

大学院生プロジェクト型研究・研究成果報告書

研究代表者：三浦直己（教育情報アセスメントコース）

■ 研究題目
個別最適な英語学習の提供に向けた研究
■ 研究代表者・分担者（氏名、コース）
三浦 直己（教育情報アセスメントコース・博士課程前期 1 年）（代表者） 舟山 弘晃（東北大学大学院情報科学研究科・博士課程後期 3 年） 岩瀬 裕哉（教育情報アセスメントコース・博士課程前期 2 年）
■ 研究成果概要（目的、実施内容、結果、今後の課題など）
<h2>1. はじめに（問題と目的）</h2> <h3>1.1 研究背景</h3> <p>GIGA スクール構想により、ICT を活用した個別最適な学びの提供が促進されている。中学・高校の外国語（英語）教育でも、1 人 1 台の端末を利用した英語学習が進んでおり、学年や進度に合わせた授業実践が試行錯誤されている [1]。一方で、英語学習では、生徒が直面する課題やつまづきは多岐にわたり、各人のニーズを丁寧に拾い上げ対応していくのが喫緊の課題であると指摘されている [2]。</p> <p>こうした背景のもと、教員が各生徒の習熟状況を詳細に把握し、指導を行う「個別最適な英語学習」の実現が目指されている。しかし、現実の教育現場においては、教員の負担が大きく、「学習の個別化」を十分に実現することが困難な状況にある。その中でも特に、記述式答案の評価は、生徒のつまづきを個別に把握し、適切なフィードバックを提供する上で有効な手法であるが、その採点業務には多大な労力を要する。</p> <p>例えば、和文英訳問題は日・英言語間の類似点や相違点の認識を促し、相応しい表現方法への理解を深めることができる。そのため、言語学習の初期段階において学習者が基本的な文法や表現を習得するのに特に効果的である [3]。しかし、和文英訳はその問題の性質上、記述式回答となるため、答案の採点やフィードバックの返却は教育者にとって大きな負担となる。結果として、言語習得において反復練習が重要であるにも関わらず [4]、実施の頻度が限定されてしまう。</p>

このような状況から、先行研究でも和文英訳問題の答案の自動採点モデルの構築が行われている。既存手法 [5] では、当時最も汎用的に使われていた言語モデルである BERT[6] をベースとした採点手法を提案した。しかしながら [5] では、問題ごとに 100 件程度の採点済みデータを用意する必要があるため、モデル構築の面で非常にコストがかかるという課題がある。この課題を受けて、[7] では、その答案の自動採点に大規模言語モデル (Large Language Model, LLM) を使用する手法を検討し、その性能を評価した。[7] は、LLM 内部のパラメーター更新を行わずに、LLM への入力文中に例示や指示を与えて即時的にタスクに対応させる文脈内学習 [8] という低コストな手法を試したが、結果として、既存手法 [5] と同等の採点性能を示すには至らなかった。さらに、エラー分析の結果、LLM においても既存手法 [5] 同様に、モデル内部のパラメーターを更新する手法である追加学習 (finetuning) を行う必要性が示唆された。

以上の背景を踏まえ、本研究では LLM を追加学習させることで採点性能の向上を目指す。また、教育現場での導入を考慮し、追加学習に用いるデータの分布や量を調整することで、提案手法が既存手法 [5] よりも低資源環境において高い採点性能を実現できるかを検討する。

1.2 先行研究

記述式の英語答案の自動評価という文脈では、記述式答案自動採点 (SAS) と文法誤り訂正 (GEC) の 2 つが主要な研究分野である。記述式和文英訳問題の自動評価の先行研究として [5] が挙げられる。この研究は、記述式和文英訳問題の自動採点を文法項目などの各採点項目に基づき個別に採点するタスクであると定義し、和文英訳問題データセットを構築した上で、言語モデルである BERT[6] をベースとした採点モデルを作成した。この採点モデルは、SAS のために [9] によって提案された採点モデルを、文法事項などの各採点項目から個別採点するように拡張したものである。この採点モデルは高い性能を出している一方で、モデル構築のために 100 件程度の採点例を要するほか、採点項目ごとにモデルを構築する必要があり、コストの高さが課題である。また、モデル構築時に使用したデータの中に含まれなかった事例には適切に対応できない場合があるという汎用性の面でも課題がある。先述の通り、この課題に対し [7] では、LLM を活用して低コストの文脈内学習による手法を提案したが、[5] を上回る採点性能は示されなかった。

最も有名な GEC システムである Grammarly1 は英語学習において重要な役割を果たしている。ただ、GEC システムの純粋に使用するだけでは、学習者の取り組みを効果的に引き出すことができないとの報告もある [10; 11]。そのため、より教育的な文脈では [12] が第二言語としての英語学習に焦点をあて、GEC によるフィードバック生成タスクを提案した。[13; 14; 15] では、学習者の書いた英文の文法エラーに対してフィードバックを生成する方法を開発している。また近年、第二言語学習の分野において、DEEP

GRADE2 や 桐原 AI エッセイライティング 3 など、主要な LLM である GPT-3.5 を活用した記述式答案の文法誤り訂正 (GEC) および内容評価を行う自動採点システムがリリースされている。しかしながら、これらのシステムは、あらかじめ用意された問題を採点するシステムであり、実現場で教師が任意に作成した問題の採点に流用することは困難であるという課題がある。

問題:
次の和文を英訳せよ

私は 一昨年に オーストラリアで 見るまで
コアラを 見た ことがなかった

学習者の回答

I hadn't seen a koala , before I saw in Australia two years ago.

O4 G4 E3

採点基準

採点項目	基準	2 (正解)	0 (不正解)
“オーストラリアで”	E3	“in Australia”と表現している	その他

“見るまで”	O4	語順が “接続詞+SVO”	語順が 不適當
	G4	“saw”を 使っている	その他

図 1 : 和文英訳問題の採点例

1.3 研究の目的

本研究では、教育現場における課題と先行研究の自動採点モデルの現状を踏まえ、教員が作成した任意の問題に対しても、低コストで採点を自動化する手法を模索する。そこで、第一歩として、LLM を活用し、図 1 に示したような部分採点項目の自動採点技術を構築する。採点モデル構築時点において、モデルの追加学習に用いるデータの分布や量を調整することで、提案手法が低資源環境において高い採点性能を実現できるかを検討する。

実際の教育現場において、和文英訳問題の採点業務の自動化により、教員の採点業務の負担を軽減・生徒の反復学習を効率化し、各生徒に個別最適な学習を提供する一助となることが期待される。また LLM を採点タスク用に追加学習させることで、[7] での性能面の課題を克服しつつ、1.2 節で挙げた [5] のコスト面と汎用性の面での課題に対応し、少量の採点済みデータのみで、汎用的に採点することが期待できる。

2. 和文英訳問題の自動採点

2.1 和文英訳問題とは

英語学習では様々なタイプの問題が扱われるが、本研究では、記述式の和文英訳問題を題材として扱う。和文英訳問題とは、与えられた問題文（和文）に対して生徒が回答（英訳）する問題形式である。図1に和文英訳問題の答案の採点例を示す。時制や語順などの文法に関する採点項目や、和文に対応する適切な単語を使用しているかといった語彙力に関する採点項目があり、これらの採点項目に基づき、回答は文法的な正確性と元の和文の意味を十分に含意しているかという基準で採点が行われる。図1ではEに当たるものが表現、Oが語順、Gが文法に関する採点項目である。各項目に対する基準から、項目ごとに答案が評価され、今回の学習者の回答では過去時制の使用などで正解の評価がなされるが、主語動詞の後に目的語が欠如しているため、語順の採点項目で不正解の評価がされている。

このように、和文英訳問題では、特定の文法知識や語彙知識を問いやすいため、教員が生徒の英語力を文法や語彙などの観点ごとに、詳細に把握することができる。したがって、この出題形式が個別最適化された学習を提供するというニーズに直結しているため、本研究では和文英訳問題に焦点をおいて研究を行う。

2.2 英語学習における和文英訳問題の利点

この節では、第二言語習得の観点から、英語学習における和文英訳問題の利点について論じる。文章の翻訳問題は、第二言語学習の初期段階、特に言語学的に離れた言語ペア間の教育ツールとしてよく利用される [3; 16]。第二言語学習における母語の使用や活用に関しては多くの議論がなされており、その効果に関する研究もある。

[3; 16] は、第二言語学習での母語を使用が、言語間の相違点と類似点の理解を促進すると示した。また [17] では、母語の使用が、学習する第二言語についての理解不能を減らし、学習の集中力を高めることが示されている。さらに [18] では、言語翻訳が生徒の4技能（話す、書く、読む、聞く）を向上させ、学習能力やコミュニケーション能力を促進する効果的な役割を果たすことが示している。

また、和文英訳問題には、学習者が特定の英語表現を用いるように、教育者の意図を反映できるという問題形式上の利点がある。第二言語学習上の課題として、例えば日本語のように英語と言語的に大きく異なる言語を母語に持つ学習者は、英語記述において母語には存在しない特定の表現を回避する傾向があることが明らかにされている [19]。

そのため、特定の文法表現の習得を学習目標とする教育現場において、非常に有効な問題形式と言える。このような利点から、文章の翻訳問題は教育現場、特に言語学習初心者の中で広く利用されている。そのため和文英訳問題は、学習者が日本語（母語）と英語（学

習言語)の類似点や相違点を認識しやすくなり、望ましい表現方法に関して理解を深めることができる。以上のことから、和文英訳問題は言語学習の初期段階で学習者が基本的な文法や表現を習得するのに効果的な問題形式である。

2.3 先行研究手法の課題

和文英訳問題の自動採点を扱った先行研究である [5] では、問題ごとに事前に 100 件程度の採点済みデータを使用した教師あり学習による手法が用いられていた。教師あり学習とは、機械学習の手法の一つであり、事前に付与されたデータを正解データとみなし、分類問題や回帰問題を学習するものである。この手法による採点モデルは正解答案に対して高い採点精度を出す一方で、いくつかの課題がある。

まずは前述の通り、採点済みデータを用意するコストの高さと汎用性の低さである。正解データとして採点済みの答案データを大量に必要とするほか、新しい問題(和文英訳問題)を扱うたびに同様の作業を行わなければならない。自動採点モデルの教育現場での実運用を意識した際に、新しい問題を扱うたびに大量の採点済みデータを事前に用意することは非現実的である。さらに一度構築した採点モデルを別の問題の採点にそのまま流用することができないことも、実運用を妨げる要因である。

また [5] では、不正解答案の採点精度に関しても課題があったと言及されている。モデル構築に用いたデータ内に含まれなかった不正解パターンが入力された際に、採点精度が低下してしまうことが問題であった。[5] ではこの課題に対し、収集したデータに加えて誤りを含んだデータを人為的に作成し、モデル構築に用いることで採点精度を向上させていた。しかしながら、不正解のパターンを網羅的にモデルに学習させることは非常に難しく、改善策としては限界がある。

さらに [7] では、GPT-3.5 や GPT-4 の文脈内学習によって低コストな採点モデルの構築を目指したが、LLM 内部のパラメータ更新を行わないため、採点自体を理解していないような出力が目立ち、良い性能を示すことができなかった。

以上が先行研究の課題である。本研究では、LLM を省資源設定で追加学習し、LLM の内部パラメータを採点用に微調整することでこれらの課題を解消し、自動採点モデルの実運用をより現実的にすることを目指す。

3. 提案手法(方法)

和文英訳問題の自動採点手法として、本研究では GPT モデル [20] を追加学習させたモデルを用いる。また提案手法に対するベースラインとして、既存手法 [5] で用いられる BERT ベースのモデルを採用する。BERT は gpt モデルリリース以前に、最も活用され

ていた言語モデルであり、内部パラメータ量は gpt-3.5 の数千分の 1 である。実験概要を図 2 に示す。2 種のモデルの追加学習に用いるデータの量や分布を調節することで、その性能差を比較し、提案手法が省資源設定下でも活用可能かを明らかにする。

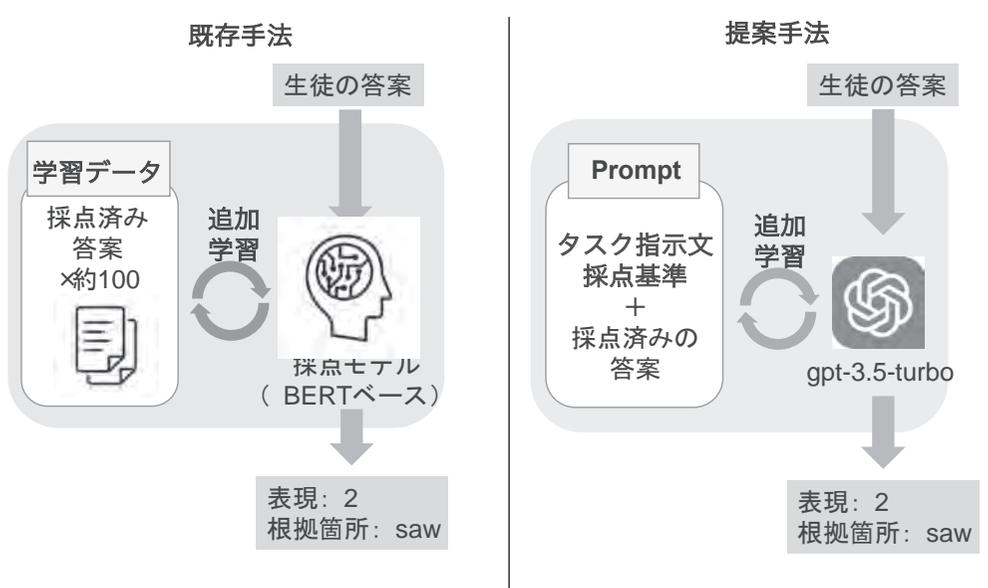


図 2 : 実験手法の概要 (左が既存手法、右が提案手法)

3.1 追加学習された GPT モデル

提案手法として、gpt-3.5 の追加学習によって採点を行うモデルを構築する。この方法により、採点項目ごとに採点モデルを構築するコストと、モデルの微調整に必要な訓練データの大幅な削減が期待できる。さらに、GPT モデルは翻訳や要約のようなタスクで優れた性能を示していることから [21; 22]、和文英訳問題の採点に必要な文法や語彙知識を潜在的に備えていることも期待される。

gpt-3.5 の追加学習は OpenAI の追加学習 (Fine-tuning) フォーマット 4 に従い、採点済みの答案と採点基準、問題文 (和文) を用いて行う。追加学習に用いた json 形式のデータは、図 3 のようになる。まず、事前入力の部分では、gpt-4o を用いて、図 1 に示すような人手で作成された採点基準をモデルにとって解釈可能な形式に書き下した。また、学習に用いる採点事例をその後に入力し、各採点アイテムのスコアに該当する答案内の単語列とスコアを順番に出力させるよう設計した。

3.2 BERT ベースモデル

和文英訳答案自動採点のタスクにおけるベースラインとして、BERT[6] ベースの採点

モデルである先行手法 [5] の項目採点モデルを用いた。実験では、モデルのエンコーダーに BERT を導入し、学習用の採点済み答案データを用いてモデル内部のパラメータを更新する教師あり学習による手法で採点を行う。モデルの微調整では、最適化関数に Adam[23] を使用し、学習率は 0.001 に設定した。Bi-LSTM の隠れ状態の次元は 128 に設定し、学習に用いるデータ量が限られているためバッチサイズは 10 とした。

事前入力
[採点タスクの説明] You are the teacher and grader for the sentence translation.. You will be provided with a .. rubric and student answer Please grade the answer according to the rubric.... ## Instructions 1. Identify the span of the sentence.. 2. Then, provide the Grade based..
[出力形式の指定] ## Output format Justification Cue: <Your Output here> Grade: <Your Output here>
[問題文(和文)]
[採点基準: 各スコアに対する採点基準とフレーズの例 (図1参照) を gpt-4o で書き下したもの]
GPTモデルの学習用入力 (生徒の答案)
GPTモデルの学習用出力 (人手の採点結果)
Justification Cue: ~ Grade: ~

図 3 : gpt-3.5 の追加学習用データ略図

4. 実験

和文英訳答案の自動採点において、省資源下でのモデルの採点性能を評価するため、実験では BERT モデルと追加学習した GPT-3.5 を使用して答案に対する得点予測を行った。得られた予測得点と人間による採点結果との F1 スコアを用いて評価した。

4.1 実験設定

今回の実験では、[5; 7] によって作成された和文英訳問題データセットを評価データとして使用する。先行研究 [5] に従い、モデルの性能を F1 スコアで評価した。各問題のデータセットを訓練セット、開発セット、評価セットに 3:1:1 の比率で分割し、5 分割交差検証を行った。BERT モデルは各訓練セットにおいて 50 エポックでモデルの微調

整を行い、開発セットにおいて最も良い性能を示した時のパラメータを用いた。また、一部の採点基準は、学習者にとって難易度が低く、不正解の答案が非常に少なかった。そのため、採点モデルの性能を適切に評価するために、10%以上の不正解の答案事例数を持つ採点項目だけを評価に使用した。

4.2 実験 A (均衡データにおける実験)

省資源設定でも高い性能を示すことを期待し、gpt・BERTの両モデルの追加学習に使用するデータに均衡データを使用した。均衡データとは、学習データ内に含まれる答案の分布を均等にしたものである。例えばある採点項目に対して、全200件の答案のうち、180件が正解答案で20件が不正解答案であるとする、20件の学習データを作る際には、正当不正から10件ずつサンプリングしてデータを整形する。このような設定にすることで、既存研究より大幅に少ない20件というデータ量でも、高い採点性能を示すことが期待される。なお、OpenAI社の追加学習回数制限等の影響で、評価データには和文英訳問題データセットの全21問中9問のみの使用に止まっている。

4.2 実験 B (不均衡データにおける実験)

実験 A に対して、現場実装を意識した、より低リソース下での性能評価を試みる。不均衡データは、オリジナルのデータセットの答案分布を保ったままサンプリングして整形されるデータのことである。先ほどと同様に、全200件の答案のうち、180件が正解答案で20件が不正解答案であるとする、9:1というデータ分布を維持して、20件の学習データをサンプリングする場合、正解答案を18件、不正解答案を2件含んだデータとなる。自動採点のタスクでは、正解答案は殆ど一意に定まる一方で、不正解答案のバリエーションは無限に想定される。そのため、不正解答案が学習データに十分に含まれないのは、モデルにとって不利な設定である。しかしながら、教育現場を考えると、教員が任意に作成した問題に対して、何種類も誤答のパターンを想定し入力することは現実的ではない。そのため、あえてこのような極端に厳しい設定でも試行し、提案手法の性能を測る。評価データは実験 A と同様に9問を用いた。

5. 結果

実験 A の結果を図4に示す。縦軸がモデルの評価値であるF1スコアであり、横軸が均衡データの事例数である。右側の正解答案に対する採点性能は、正解答案を正しく正解と分類できているかという性能であり、左側が不正解答案に対する採点性能である。図中の黒い点線は[7]で示された、本研究の提案手法よりも低コストな手法である。

図3より、正解答案への採点性能は少ない学習データ量でも十分に高いことがわかる。これは、正解答案のバリエーションが数パターンしかないため、少ないデータ量でも正当

パターンが十分に含有できることを示す。その一方で、不正解答案への採点性能においては、省資源設定においては、BERT と gpt モデルで採点性能に明確な差が示された。gpt モデルは省資源設定である 20 事例や 30 事例での学習でも高い性能を示しているが、BERT モデルは省資源設定では十分な採点性能が示されなかった。この原因として、少量の学習データでは、あらゆる不正解答案のパターンを網羅できておらず、BERT モデルが正解不正解の境界を十分に学習することができなかつたことが示唆される。その一方で、gpt モデルはあらゆるタスクで高い性能を示しているように、今回の採点タスクにおいても、膨大な内部パラメーターを活用し、少量のデータでも高い採点性能を示した。

実験 B の結果は図 5 に示す。不均衡データはより現場実装を意識した極端に低リソースな設定であり、gpt モデルにとっても困難な設定であった。特に不正解答案への採点性能はベースラインである BERT モデルを下回る結果となった。原因として、不均衡データ内はマジョリティラベルである正解事例が大半を占めるため、その影響を過敏に受けてしまった。そのため特定ラベルだけを出力するように学習してしまうこともあり、性能が向上していかなかったことが示される。

以上より、gpt モデルに対し、学習データの答案分布の均衡を保つことで、省資源設定でも高い性能を示すことがわかった。一方で不均衡データでの実験より、学習データにおいて答案分布の均衡を保たなければ、マジョリティラベルの影響を大きく受け、採点性能に悪影響を及ぼすことがわかった。

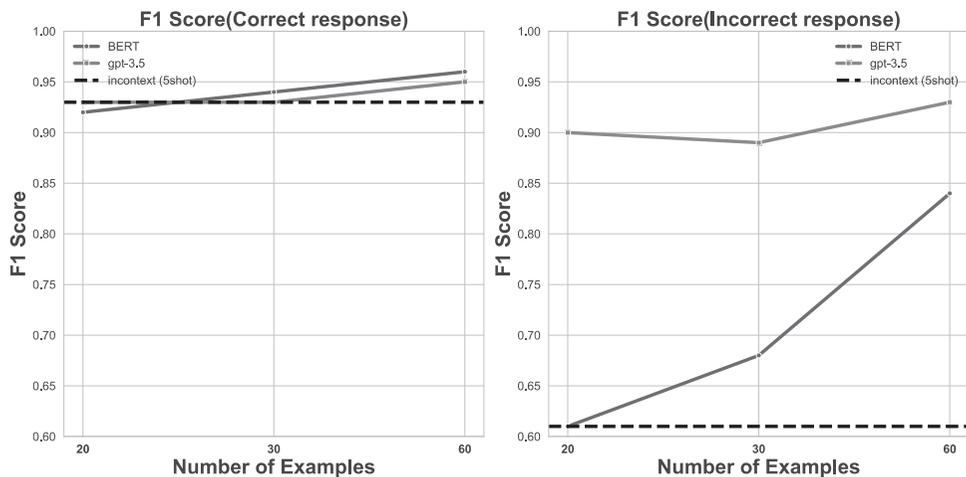


図 4：均衡データの事例数を増やすことによる gpt-3.5 の採点性能の変化

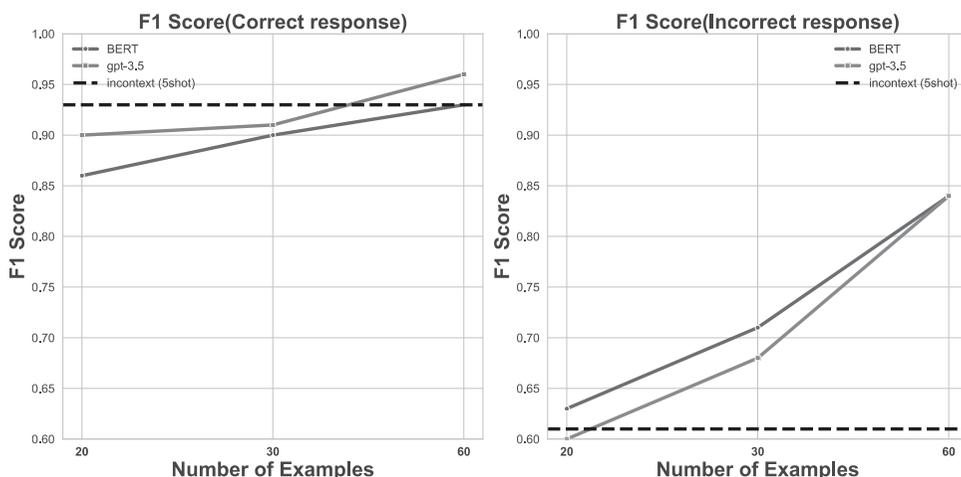


図 5 : 不均衡データの事例数を増やすことによる gpt-3.5 の採点性能の変化

6. 今後の方針

本研究では、従来手法よりも学習データ量が省資源設定下において、均衡を保った学習データならば gpt モデルが非常に高い性能を示すことを明らかにした。その一方で元のデータセットの答案分布を反映させた不均衡な学習データで構築したモデルは、マジョリティラベルの影響を大きく受け、望ましい性能が得られなかった。本研究が示唆するところとして、gpt モデルを追加学習させることで非常に高い性能を示す一方で、より現実的な低リソース設定では、gpt モデルにとっても採点タスクは挑戦的なタスクである。

現場実装を意識する上で、教員がその場で任意に作成した問題に対して、不正解答を複数個列挙することは労力がかかることである。今後の方針としては、学習データの不均衡を補填するべく、不正解答を擬似的に生成する方法などを試行し、モデルの学習データの均衡を保った上で構築することなどが挙げられる。

引用文献

- [1] 和田憲明. Giga スクール構想と英語教育における ict の活用. 姫路大学教育学部紀要/姫路大学教育学部紀要編集委員会 編, No. 15, pp. 47-52, 2022.
- [2] 嵐谷恭子, 川谷のり子, 鎌田真由美, 三成拓重, 大谷みどり, 猫田英伸. 中学校英語科における学びの多様性に応じた個別最適な授業づくり. 中国地区英語教育学会誌, Vol. 52, pp. 39-51, 2022.
- [3] Guy Cook. *Translation in Language Teaching: An Argument for Reassessment*. Oxford University Press, Oxford, March 2010.
- [4] Diane Larsen-Freeman. On the roles of repetition in language teaching and learning. *Applied Linguistics Review*, Vol. 3, No. 2, pp. 195–210, 2012.
- [5] 菊地正弥*, 尾中大介*, 舟山弘晃*, 松林優一郎, 乾健太郎. 項目採点技術に基づいた和文英訳答案の自動採点. 言語処理学会 第 27 回年次大会 発表論文集, p. 691, 2021.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [7] Naoki Miura, Hiroaki Funayama, Seiya Kikuchi, Yuichiroh Matsubayashi, Yuya Iwase, and Kentaro Inui. Japanese-english sentence translation exercises dataset for automatic grading. arXiv preprint arXiv:2403.03396, 2024.
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.
- [9] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. Analytic score prediction and justification identification in automated short answer scoring. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 316–325, Florence, Italy, August 2019. Association for Computational Linguistics.
- [10] Svetlana Koltovskaia. Student engagement with automated written corrective feedback (AWCF) provided by grammarly: A multiple case study. *Assessing Writing*, Vol. 44, p. 100450, April 2020.
- [11] Jim Ranalli. L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, Vol. 52, p. 100816, June 2021.
- [12] Ryo Nagata. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3206–3215, Hong Kong, China, November 2019. Association for Computational Linguistics.

- [13] Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. Exploring methods for generating feedback comments for writing learning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 9719–9730, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [14] Steven Coyne. Template-guided grammatical error feedback comment generation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 94–104, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.
- [15] Yi-Huei Lai and Jason Chang. TellMeWhy: Learning to explain corrective feedback for second language learners. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP):System Demonstrations, pp. 235–240, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [16] Wolfgang Butzkamm and John Caldwell. The Bilingual Reform. A Paradigm shift in Foreign Language Teaching. Narr Dr. Gunter, 01 2009.
- [17] Virginia Scott and Mar'ia De la Fuente. What's the problem? l2 learners' use of the l1 during consciousness - raising, form - focused tasks. The Modern Language Journal, Vol. 92, pp. 100 – 113, 03 2008.
- [18] Muhammet Yasar Yuzlu and Kenan Dikilitas. Translanguaging in the development of efl learners' foreign language skills in Turkish context. Innovation in Language Learning and Teaching, Vol. 16, No. 2, pp. 176–190, 2022.
- [19] Jacquelyn Schachter. An error in error analysis. Language Learning, Vol. 24, No. 2, pp. 205–214, 1974.
- [20] OpenAI. Gpt-4 technical report, 2023.
- [21] Serge Gladkoff, Gleb Erofeev, Lifeng Han, and Goran Nenadic. Predicting perfect quality segments in mt output with fine-tuned openai llm: Is it possible to capture editing distance patterns from historical data? arXiv preprint arXiv:2308.00158, 2023.
- [22] Abdulkader Helwan, Danielle Azar, and Dilber Uzun Ozsahin. Medical reports summarization using text-to-text transformer. In 2023 Advances in Science and Engineering Technology International Conferences (ASET), pp.01–04, 2023.
- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.