一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NLC2023-23(2024-03)

複数モデルの統合とデータ拡充による議論評価

橋口　駿亮†　　嶋田　和孝††

† 九州工業大学大学院情報工学府 情報創成工学専攻 〒820-8502 福岡県飯塚市川津 680-4
†† 九州工業大学大学院情報工学研究院 知能情報工学研究系 〒820-8502 福岡県飯塚市川津 680-4
E-mail: †hashiguchi.shunsuke739@mail.kyutech.jp, ††shimada@ai.kyutech.ac.jp

あらまし　近年，大学入試や就職試験において人とのコミュニケーション能力を測る手段としてグループディスカッションが活用されている．このような議論を試験として公平に評価することは難しい．そのため，議論に対して適切に評価するシステムが求められている．本研究では日本語の議論を対象とした品質評価タスクに取り組む．議論を対象としたタスクは，主に英語圏を中心に研究が行われており，日本語の議論を対象とした研究は少ない．そのため，日本語の議論データは英語に比べてデータ量は少ない．このような少量データに対処するため，対話特化モデルを利用した手法と GPT-4 によるデータ拡充の 2 つのアプローチを提案する．実験の結果，対話特化モデルを利用した手法は統計的な検定から有意差は得られなかったが，GPT-4 によるデータ拡充は本手法の有用性を確認した．
キーワード　議論，品質評価，対話特化モデル，GPT-4，疑似データ

Quality Assessment for debate using combined models and data augmentation

Shunsuke HASHIGUCHI† and Kazutaka SHIMADA††

† Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
†† Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
E-mail: †hashiguchi.shunsuke739@mail.kyutech.jp, ††shimada@ai.kyutech.ac.jp

Abstract　Recently, the incorporation of group debates has emerged as a strategic approach for measuring communication ability within the realms of entrance and employment examinations. However, the equitable evaluation of debates for such kinds of examinations proves challenging. Therefore, a system is required to assess debates appropriately. In this study, we work on the quality assessment of debates focused on the Japanese language. The studies related to debate primarily focused on English, while those targeted at Japanese are limited. Consequently, the availability of debate data in Japanese is more constrained compared to English. To handle the low-resource data, we propose two methods: the utilization of a dialogue-specific model and data augmentation using GPT-4. We demonstrate that employing a dialogue-specific model does not yield statistically significant scores, while data augmentation using GPT-4 enhances the performance.
Key words　debate, quality assessment, dialogue-specific model, GPT-4, pseudo data,

1. はじめに

近年，入試や就職試験で知識以外の能力を測る機会が増えてきている．その中でもコミュニケーション能力は，個人や組織との意見共有や問題解決など様々な場面において重要なスキルである．このコミュニケーション能力を測る手段の 1 つとしてグループディスカッションがある．採点者は，各参加者に対して協調性やリーダシップなどの評価項目を評価，あるいはグループ全体として良い議論ができたかを評価する．しかし，採点者によって議論を評価する基準が異なるため，入試や就職試験といった試験で公平な評価を行うことは難しい．そのため，議論に対して自動で評価を行うシステムを導入する必要がある．

本研究では，日本語の議論を対象とした品質評価タスクに取り組む．議論の品質評価タスクは，図 1 のようにグループディスカッションからグループ全体の議論の品質を評価するタスクである．議論の品質は，Wachsmuth ら [1] によって 3 種類の評価軸（合理性，有効性，適切性）から評価できると示されており，これらの評価が高いほどより品質の高い議論となる．本タスクのような議論を対象としたタスクは，主に英語圏を中心に研究が行われており，日本語の議論を対象とした研究は少な

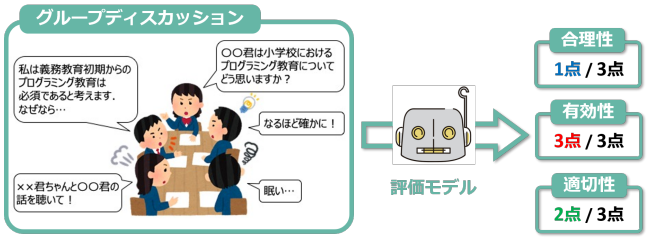

図 1　議論の品質評価タスクの概要．

い．そのため，日本語の議論データは英語に比べてデータ量は少なく，大規模な学習データが必要となる深層学習モデルを本タスクに直接利用するのは難しい．そこで，少量の日本語の議論データにおける議論評価モデルを構築するために 2 つのアプローチを提案する．

1 つ目のアプローチは，対話特化モデルの利用である．本タスクのような議論や対話の評価タスクを解く際，議論や対話の構造・特性を理解していることが望ましい．そこで，対話構造をモデル内部に組み込んだ対話特化モデルを利用する．しかし，日本語の対話データや議論データは少ないという現状から，日本語ベースの対話特化モデルは提案されていない．この問題に対応するために，英語ベースの対話特化モデルを利用する．英語圏の研究では，英語の対話データ量が多いため，英語ベースの対話特化モデルがいくつか提案されている [2] [3] [4]．英語の対話データを対話特化モデルに追加学習（ファインチューニング）し，追加学習したモデルを本タスクに適応することで，議論の構造・特性を考慮した品質評価ができると考えられる．しかし，英語の対話データで学習したモデルは日本語に関する知識を有していない．本研究では，日本語の知識を持つ汎用モデルと英語ベースの対話特化モデルを統合した手法を提案し，本タスクでの有用性を検証する．

2 つ目のアプローチは，GPT-4 によるデータ拡充である．データ拡充は，既存のデータセットから人為的にデータを拡張する手法である．対話形式のデータを生成する手法はいくつか提案されている [5] [6] が，多くの手法は対話システムの構築を目的としたデータ拡充手法である．つまり，ユーザとシステムの 1 対 1 による対話を生成することを目的としているため，議論のような複数人による対話を生成することは難しい．そこで，近年様々な生成タスクで高い精度を獲得している GPT-4 [7] を利用する．GPT-4 は OpenAI によって開発された大規模言語モデルで，自然な会話を生成することを得意としている．また，GPT-4 は単純なプロンプトでも高品質なデータを生成できる [8] ため，議論テーマをプロンプトとして GPT-4 に与えることで，高品質でより自然な議論を生成できることが期待される．本研究では，議論の品質評価タスクにおいて GPT-4 によるデータ拡充の有用性について調査する．

2. 関 連 研 究

本節では，本アプローチの関連研究について述べる．2.1 節では，品質評価タスクについての関連研究を述べる．2.2 節では対話特化モデルについて，2.3 節ではデータ拡張についての関連研究を述べる．

2.1　品質評価タスクに関する関連研究

議論・論の品質評価タスクの研究として，Wachsmuth ら [1] のものがある．Wachsmuth らは，過去の品質評価に関する研究を基に，議論・論の品質を評価する基準を定めた．Wachsmuth らの定義では，議論・論の品質は合理性，有効性，適切性の 3 つの評価軸から評価できると示されている．合理性が高い議論・論は，十分に容認できる形で議題の解決に貢献している議論・論のことを指す．有効性が高い議論・論は，主張が聞き手に納得・同意させる議論・論のことを指す．適切性が高い議論・論は，主張に含まれる前提が結論に対して十分，かつ関連している議論・論のことを指す．Wachsmuth らは，論に対して合理性，有効性，適切性の評価値が付与されたコーパスを構築した．

Shiota ら [9] は，日本語による対面対話を収録した複数人議論コーパス（Kyutech Debate Corpus）を構築し，議論データに対して合理性と有効性の評価ラベルを付与した．また，議論の品質を自動で評価するために，SVM や Attention 機構有りの LSTM などのモデルを構築した．本研究では，Shiota らによって構築した Kyutech Debate Corpus を使用する．しかし，Kyutech Debate Corpus は議論データが数百個と少量のデータセットである．本研究では，少量のデータセットに対して 2 つのアプローチを提案し，その有用性を調査する．

2.2　対話特化モデルに関する関連研究

近年では，深層学習を用いた対話特化モデルが提案されている．Majumder ら [2] は，対話から各話者の状態を考慮した RNN(Recurrent Neural Network) を構築した．しかし，RNN はテキストのような長い系列データを学習する際，誤差逆伝播による勾配消失が発生してしまう．この問題に対処するために，Ghosal ら [3] は，GCNN(Graph Convolutional Neural Network) を用いて話者自身・話者間の発話関係を考慮した枠組みを構築し，対話タスクにおいて高い精度を獲得した．しかし，ニューラルネットワークベースのモデルは大量の教師データが必要となるため，学習面でコストがかかってしまう．近年では BERT [10] を始めとした大規模言語モデルが主流となってきており，ニューラルネットワークよりも少ない学習データで，様々なタスクで高い精度を獲得している．Shen ら [4] は，事前学習済みモデルである XLNet [11] 内部に話者自身・話者間の発話関係を考慮した機構を組み込んだ手法 (DialogXL) を提案し，対話タスクにおいて RNN や GCNN をベースとした対話特化モデルよりも高い精度を獲得した．本研究では，この DialogXL を利用し，日本語の汎用モデルと DialogXL を統合した手法を提案する．

2.3　データ拡充に関する関連研究

データ拡充手法において，対話形式のデータを生成する手法がいくつか提案されている．Gao ら [5] は，言い換えによる対話の疑似データを生成するアプローチを提案し，少量データでの対話生成タスクにおいて高い精度を獲得している．また，Gritta ら [6] は，グラフ構造を用いて対話データを生成するアプローチを提案している．しかし，これらの手法は 1 対 1 の対話を目的としたデータ拡充手法であるため，議論のような複数人による対話を生成するのは難しい．近年では，自然な会話を

生成することを得意としている生成 AI が注目を集めている．Chan ら [12] は，質問と 2 つの回答に対して，どちらの回答がより質問の答えとして適切かを複数の GPT-4 を用いて議論を行い，自動的に評価した．また，生成 AI によるデータ拡充は，様々なタスクに応用されている．Anders ら [8] は，GPT-4 を用いることで高品質なラベル付きデータを生成し，少量データでのテキスト分類タスクにおいてデータ拡充の有用性を示した．また，Cochran ら [13] は，GPT-3.5 を用いて小論文データを生成し，小論文の自動採点タスクにおいてデータ拡充の有用性を示した．本研究では，GPT-4 を用いて議論データを生成し，学習データの拡充を行う．

3. データセット

本研究では，議論の品質評価タスクのデータセットとして，Shiota ら [9] によって構築された Kyutech Debate Corpus を使用する．Kyutech Debate Corpus には，5 つの議論テーマ (表 1) について，大学生および大学院生 4 人 1 組による討論・合意形成を含む計 10 対話分の議論データが収録されている．そして，各対話においてトピックセグメンテーションを行い，10 対話の議論データを 178 個の議論セグメントに分割した．トピックセグメンテーションは，与えられたテキストデータに対して異なるトピックやカテゴリに分割する手法である．例えば図 2 では，「飲酒可能年齢は 20 歳から下げられるべきである」という議題について議論を行ったとき，複数の議論セグメントが内在していることを示している．図 2 では，1 つの対話内に飲酒に関するトピックが 4 つ含まれており，各トピックに分割された議論が議論セグメントとなる．そして，各議論セグメントには，議論の評価基準である合理性・有効性について，どの程度満たしているかを 3 つのラベル（Low，Middle，High）に割り当てている．178 個の議論セグメントについて，2 つの評価基準をラベリングした結果，表 2 のようなラベル分布が得られた．本研究では，2 つの評価軸のうち有効性のラベルが付与された議論セグメントを使用する．

表 1　議論テーマの一覧.

(小中高の) 生徒は制服を着用すべきである
成人の拳銃所持・携帯の権利を認めるべきである
小中高の教材はタブレットに置き換えるべきである
飲酒可能年齢は 20 歳から下げられるべきである
未成年の暴力的ゲームのプレイを禁止すべきである

図 2　議論テーマ「飲酒可能な年齢は 20 歳から下げられるべきである」についての議論をトピックセグメンテーションしたときの議論セグメント.

4. 提 案 手 法

本節では，議論の品質評価タスクを解くための 2 つのアプローチについて述べる．4.1 節では日本語の汎用モデルと英語ベースの対話特化モデルを統合した手法について，4.2 節では GPT-4 によるデータ拡充について説明する．

表 2　各評価基準の正解ラベル分布.

評価軸	Low	Middle	High
合理性	13	89	76
有効性	9	97	72

4.1 複数モデルの統合

本節では，日本語の汎用モデルと英語ベースの対話特化モデルを統合したアプローチについて説明する．本アプローチでは，事前学習済みモデルとして日本語の汎用モデルである BERT [10] と，英語ベースの対話特化モデルである DialogXL [4] の 2 つのモデルを使用する．BERT は大規模なコーパスで事前学習された言語モデルであり，12 層の Transformer 層によって文脈を理解することができる．DialogXL は XLNet [11] と呼ばれる言語モデルをベースとした対話特化モデルである．議論のような対話形式の場合，話者間および話者自身の発話の依存関係が存在する．例えば議論では，ある人の主張や意見によって同意・納得の発言をしたり，反論したりすることで議論が進められる．しかし，BERT や XLNet といった既存の事前学習済みモデルは，これらの議論特有の性質を学習するのは難しい．DialogXL は，XLNet 内部の Transformer 層に話者間・話者自身の発話関係を考慮した機構が組み込まれているため，BERT に比べて議論や対話の構造・特性を理解することができる．

本アプローチでは，BERT と DialogXL の 2 つのモデルの損失を結合し，BERT あるいは DialogXL に結合した損失を誤差逆伝播する手法を提案する．本手法は，Qin ら [14] が提案した Knowledge Inheritance を模倣した手法である．Knowledge Inheritance は，教師モデルの知識を生徒モデルに継承する手法である．教師モデルと生徒モデルの損失を結合し，結合した損失を生徒モデルに誤差逆伝播を行うことで，生徒モデルに知識を継承する．本手法でも Qin らの手法に倣い，DialogXL の対話構造に関する知識を BERT に付与する手法（$\mathrm{Loss}_{\mathrm{BERT}}$）と，BERT の日本語に関する知識を DialogXL に付与する手法（$\mathrm{Loss}_{\mathrm{DialogXL}}$）を提案する．

$\mathrm{Loss}_{\mathrm{BERT}}$ の概略図を図 3 に示す．まず，日本語の議論コーパスを BERT と DialogXL の 2 つのモデルに入力する．DialogXL は英語ベースのモデルであるため，日本語の議論コーパスを英語に機械翻訳し，機械翻訳した議論コーパスを DialogXL に入力する．次に，2 つのモデルの損失を計算する．本研究では，3 つのラベルを分類するため，多クラス分類で用いられる CrossEntropy で損失を計算する．そして，2 つのモデルの損失を式 1 のように結合する．

$$\mathcal{L}_{\mathrm{All}} = \alpha \mathcal{L}_{\mathrm{DialogXL}} + (1-\alpha)\mathcal{L}_{\mathrm{BERT}} \tag{1}$$

ここで，$\mathcal{L}_{\mathrm{All}}$ は全体の損失，α は重み係数，$\mathcal{L}_{\mathrm{BERT}}$ は BERT の損失，$\mathcal{L}_{\mathrm{DialogXL}}$ は DialogXL の損失を表す．全体の損失 $\mathcal{L}_{\mathrm{All}}$ は，重み係数 α によって，どちらのモデルの予測を重視するかを決定することができる．例えば，α の値が 0.1 のとき，BERT

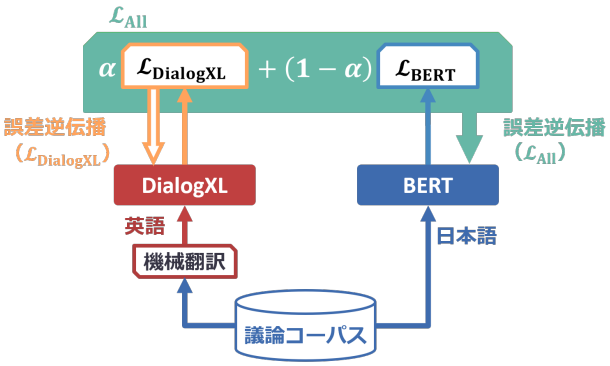

図 3　$\text{Loss}_{\text{BERT}}$ の概略図.

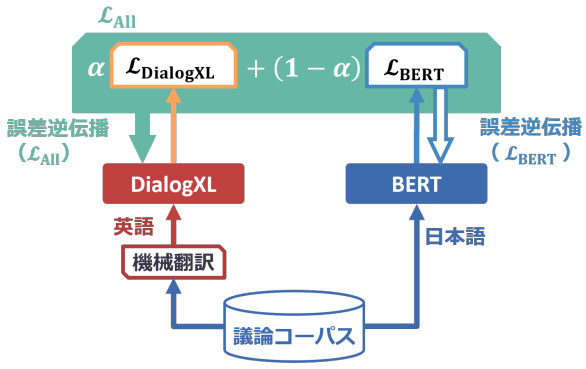

図 4　$\text{Loss}_{\text{DialogXL}}$ の概略図.

の損失 $\mathcal{L}_{\text{BERT}}$ の方が DialogXL の損失 $\mathcal{L}_{\text{DialogXL}}$ よりも大きくなるため，タスクを解く上で BERT の予測が重視される．そして，計算した全体の損失 $\mathcal{L}_{\text{All}}$ を BERT に誤差逆伝播し，BERT 内部のパラメータを更新する．これにより，DialogXL の対話構造に関する知識を BERT に付与することを期待する．

一方，$\text{Loss}_{\text{DialogXL}}$ では図 4 のように，全体の損失 $\mathcal{L}_{\text{All}}$ を DialogXL に誤差逆伝播し，DialogXL 内部のパラメータを更新する．これにより，BERT の日本語に関する知識を DialogXL に付与することを期待する．

4.2　GPT-4 によるデータ拡充

本節では，GPT-4 によるデータ拡充について説明する．GPT-4 は自然な対話を生成することを得意としている．しかし，複数の疑似データを生成する際，特定のトピックを与えると，議論内容が類似した疑似データが生成されてしまう．そこで，本研究では生成した疑似データの多様性を保つために，2 段階の処理を行う．GPT-4 による疑似データ生成の流れを図 5 に示す．第 1 段階では議論テーマに関するトピックの生成，第 2 段階ではトピックに関する議論の疑似データ生成を行う．

第 1 段階では，議論テーマに関連するトピックを生成する．プロンプトを図 6 に示す．ここで，<<theme>>は議論のテーマを表す．プロンプトには，議論テーマからディベート大会で使われるトピックを 20 個生成するように指示する．議論テーマは，Shiota らが構築した Kyutech Debate Corpus の 5 つの議論テーマ (表 1) を使用する．

第 2 段階では，第 1 段階で生成したトピックから議論の疑似データを生成する．プロンプトは，有効性ラベルが Low・High の場合と Middle の場合で異なる設計を行っている．その中で，有効性ラベルが Low，あるいは High の疑似データを生成するプロンプト（$\text{Prompt}_{\text{Low\&High}}$）は図 7 のように設計した．<<theme>>は議論のテーマ，<<topic>>は生成したトピックを表す．$\text{Prompt}_{\text{Low\&High}}$ は，タスク説明，有効性の定義，出力形式の 3 つの要素から構成されている．タスク説明では，GPT-4 に対してどのようなタスクを行わせるのかを指示する．ここでは，有効性の低い・高い議論を生成する指示以外に，議論の参加者，話者の設定などもプロンプトとして指示する．有効性の定義では，Wachsmuth ら [1] が定めた有効性の定義に従い，有効性を決める 5 つの要素（信用性：Credibility，情動性：Emotional Appeal，明瞭性：Clarity，妥当性：Appropriateness，順序性：Arrangement）をプロンプトとして指示する．出力形式では，GPT-4 に対してどのような出力形式でレスポンスさせるかを指示する．

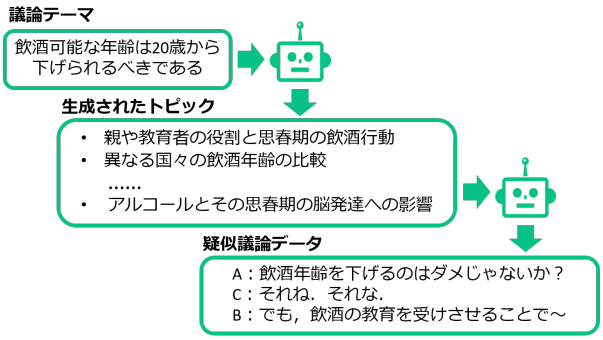

図 5　GPT-4 による疑似データ生成までの流れ図.

```
Please create 20 subtopics that can be used in debate
contest from the main topic.
The main topic is "<<theme>>".
The output format should be in a list:
Output: ["subtopic1", "subtopic2", "subtopic3",...]
Output:
```

図 6　トピック生成のプロンプト.

有効性ラベルが Middle の議論は，有効性ラベルが Low・High のどちらにも含意しない議論を生成する必要があるため，プロンプトとして定義するのは難しい．そのため，$\text{Prompt}_{\text{Low\&High}}$ の有効性の定義を表すプロンプトを除外し，単にトピックに関する議論を生成するように指示する．つまり，GPT-4 が考える一般的な議論を，有効性ラベルが Middle の議論とする．

5.　実　　験

本節では，前節で述べた 2 つのアプローチの実験について述べる．5.1 節では対話特化モデルを利用した手法について，5.2 節では GPT-4 によるデータ拡充についてそれぞれ実験を行う．

5.1　対話特化モデルの利用

本節では，4.1 節で説明した対話特化モデルを利用した手法の実験について述べる．5.1.1 節では実験設定について，5.1.2 節では実験結果と考察について述べる．

5.1.1　実 験 設 定

議論の品質評価タスクの評価方法について説明する．本研究では，Kyutech Debate Corpus に含まれる 10 対話の対話データ

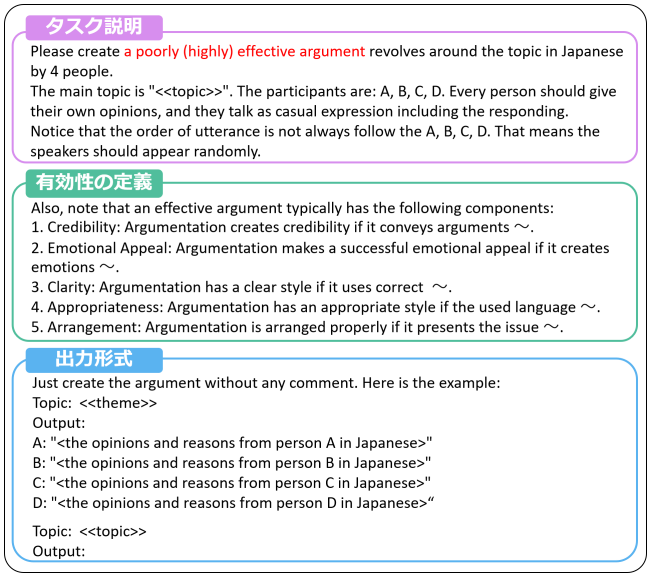

図 7　有効性ラベルが Low，High の疑似データを生成するためのプロンプト（$\text{Prompt}_{\text{Low\&High}}$）.

を，訓練データ 8 対話，検証データ 1 対話，テストデータ 1 対話に分割し，10 対話交差検証を 1 実験の評価とする．そして，結果の頑健性を保つために実験を 5 回行い，そのマクロ平均を報告する．本実験でのベースライン手法として BERT を用いる．評価指標には分類タスクで用いられる F 値を使用する．

本実験では対話特化モデルを利用した手法として，BERT と DialogXL を統合した手法以外に，DialogXL だけで品質評価を行った結果も示す．DialogXL は英語ベースのモデルであるため，日本語の議論コーパスを入力するとき，機械翻訳を通して英語に翻訳する必要がある．本実験では，機械翻訳器として Facebook が公開している多言語 mBART[1]を使用する．また，$\text{Loss}_{\text{BERT}}$ と $\text{Loss}_{\text{DialogXL}}$ に関しては，損失の重み係数 α によって，BERT と DialogXL どちらのモデルの予測を重視するかが決まる．本実験では，$\alpha = 0.1, 0.5, 0.8$ の 3 つの値を設定し，精度比較を行う．また，評価の際，$\text{Loss}_{\text{BERT}}$ は BERT のみ，$\text{Loss}_{\text{DialogXL}}$ は DialogXL のみで評価を行う．

モデルの設定について説明する．BERT は東北大学が公開しているモデルを使用する[2]．DialogXL はより対話構造の特徴を捉えるために，英語の対話データで追加学習する．対話データは IEMOCAP[3]を用いる．IEMOCAP は，2 人による 1 対 1 の会話が収録されており，各発話に感情を表すラベルが付与されている．この IEMOCAP で追加学習した DialogXL を本研究で利用する．損失関数は CrossEntropy，最適化関数は AdamW，学習率は 1e-5，バッチサイズは 8，エポック数は 50 と設定して実験を行う．

5.1.2　実験結果および考察

本節では，対話特化モデルを利用した手法の実験結果と考察について述べる．実験結果を表 3 に示す．Low，Middle，High は各ラベルの F 値，Ave は Low，Middle，High の F 値の重み付き平均を表す．表内の数字の太字は，各ラベルの F 値および重み付き平均において最も高い精度を表す．

表 3　対話特化モデルを利用した手法の実験結果.

手法	Low	Middle	High	Ave.
BERT	0.000	0.634	0.466	0.534
DialogXL	0.000	0.641	0.500	0.551
$\text{Loss}_{\text{BERT}}(\alpha{=}0.1)$	0.000	**0.659**	0.477	**0.552**
$\text{Loss}_{\text{BERT}}(\alpha{=}0.5)$	0.000	0.646	0.467	0.541
$\text{Loss}_{\text{BERT}}(\alpha{=}0.8)$	0.000	0.642	0.488	0.547
$\text{Loss}_{\text{DialogXL}}(\alpha{=}0.1)$	0.000	0.623	0.402	0.502
$\text{Loss}_{\text{DialogXL}}(\alpha{=}0.5)$	0.000	0.625	0.417	0.509
$\text{Loss}_{\text{DialogXL}}(\alpha{=}0.8)$	0.000	0.602	0.444	0.508

今回提案した $\text{Loss}_{\text{BERT}}$，$\text{Loss}_{\text{DialogXL}}$ について実験結果と考察を述べる．$\text{Loss}_{\text{DialogXL}}$ は，BERT と比べて精度の低下が見られる．DialogXL は，モデル内に発話関係を考慮した機構が組み込まれている．BERT による日本語の知識付与が，発話関係を学習する上でノイズとなっていることが考えられる．$\text{Loss}_{\text{BERT}}$ は，BERT や他の手法よりも良い結果となっているが，統計的な検定から有意差はなかった.[4] 日本語の汎用モデルと英語の対話特化モデルの統合は，互いの学習に影響を及ぼしてしまうことが考えられる．

本実験において，全ての手法で Low の F 値が 0 となっている．表 2 より，有効性ラベルの Low データが他のラベルと比較して極端に少ないため，全ての手法で正しく予測することができなかったと考えられる．

5.2　GPT-4 によるデータ拡充

本節では，4.2 節で説明した GPT-4 によるデータ拡充の実験について述べる．5.2.1 節では実験設定，5.2.2 節では実験結果と考察を述べる．

5.2.1　実 験 設 定

議論の品質評価の評価方法，損失関数や最適化関数などのモデル設定については，5.1.1 節の実験設定に倣う．

本実験では，各有効性ラベルに 100 個の疑似データを生成し，学習データの拡充を行う（Mix_{All}）．生成された疑似データは，検証データには用いず，全て訓練データとして用いる．また，表 2 の有効性ラベルの分布より，Kyutech Debate Corpus のラベルにはデータの偏りがある．そこで，データの偏りを考慮した手法として，Low ラベルと High ラベルの疑似データを訓練データに加えた手法（$\text{Mix}_{\text{Low\&High}}$），Low ラベルの疑似データのみを訓練データに加えた手法（Mix_{Low}）の 2 つのデータ拡充手法も行う．そして，BERT を用いて各データ拡充手法を適用し，精度の比較を行う．

5.2.2　実験結果および考察

本節では，GPT-4 によるデータ拡充の実験結果と考察について述べる．実験結果を表 4 に示す．表内の数字の太字は，各ラベルの F 値および重み付き平均において最も高い精度を表す．3 つのデータ拡充手法全てにおいて，Low ラベルの疑似データを加えたことで，Low ラベルの F 値が向上した．少量のラベル

（注1）：https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
（注2）：https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
（注3）：https://sail.usc.edu/iemocap/index.html
（注4）：有意水準 5% でマクネマー検定を実施.

表 4　GPT-4 によるデータ拡充の実験結果.

手法	Low	Middle	High	Ave.
BERT	0.000	**0.634**	0.466	0.534
Mix_{All}	0.138	0.605	0.486	0.534
$\text{Mix}_{\text{Low\&High}}$	0.110	0.618	0.505	0.547
Mix_{Low}	**0.156**	**0.634**	**0.510**	**0.560**

データに対して，データ拡充を行うことの有効性が示された．

各データ拡充手法についての実験結果と考察を述べる．Mix_{All} は全ラベルの疑似データを加えても，精度の向上は見られなかった．Middle ラベルの疑似データは，他ラベルの疑似データと比べて有効性に沿った議論となっていない．そのため．Middle の疑似データは学習のノイズとなり，Middle の F 値が低下したと考えられる．$\text{Mix}_{\text{Low\&High}}$ は Middle ラベルの疑似データを除くことで，BERT および Mix_{All} よりも精度の向上が見られる．しかし，依然として Middle ラベルの F 値は BERT と比べて低下している．High ラベルの疑似データを入れたことで，High ラベルにデータが偏っている．これにより，Middle ラベルの F 値が低下したと考えられる．Mix_{Low} は，BERT および他のデータ拡充手法と比べて精度の向上が見られる．Low ラベルの疑似データのみデータ拡充したことで，ラベル分布は均一になった．その結果，他のデータ拡充手法よりも Middle ラベルの F 値が向上した．ラベルの均一および希少ラベルに対するデータ拡充は重要であることがいえる．

6. おわりに

本論文では，少量の日本語の議論データを用いて，議論の品質評価タスクに取り組んだ．そして，少量の日本語の議論を評価するために，対話特化モデルを利用した手法と GPT-4 によるデータ拡充の 2 つのアプローチを提案し，本タスクの有用性を検証した．

1 つ目の対話特化モデルを利用した手法では，日本語の汎用モデルと英語ベースの対話特化モデルを統合し，議論の品質評価を行った．結果として，提案した 2 つの手法は，ベースラインと比較して精度の向上は見られない，あるいは統計的に有意な差はなかった．異なる言語を持つモデルを組み合わせて学習するのは困難であることがいえる．

2 つ目のデータ拡充手法では，生成 AI である GPT-4 で疑似データを生成し，学習データの拡充を行った．結果として，希少なラベルデータに対して疑似データを付与することによって，最高精度を得た．

今後の課題として，本論文ではテキストのみを用いて議論の品質評価を行ったが，Shiota ら [9] による研究では音声や画像を含めたマルチモーダルによる品質評価を行っている．話者による意見・主張を述べる際の声の大きさや，話者の表情・しぐさなどの情報は議論の品質評価に寄与することが考えられる．本論文で提案したデータ拡張手法をマルチモーダルに組み入れ，さらなる品質評価の精度向上を目指す．

謝辞　本研究は科研費 23K11368 の一部です．

文　献

[1] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 176–187, 2017.

[2] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6818–6825, 2019.

[3] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. 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. 154–164, 2019.

[4] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 13789–13797, 2021.

[5] Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 639–649, 2020.

[6] Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. Conversation graph: Data augmentation, training, and evaluation for non-deterministic dialogue management. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 36–52, 2021.

[7] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[8] Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*, 2023.

[9] Tsukasa Shiota and Kazutaka Shimada. Annotation and multi-modal methods for quality assessment of multi-party discussion. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pp. 175–182, 2022.

[10] 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, 2019.

[11] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, Vol. 32, , 2019.

[12] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[13] Keith Cochran, Clayton Cohn, Jean Francois Rouet, and Peter Hastings. Improving automated evaluation of student text responses using gpt-3.5 for text data augmentation. In *Artificial Intelligence in Education*, pp. 217–228, 2023.

[14] Yujia Qin, Yankai Lin, Jing Yi, Jiajie Zhang, Xu Han, Zhengyan Zhang, Yusheng Su, Zhiyuan Liu, Peng Li, Maosong Sun, et al. Knowledge inheritance for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3921–3937, 2022.