, vol. 34, pp.46-49 (2013)
9. Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever. “Improving Language Understanding by Generative Pre-Training”, Technical report OpenAI, (2018)
10. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp.4171-4186 (2019)

11. 栗原健太郎, 河原大輔, 柴田知秀: ”JGLUE:日本語言語理解ベンチマーク”, 自然言語処理学会, vol. 30 No. 1, pp.63-87 (2023)
12. 元田浩, 津本周作, 山口高平, 沼尾正行: “データマイニングの基礎”, オーム社, (2006)
13. “Language Media Processing Lab at Kyoto University” <https://huggingface.co/ku-nlp/deberta-v2-base-japanese/tree/main> (2024 年 1 月 22 日取得)