アスペクト文判別のための大規模言語モデルを用いたデータ拡張の有効性

川嵜 慎乃介[1]　嶋田 和孝[2]

概要：レビューから評価項目（アスペクト）を自動的に抽出するタスクをアスペクト推定と呼ぶ．アスペクト推定は，詳細なレビュー分析などの為に重要なタスクである．しかし，データセットによってはアスペクトごとのデータのばらつきやデータの不足が起きる場合がある．その結果，推定精度が十分ではない場合が存在していた．その解決策としてデータを拡張することが考えられるが，旧来は単語置換などの手法によって行われており，文意が変わってしまうなどの問題があった．本研究では生成 AI である ChatGPT を利用してデータの拡張を行う．本論文では拡張データの有無によりアスペクト推定精度がどのように変化するのかを実験的に検証し，考察する．

キーワード：分類学習，評判・感情解析，自然言語生成・言い換え

Effectiveness of Data Augmentation Using LLM for Aspect Classification

Shinnosuke Kawasaki[1]　Kazutaka Shimada[2]

Abstract: Aspect classification is an important task for review analysis. However, depending on the dataset, there may be a lack of data for certain aspects. Consequently, classification accuracy can sometimes be insufficient. One solution is to augment the data, but this could lead to issues such as altering the meaning of the text. In this study, we employ ChatGPT, an AI text generation tool, to augment the data. We experimentally investigate and discuss how aspect classification accuracy is affected by the presence or absence of augmented data.

Keywords: Classification Training, Review/Emotion Analysis, Natural Language Generating

1. はじめに

近年，通販サイトの普及などによりインターネット上には日夜多くのレビュー文が投稿されている．レビュー文の中には機能に対する意見や値段に対する意見など，複数の評価項目に対する意見が混在している．これらの評価項目を自然言語処理分野においてはアスペクトという．多量のレビュー文集合から特定のアスペクトに関するレビュー文を抜き出すことはユーザにとって有意義なことであるが，大量のレビュー文を人手で分類することは非常にコストがかかってしまう．そのため，システムによって自動的にレビュー文中に含まれるアスペクトを特定し，ラベル付けすることが望ましい．そのようなタスクをアスペクト推定という．アスペクト推定によってラベル付けされたレビュー文集合から，任意のアスペクトに関するレビュー文を抜き出すことは非常に容易となり，ユーザにとって大きなコスト削減につながる．

Nakamuta ら [1] は，複数のアスペクト情報が内在する

[1] 九州工業大学 大学院情報工学府
Department of Creative Informatics, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
[2] 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
Department of Artificial Intelligence, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN

ゲームレビュー文を対象にアスペクト推定を行っている．Nakamuta らの手法では，レビュー文中の単語情報からアスペクトの特定に繋がりやすい語彙に絞ってベクトル化を行ったのち，SVM という機械学習モデルを用いてアスペクトの推定を行っている．Nakamuta らによる手法の問題点として，元のレビュー文集合の数が不十分なため訓練が十分に行うことができず，いくつかのアスペクトにおける低い推定精度に繋がっていることが論文中に指摘されている．本研究では，この問題点を踏まえたうえでゲームレビュー文に対するアスペクト推定の精度を向上させることを目的とする．

図 1 に，本研究で提案する手法を表した概略図を示す．前述の問題点への対応として，本研究では 2 段階のアプローチを提案する．1 つ目は，推定モデルの変更である．Nakamuta らが使用している SVM は訓練に使用したゲームレビュー文から得られる情報のみを用いてアスペクトの推定を行っていた．しかし，近年の自然言語処理分野で広く用いられている事前学習済みモデル BERT[2] を用いることで，SVM と同量の訓練データでもより高い推定精度を期待する事ができる．事前学習済みモデルとは，事前に大規模コーパスによる事前学習により汎用的言語特徴をとらえている言語モデルである．事前学習を行った後，特定のタスクに適応させる学習が必要となるが，この際に用いるデータセットはたとえ少量であっても従来の機械学習モデルより高精度な推論を行えることが知られている．そのため，本研究で用いるゲームレビューデータセットに対しても SVM より高精度な推論を行えることを期待して実装する．

2 つ目は，拡張データの生成である．自然言語処理分野においては，データセットの量的不足に対して，単語置換などのデータ拡張手法を用いることで訓練データ量を増加させることが一般的である．しかし，本研究で使用するゲームレビュー文は，各ゲームに関する用語や登場人物等を置換してしまうと本来のゲームに関するレビューですらなくなってしまい，推定精度を一方的に下げてしまうノイズとなってしまう恐れがある．そこで，本研究では拡張データの生成に大規模言語モデルである ChatGPT を使用する．ChatGPT とは，OpenAI 社の AI サービスの 1 つである．その特徴として，まるで人と対話するように応答することができる点，人間から見ても非常に自然な文を生成することができる点が挙げられる．また，文章生成モデルを用いた拡張データの有効性を示す研究も示されている [3]．しかし，ChatGPT システム内部に存在する言語モデル GPT-3.5 や GPT-4.0 の詳細な性能は公開されていない．そのため，研究分野における実用性について，様々な研究が盛んに行われている [4] [5]．本研究では ChatGPT をゲームレビューデータにおける拡張データ生成に使用することで，推定精度に与える変化や実際に生成された文章

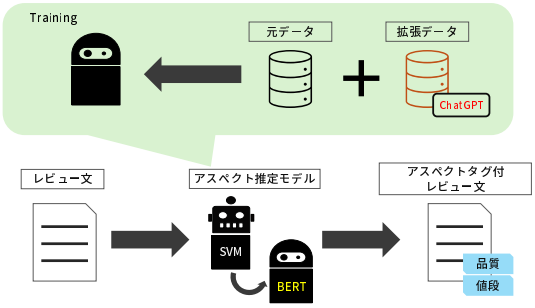

図 1 本研究における精度向上のための取り組みの概要．
Fig. 1 Summary of efforts for improving accuracy in this study.

からその有効性を検証する．また，生成モデルに分類結果を出力させることで，分類モデルと同等の精度を示す研究が示されている [6]．高性能な生成 AI である ChatGPT も同様に分類問題に適応できるのではないかと考え，分類モデルとして実装し，SVM や BERT との精度比較を行う．

2. 関連研究

自然言語処理分野においてアスペクト推定は ABSA（Aspect Based Sentiment Analysis）と呼ばれる感情分析タスクのサブタスクとして位置づけられている．ABSA タスクは，テキスト中の様々な側面（アスペクト）に対する感情極性（ポジティブ，ネガティブ，中立）を判定するタスクである．過去に様々なドメイン（特定の分野やテーマ）のデータセットが公開されており，例としてレストランやホテルのレビューを収集した SemEval-2016[7] や，金融関係のツイートデータを収集した FiQA-2018[8] などが存在する．ABSA タスクにデータ拡張手法を取り込んだ研究は数多く行われている．Wei ら [9] は，ランダムに文中単語の置換や削除を行うデータ拡張手法 EDA を提案し，精度の向上を確認した．Kobayashi ら [10] は，ラベル情報を考慮しつつ言語モデルを使用した単語置換を行うデータ拡張手法 Contextual Augmentation を提案し，精度の向上を確認した．しかし，これらのデータ拡張手法ではそれ以外のコーパスではあまり見られないゲーム用語に影響を与え，ゲームレビューデータという属性を変更してしまう恐れがある．また，拡張手法に ChatGPT を使用した研究はまだ数少なく，議論の余地がある．

Nakamuta らと同様のデータセットを用いてアスペクト推定の後段タスクに取り組んでいる研究は複数存在している．Takeo ら [11] はデータ中のアスペクト情報を使用し，各アスペクトの評価値推定を行った．Tadano ら [12] はクラスタリング手法を用いて複数アスペクトの要素を含む要約文生成を行った．アスペクト推定の精度を向上させることで，これらの後段タスクにおいてアスペクト情報を活用することが容易となる．

ChatGPT の推定モデルや生成手法としての有効性を検証する研究は，徐々に増加している．Tang ら [13] は金融ド

アスペクト	熱中度 (a)	快適さ (c)	難易度 (d)	グラフィック (g)	音楽 (m)	オリジナリティ (o)	満足度 (s)
評価値	3	4	4	4	4	4	1
良い所	・<s,d>初心者～上級者まで楽しめる</>。 ・<s>今までのマリオカートの中で一番だと思う</>。 ・<o>カートの種類が多い</>。 ・<s>今までのコースがあるのはうれしい</>。						
悪い所	・ワイファイをやっているとき、突然電源を切る人が多い。 ・<c>ここは消しても通信切断ではなく、CPUがでてきてくれればよかった</> ・あと、ワイファイでは直線ドリフトが多すぎ ・直線ではドリフトなしのほうがよかったと思う。						
感想など	いい所でもいいましたが<s>今までのマリオカートの中で一番だと思う</>。 ただ、直ドリを改善すればもっと良い作品が出来たと思う。 次回作にすごく期待します。						

図 2 実際のレビュー例.
Fig. 2 Example of actual review.

メインの感情分析タスクにおいて，ChatGPT を含むいくつかの推定モデルの精度を検証した．その結果 Zero-shot, Few-shot 設定における ChatGPT の推定精度は BERT を下回ることが示されている．Piedboeuf ら [14] は多クラスの文書分類タスクに対して，元データに ChatGPT によって生成された拡張データを追加し，推定精度の変化を検証した．Cegin ら [15] は多クラスの文書分類タスクに対して，複数のデータ拡張手法に対して ChatGPT など複数の大規模言語モデルを使用し，推定精度の変化を検証した．これらの研究は本研究と同様に多クラス分類タスクにおける ChatGPT の拡張データの有効性を検証している．しかし，ゲームレビューという特異性の高いドメインのデータセットに対しての ChatGPT の有効性は検証できていない．

3. データセット

本研究では，Nakamuta らが使用したものと同様のゲームレビューデータセットを使用する．本データセットは Nakamuta らにより Web サイト[*1]から収集されたデータである．Web サイトには任天堂のゲーム機であるニンテンドー DS を対象に発売されたゲームソフトへのレビューが投稿されている．図 2 に，実際のレビューの例を示す．1 つの投稿につき 1 つのゲームソフトへのレビューが投稿されており，上部のアスペクトごとの評価値欄と，ゲームの感想を記述する記述欄に分かれている．記述欄には各アスペクトの評価値を裏付けるための良かった点や悪い点などが記されている．本研究では，これら記述欄のみを使用し，文単位のアスペクト推定を行う．また，今回のアスペクト推定ではあくまで文中のアスペクトに関する記述を抜き出すことが目的であり，内容が良い評価であるか悪い評価であるかという違いは影響しないため，それらの区別はしないものとする．

図 2 のように，本データセットには記述欄の文章に対して，どのアスペクトに対する記述であるかを示すためのアスペクトラベルがタグ形式で人手によって付与されている．図 2 の 1 文目のように，1 文中で複数のアスペクトについて言及している場合は，その全てのアスペクトに対してラベル付けが行われている．本研究ではこれらアスペクトラベルが付与されている文章のみを使用する．具体的には，タグで囲まれている部分を抜き出して使用する．そのため，実験で使用するレビュー文は全部で 4719 文となる．アスペクトラベルの付与は文単位で行われており，本研究ではそれらを正解ラベルとして用いる．アスペクトラベルは評価値と同様に 7 種類存在している．それぞれアスペクトラベルが付与されているデータ数を表 1 に示す．前述の通り 1 文に複数のアスペクトラベルが付与されている場合もあるため，アスペクトごとのデータ数の合計が全データ数と等しくならないことに注意が必要である．

[*1] http://ndsmk2.net

表 1 各アスペクトラベルが付与されているデータ数.
Table 1 Number of data instances with each aspect label.

アスペクト名	データ数
熱中度	429
快適さ	354
難易度	353
グラフィック	229
音楽	258
オリジナリティ	2339
満足度	2252
合計	4719

4. アスペクト推定

本節では，アスペクト推定に使用する機械学習モデル及び，拡張データを使用しない状態でのアスペクト推定実験の設定及び結果について述べる．4.1 項では，本研究で使用するアスペクト推定のフレームワークと機械学習モデルについて述べる．4.2 項では，アスペクト推定実験の実験設定について述べる．4.3 項では，アスペクト推定実験の結果について述べる．

4.1 推定モデル

図 3 にアスペクト推定のフレームワークを示す．本研究では，Nakamuta らの先行研究に倣い，与えられたレビュー文に対し各アスペクトの内容が含まれているか否かを判定する 2 値分類器を実装する．このとき，各アスペクトの推定モデルは独立しており，パラメータの共有等は一切行わない．以下の項では，図 3 中の 2 値分類器に用いられる言語モデルについて説明する．

4.1.1 SVM

Nakamuta らの研究においてアスペクト推定に用いられた SVM (Support Vector Machine) を再現して実装する．入力文をベクトル化する際に用いる素性は Nakamuta らにならい，条件付き BoW (Bag of Words) とする．素性の条件付き BoW を表した式を式 1 に示す．

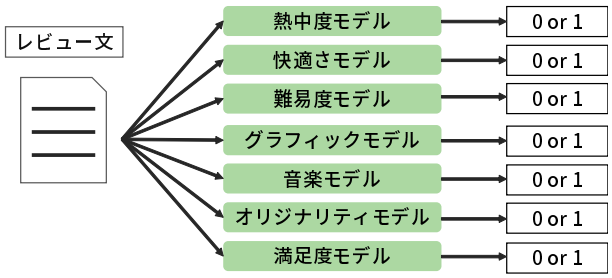

図 3 アスペクト推定モデルのフレームワーク．
Fig. 3 Framework of aspect estimation model.

$$f = \{w_1^a, w_2^a, ..., w_{na}^a, w_1^c, ..., w_{nc}^c, ..., w_1^s, ..., w_{ns}^s\} \quad (1)$$

ここで，w_j^i は単語 w_j のアスペクト i における値 Val を表す．値 Val を求める式を式 2 に示す．

$$Val\,(asp_i, w_j) = \frac{num_{ij}}{sent\,(asp_i)} \quad (2)$$

ここで，num_{ij} はアスペクト i における単語 w_j の頻度を表し $sent\,(asp_i)$ はアスペクト i ラベルが付与された文章の数を表す．各アスペクトの総データ数は異なるため，正規化のために $sent\,(asp_i)$ が用いられる．これら w_i の語彙は訓練データ中の単語によって作成され，テストデータにおいては訓練時に作成された語彙を使用する．また，素性に含まれる語彙の品詞は名詞，形容詞，動詞のみとする．

4.1.2 BERT

図 4 に BERT の処理の流れを示す．BERT (Bidirectional Encoder Representations from Transformers) は，Google 社によって開発された事前学習済み言語モデルである．12 層の Transformer によって構成されており，文章の先頭の特殊トークン [CLS] を使用して 2 値分類を行う．事前に大規模コーパスによって訓練されているため，文章中の汎用的言語特徴をタスク適応前に捉えることができる．また，文脈情報を捉えた推測を行えることも BERT の特徴として知られている．BERT は事前学習の後ファインチューニングと呼ばれる適応学習を行うことによって特定のタスクに適応させることができる．ファインチューニングの際には大規模なデータセットを必要とせず，小規模なデータセットであっても高精度な推定が行えることが知られている．Nakamuta らの研究において，幾つかのアスペクトデータが少量であったためにアスペクト推定精度が不十分であったことが示唆されている．そのため，本研究では Nakamuta らと同量の訓練データ量であってもより高い推定精度であることを期待する．本研究では東北大学が提供している日本語用事前学習済み BERT モデル[*2]を使用する．

4.1.3 ChatGPT

生成 AI である ChatGPT の推定モデルとしての有効性

[*2] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

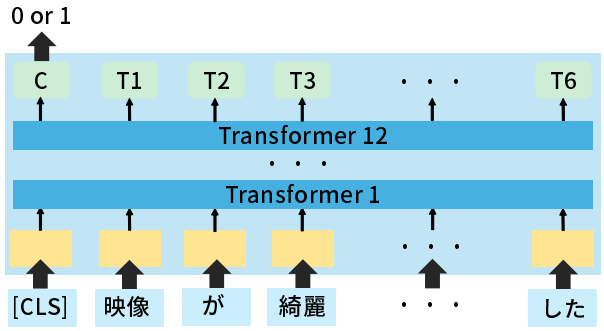

図 4 BERT による処理の流れ．
Fig. 4 BERT.

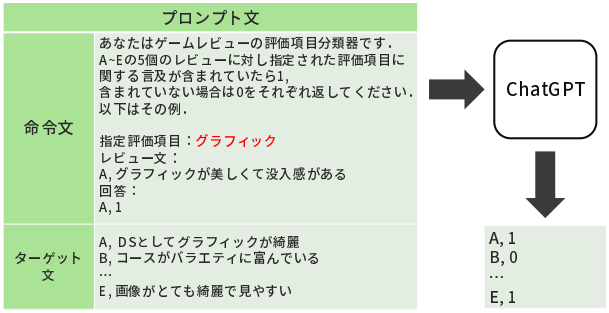

図 5 ChatGPT によるアスペクト推定の流れ図．
Fig. 5 Flowchart of aspect estimation with ChatGPT.

を検証するために，ChatGPT によるアスペクト推定を実装する．ChatGPT によるアスペクト推定には OpenAI 社が提供する API を使用し，そのモデルは GPT-3.5-turbo とする．ChatGPT によるアスペクト推定の流れを図 5 に示す．API を利用する際には ChatGPT に対して入力となるプロンプト文が必要となる．プロンプト文の内容は図 5 に示すようにタスクの説明となる命令文と推測を行うターゲット文からなる．命令文はタスクの説明部分と出力形式を誘導するための実例部分に分かれる．タスクの説明部分はいずれのアスペクトについてもほぼ同様の文を使用するために，図 5 に赤字で示されているアスペクト名のみを変更して各アスペクトの推定を行うよう指示する．実例部分は実データの中から 1 文とその文中にアスペクトが存在しているかを示すラベル（1 もしくは 0）をペアとして与える．ターゲット文は 5 文単位でテストデータから抜き出され，ChatGPT に与えられる．ターゲット文と命令文の実例部分は同じ文が与えられないようになっている．ChatGPT が生成する文章のランダム性を設定するパラメータ Temperature は 0 に設定する．

4.2 実験設定

アスペクト推定実験を行う際，Nakamuta らの実験に倣い，訓練データ中のアスペクトデータ数の偏りによって不十分な訓練が行われないようにするため，訓練データに対してデータ数の調整を行う．訓練データに使用するデータ

表 2 “熱中度”アスペクトのデータ調整結果.
Table 2 Adjustment results of “熱中度” aspect data.

	元データ数	調整後データ数	
熱中度	429	429	
快適さ	354	26	427
難易度	353	26	
グラフィック	229	17	
音楽	258	19	
オリジナリティ	2339	173	
満足度	2252	166	

数 $uses$ を求める式を式 3 に示す.

$$use_s(asp_i, asp_j) = read_s(asp_j) \times \frac{real_s(asp_i)}{all_s - real_s(asp_i)} \quad (3)$$

$real_s(asp_i)$ はアスペクト asp_j の文の数、all_s は全データ数 4719 文である．式 3 をアスペクト“熱中度”に適応した場合を表 2 に示す．アスペクト“熱中度”のデータ数 429 文に合わせてそれ以外のアスペクトのデータ数の合計が 427 文となる．

SVM のカーネルは RBF カーネルを使用する．BERT の学習率は 0.00001，epoch 数は 5，バッチサイズは 16 を使用する．評価指標は F1 値を使用する．実験の際には 10 分割交差検証を行い，その平均値を評価値とする．データの分割は SVM においては訓練：テスト = 9 : 1 とし，BERT においては訓練：検証：テスト = 8 : 1 : 1 とする．

4.3 実験結果

表 3 にアスペクト推定実験の結果を示す．全てのアスペクトにおいて，BERT が最も精度が高く，ChatGPT が最も精度が低い結果となった．全モデルに共通する特徴として，“グラフィック”と“音楽”は精度が高くなる傾向があった．これらのアスペクトでは，そのアスペクト内容を表す特徴的な語彙（アスペクト語）が少なく，推定がしやすかったためであると考えられる．また，全モデルにおいて“オリジナリティ”と“満足度”に対しては精度が下がる傾向が見られた．これらのアスペクトは，前述の“グラフィック”等と比べ内容の抽象度が高く，推定難易度の高いアスペクトであったと考えられる．さらに抽象度が高いアスペクトではレビュー記述者ごとに用いる単語が異なる．その結果，推定のための有効な語彙が広範となり，素性に BoW を使用している SVM において“熱中度”や“満足度”の精度は下がる傾向にあったと考えられる．

5. データ拡張

本研究では，Nakamuta らの先行研究によって言及されたアスペクトデータの不足を補うために，ChatGPT によって生成した拡張データを追加する．本節では，その手法と実験結果及び考察について述べる．5.1 項では ChatGPT を利用した拡張データ生成の手法について述べる．5.2 項では実験設定について述べる．5.3 項では拡張データを追加したアスペクト推定実験の結果とその考察を述べる．

表 3 アスペクト推定実験結果.
Table 3 Results of aspect classification.

	SVM	BERT	ChatGPT
熱中度	0.772	**0.901**	0.688
快適さ	0.874	**0.895**	0.659
難易度	0.903	**0.926**	0.753
グラフィック	0.931	**0.963**	0.832
音楽	0.926	**0.947**	0.792
オリジナリティ	0.829	**0.874**	0.633
満足度	0.787	**0.834**	0.660

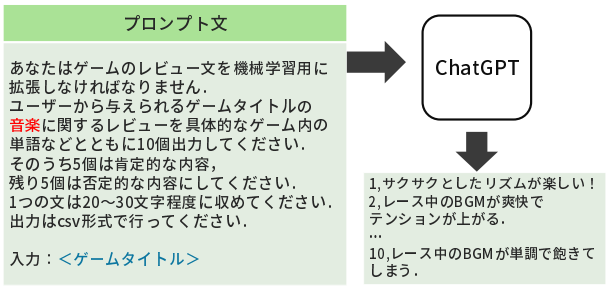

図 6 ChatGPT による拡張データ生成の流れ図.
Fig. 6 Flowchart of augmented data generation with ChatGPT.

5.1 提案手法

図 6 に，ChatGPT を使用した拡張データ生成の流れを示す．ChatGPT への入力としてプロンプトを使用し，生成するべきデータの設定や出力形式を指定する．拡張データの多様性を担保するために，生成させる 10 文のうち肯定的な文と否定的な文が 5 文ずつとなるように指定する．1 リクエストごとにゲームタイトルを 1 種指定し，1 文 20〜30 文字程度の拡張データを 10 文生成させる．ゲームタイトルは，元データ中に含まれているゲームタイトルの中から 10 個にしぼり，各ゲームタイトルに対し 10 文ずつ拡張データ生成を行う．このようにして 100 文の拡張データを生成した時点で 1 サイクルとする．図 6 のプロンプト中に赤字で示されているアスペクト名のみを切り替えてこのサイクルを n 回実行し，拡張データをアスペクトごとに同数生成する．生成の際には OpenAI 社の提供する API を使用し，モデルは GPT-3.5-turbo を使用する．GPT が生成する文章のランダム性を指定するパラメータ Temperature は 1.0 に設定する．

5.2 実験設定

使用する推定モデルは，4.3 項にて最も精度の高かった BERT に絞り拡張データを追加し，精度の変化を検証する．BERT のパラメータは 4.3 項と同じく，学習率 0.00001，バッチサイズ 16，epoch 数 5 で訓練したものを使用する．

表 4 200 文の拡張データを追加した場合の訓練データ中に占める拡張データの割合.

Table 4 The proportion of augmented data in the training data when adding 200 sentences of augmented data.

	訓練データ中の拡張データ割合
熱中度	23.0% (200 / 868)
快適さ	26.4% (200 / 757)
難易度	26.4% (200 / 755)
グラフィック	36.3% (200 / 550)
音楽	33.0% (200 / 606)
オリジナリティ	6.0% (200 / 3324)
満足度	6.1% (200 / 3255)

ChatGPT によって生成した拡張データにはすべて正解ラベルを付与し，訓練データにのみ追加する．拡張データ生成のサイクル数 n は $n = 2$ とし，アスペクトごとに 200 文の拡張データを生成する．各アスペクトの訓練データに対しそれぞれ 200 文の拡張データを追加した場合の，訓練データ中における拡張データの割合を表 4 に示す．しかし，3 節中の表 1 で示したように，アスペクトのデータ数には偏りが存在している．ゆえに元データの少ない“グラフィック”では拡張データが訓練データ中の 36.4%を占めるのに対し，元データの多い“満足度”ではわずか 6%となる．そのため，200 文の拡張データを全アスペクトにおいて一様に追加してしまうと，元データ数の少ないアスペクトほど元データが少なくなり，適切な学習が行えない恐れがある．そこで，本研究では訓練データ中の拡張データ割合を＜少＞と＜多＞の 2 つの場合に分け，拡張データの効果検証を行う．

各アスペクトにおける＜少＞と＜多＞それぞれの場合における訓練データ中の拡張データの割合を表 5 に示す．＜少＞においては，表 4 の下限である“オリジナリティ”の 6%付近に統一し，＜多＞においては，全アスペクトにおいて拡張データの割合が 25%以上となるように統一する．表 4 においてこれらの条件に当てはまるアスペクトは 200 文のまま使用する．拡張データが不足しているアスペクトは ChatGPT を使用した拡張データを追加で生成したのち，それらの中から無作為に抽出して割合を調節する．また，拡張データが過剰に存在しているアスペクトは 200 文中から無作為に拡張データを抽出し，割合を調節する．＜少＞及び＜多＞という 2 つの場合における精度の変化を確認し，ChatGPT によって生成した拡張データの質や有効性を検証する．評価値は F1 値を使用する．4.3 項と同様に 10 交差検証を行い，その平均値を評価値とする．

5.3 実験

拡張データを追加したアスペクト推定実験の結果を表 6 に示す．拡張データが＜多＞の場合ではすべてのアスペクトにおいて推定精度が低下する結果となった．また，

表 5 ＜少＞及び＜多＞設定における各アスペクト訓練データ中に占める拡張データの割合.

Table 5 The proportion of augmented data in the training data for each aspect setting of ＜少＞ and ＜多＞.

	＜少＞	＜多＞
熱中度	6.0% (43 / 711)	26.0% (235 / 903)
快適さ	6.0% (36 / 593)	26.4% (200 / 757)
難易度	6.0% (36 / 591)	26.4% (200 / 755)
グラフィック	6.0% (23 / 373)	36.3% (200 / 550)
音楽	6.0% (26 / 432)	33.0% (200 / 606)
オリジナリティ	6.0% (200 / 3324)	26.0% (1097 / 4221)
満足度	6.1% (200 / 3255)	26.0% (1073 / 4128)

表 6 データ拡張実験結果.

Table 6 Results of aspect classification with data augmentation.

	拡張データ割合		
アスペクト	0%	＜少＞	＜多＞
熱中度	**0.901**	0.888	0.875
快適さ	**0.895**	0.886	0.832
難易度	**0.926**	0.907	0.873
グラフィック	**0.963**	0.957	0.917
音楽	0.947	**0.954**	0.946
オリジナリティ	**0.874**	**0.874**	0.865
満足度	0.834	**0.878**	0.816

ChatGPT によって生成された拡張データ内に指定したアスペクトに該当しないデータがどのアスペクトにおいても見られた．このことから，ChatGPT によって生成される拡張データは質が悪く，多量に追加することでアスペクト推定に悪影響となることが考えられる．

“熱中度”について，拡張データ割合が増加するにつれて推定精度が減少した．この傾向は“難易度”や“グラフィック”など多くのアスペクトにおいて見られた傾向である．しかし，これらのアスペクトは拡張データ追加前から高い推定精度を示しており，少量であっても拡張データがノイズとなってしまったと考えられる．そのため，精度の向上につながる拡張データを生成するために，プロンプトの改良や生成データの前処理の改善に取り組む必要がある．“満足度”について，＜少＞では精度が向上したが，＜多＞では精度が大きく減少した．＜少＞では満足度という抽象的なアスペクトを捉えるというタスクに，ChatGPT の生成した拡張データがデータの多様性に貢献したと考えられる．一方で，＜多＞では質の悪い拡張データによってノイズとしての悪影響が大きくなり，精度が低下したのではないかと考えられる．“音楽”のように，ほぼ精度が変化しないアスペクトも見られた．

6. おわりに

本研究では，ゲームレビュー文を対象とするアスペクト

推定の精度向上を目的として推定モデルの変更と，拡張データの追加に取り組んだ．複数の推定モデルを使用したアスペクト推定実験の結果から，事前学習済みモデルである BERT の有効性がすべてのアスペクトにおいて確認された．また，ChatGPT の推定結果から，SVM と同様に特徴的なアスペクト語が出現しやすい“グラフィック”などの推定は高精度である一方，アスペクトの粒度が広く広範なアスペクト語が出現する“オリジナリティ”などのアスペクト推定は低精度であるという特徴が見られた．拡張データ追加実験の結果から，ChatGPT によって生成された拡張データは多くのアスペクトにおいて悪影響を与えるノイズとなってしまうことが確認された．その理由としては ChatGPT によって生成される拡張データの内容が本来指定したアスペクトと異なっていることで，誤った正解データとして訓練されている事が考えられる．

今後の課題として，ChatGPT 以外の大規模言語モデルによって生成された拡張データを利用した場合の精度変化や，ChatGPT による拡張データが精度向上に寄与できるアスペクトの特徴や要件の分析，またはプロンプトの改良などが挙げられる．

参考文献

[1] Takuto Nakamuta and Kazutaka Shimada. Multi-aspects rating prediction using aspect words and sentences. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pp. 513–521, 2015.

[2] 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.

[3] An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. Generative data augmentation for aspect sentiment quad prediction. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pp. 128–140, 2023.

[4] Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 431–469, 2023.

[5] Bartłomiej Koptyra, Anh Ngo, Łukasz Radliński, and Jan Kocoń. Clarin-emo: Training emotion recognition models using human annotation and chatgpt. In *Computational Science – ICCS 2023: 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part I*, p. 365–379, 2023.

[6] Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. Solving aspect category sentiment analysis as a text generation task. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4406–4416, November 2021.

[7] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 19–30, 2016.

[8] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. p. 1941–1942, 2018.

[9] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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. 6382–6388, 2019.

[10] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 452–457, 2018.

[11] Masaki Takeo, Shinnosuke Kawasaki, and Kazutaka Shimada. Rating prediction of multi-aspect reviews using simultaneous learning. In *2023 International Conference on Asian Language Processing (IALP)*, pp. 358–363, 2023.

[12] Ryosuke Tadano, Kazutaka Shimada, and Tsutomu Endo. Multi-aspects review summarization based on identification of important opinions and their similarity. In *Proceedings of the 24th Pacific Asia conference on language, information and computation*, pp. 685–692, 2010.

[13] Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15465–15471, 2023.

[14] Frédéric Piedboeuf and Philippe Langlais. Is ChatGPT the ultimate data augmentation algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15606–15615, 2023.

[15] Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. Effects of diversity incentives on sample diversity and downstream model performance in llm-based text augmentation. *arXiv preprint arXiv:2401.06643*, 2024.