修士論文

文書と属性に対する感情分析のマルチタスク学習に関する研究

小泉さやか

主指導教員　白井清昭

北陸先端科学技術大学院大学
先端科学技術研究科
(情報科学)

令和 7 年 9 月

Abstract

In recent years, with the proliferation of e-commerce sites and social media, users have more opportunities to express their opinions about products and services. For companies providing products and services, as well as for people considering the purchase of those products or utilization of those services, reviews that include users' impressions and opinions serve as extremely valuable sources of information. Given this background, sentiment analysis, which is a task to classify subjective evaluations within text, has attracted significant attention across a wide range of fields, including marketing, recommendation systems, and customer satisfaction analysis. Among the various subtasks of sentiment analysis, Aspect-Based Sentiment Analysis (ABSA), which determines the polarity (positive, negative, or neutral) for specific aspects such as "price," "customer service," or "cleanliness," enables more fine-grained analysis of user opinions. However, to train models for ABSA, training data is required in which a polarity label is assigned to each aspect, and constructing such data demands substantial effort. In particular, for the Japanese language, the development of datasets for ABSA remains insufficient, and the lack of training data poses a significant barrier.

The goal of this research is to explore methods for ABSA with high accuracy in scenarios where only a limited amount of training data is available. Specifically, the study addresses the problem of data sparseness by leveraging a dataset labeled with polarity at the document level as auxiliary data. On many websites that facilitate user reviews, users can assign evaluation scores for target products or services. It is relatively easy to assign polarity labels to entire reviews based on users' scores; thus, a large-scale dataset can be relatively easily constructed. Regarding ABSA as the main task and document-level sentiment analysis as the auxiliary task, this research employs multi-task learning to improve the performance of ABSA by utilizing the information obtained from the auxiliary task's labeled data as external knowledge.

This study performs multi-task learning of aspect-based and document-level sentiment analysis. In our multi-task learning framework, BERT is used as an encoder shared across these two tasks. Furthermore, two models for multi-task learning are proposed. The first is a basic multi-task learning model, referred to as MTL-Basic. An input text is encoded by BERT, and the encoded output is then fed into two distinct classification layers corresponding to each task. Sharing the encoder enables learning of the classification model from diverse information contained in multiple datasets, which is expected to offer benefits such as suppression of overfitting and the enhancement of the generalizability of abstract representations of review texts. The second model is a multi-task learning model

that uses a shared intermediate layer, referred to as MTL-Shared. To further enhance information sharing between the two tasks, the intermediate layer shared between tasks is added after the BERT encoder. The integration of features common to both document-level and aspect-level sentiment analysis within this shared layer is expected to enhance the performance of both the main and auxiliary tasks.

In addition to the above two multi-task learning models, a filtering method is proposed to enhance the quality of the dataset used for document-level sentiment analysis, which serves as the auxiliary task. Only one polarity label is assigned to each document in the dataset of document-level sentiment analysis. It causes inconsistency between the content of a review and its polarity label. For example, even when both positive and negative opinions are included in a review, either a "positive" or "negative" label is assigned. Even the "neutral" label is assigned to such a review, as the reviewer may give a neutral score due to holding both positive and negative opinions across different aspects. To address this issue, reviews that contain both positive and negative expressions are removed from the dataset. Specifically, a sentiment lexicon is employed for the filtering. A review is discarded when both positive and negative words in the sentiment lexicon appear in it. After applying this filtering process, a polarity classification model is trained using MTL-Basic. This model is referred to as MTL-Refined.

Two datasets were used in the experiments to evaluate the effectiveness of the proposed method. For ABSA, the "Rakuten Travel Review: Aspect and Sentiment-tagged corpus" ("ABSA corpus" in short) was used. This dataset contains polarity labels for seven types of aspects, including "location," "room," "service," and so on. For document-level sentiment analysis, a collection of reviews posted on Rakuten Travel, which is a part of the Rakuten dataset, was used. Polarity labels were automatically assigned based on the 5-point rating scores provided by users at the time of review submission. For both datasets, balanced datasets are constructed by randomly selecting an equal number of positive, neutral, and negative samples. The balanced datasets consist of 10,000 samples in total. The balanced dataset of the ABSA corpus is not perfectly balanced as a number of samples of the minor class is less than one-third of 10,000. To conduct experiments, the datasets are subdivided into 70% training data, 20% development data, and 10% test data. In addition, to verify how the size of the training data influences the model's performance, the number of samples in the ABSA dataset was diminished. Specifically, we prepared datasets of four different sizes, i.e., 100%, 50%, 25%, and 10% of the original dataset. While the size of the training and development data is reduced, the test data remains constant to ensure a fair comparison of the models trained from the different-sized training data. Besides, the size of the dataset for document-level sentiment analysis

is not decreased, considering that it can be relatively easily constructed. The performance of ABSA models was evaluated using the accuracy and F1 score. A single-task learning model was used as the baseline and compared with the three proposed multi-task learning models.

Experimental results showed that there was no significant difference in performance between the multi-task learning models and the single-task learning model when the training data was sufficiently large. On the other hand, when the size of the training data was limited to 10%, the multi-task learning models outperformed the single-task model. In particular, for polarity labels that are difficult to classify, such as “negative” and “neutral”, MTL-Basic showed improvements of up to 3 points in both the accuracy and F1 score compared to the single-task learning model. Moreover, multi-task learning models demonstrated smoother transitions in the accuracy and F1 score during training, with a tendency to attain high performance even in the early stages of training (with fewer epochs). MTL-Shared was effective for some aspects, but showed a strong bias toward the “positive” label, i.e., MTL-Shared classified many test samples as positive. MTL-Refined improved the accuracy for the “neutral” label, although improvements for the other labels were limited. A case study analyzing correctly and wrongly predicted samples revealed that the multi-task learning models could correctly predict polarity even for short reviews or reviews including euphemistic expressions.

Future work will focus on further improving the flexibility and performance of multi-task learning models through dynamic adjustment of task loss weights, refinement of parameter-sharing strategies between tasks, and incorporation of attention mechanisms. Furthermore, efforts will be made to enhance the interpretability of ABSA models and to explore their practical applications.

概要

近年は EC サイトや SNS の普及により、利用者が商品やサービスに対して意見を発信する機会が多くなっている。商品を提供している企業や、購入・サービスの利用を検討している人々にとって、利用者の感想や意見を含むレビューは極めて有用な情報源である。こうした背景のもと、テキストにおける主観的評価を分類する「感情分析」は、マーケティング、推薦システム、顧客満足度の分析など幅広い分野で注目されている。中でも、評価対象の特定の属性（「価格」「接客」「清潔さ」など）ごとに感情の極性（肯定・否定・中立）を判定する「属性に対する感情分析（Aspect-Based Sentiment Analysis: ABSA）」は、利用者の意見のより精緻な分析を可能にする。しかし、属性に対する感情分析のモデルを学習するには属性に対して極性ラベルが付与された訓練データが必要だが、その作成には多大な労力を要する。特に日本語では属性に対する感情分析のデータセットの整備が十分に進んでおらず、訓練データの不足が大きな障壁となっている。

本研究の目的は、少量の訓練データしか利用できない状況下で高い精度で属性に対する感情分析を行う方法を探求することにある。具体的には、文書全体に対して極性ラベルが付与されたデータを補助データとして活用することで訓練データ不足の問題を解決する。多くのレビューサイトでは、利用者がレビューを投稿する際、製品やサービスの評価スコアを与えることができる。この評価スコアを基にレビュー全体に極性ラベルを付与することで、大規模なデータセットを比較的容易に構築できる。属性に対する感情分析を主タスク、文書全体に対する感情分析を補助タスクとするマルチタスク学習を行い、補助タスクのラベル付きデータから得られる情報を補完することで、属性に対する感情分析の性能を向上させることを狙う。

本研究では、BERT を共通エンコーダとして用い、属性に対する感情分析のモデルと文書全体に対する感情分析のモデルを同時に学習するマルチタスク学習のアーキテクチャとして 2 つのモデルを提案する。一つ目は基本マルチタスク学習モデル MTL-Basic である。入力テキストを BERT でエンコードし、その出力をそれぞれのタスクに対応する独立した分類ヘッドの入力とする。エンコーダを共有することで複数のタスクのデータセットから得られる多様な情報を統合的に学習でき、過学習の抑制やレビュー文の抽象表現の汎化といった利点が期待される。二つ目は中間共有層を用いるマルチタスク学習モデル MTL-Shared である。両タスク間の情報共有をより強化するために、BERT によるエンコーダ出力の後にタスク間で共有する中間層を追加する。文書に対する感情分析と属性に対する感情分析とで共通する特徴をこの中間層に反映させることで、主タスクならびに補助タスクの性能の向上が期待できる。

上記 2 つのマルチタスク学習モデルに加えて、補助タスクである文書に対する感情分析のデータセットの品質を向上させるためのフィルタリング手法を提案する。同データセットはレビュー文書全体に対して極性ラベルが 1 つ付与されるため、レビュー内容とその極性の間に不整合が生じる場合がある。例えばレビュー対象に肯定的な感想と否定的な感想の両方が含まれていても「肯定」「否定」いずれかの極性ラベルが付与される。また、レビュー投稿者はそのような場合に中立的なスコアをつける傾向にあり、このときにはレビューが肯定的・否定的な感想を両方含んでいても「中立」のラベルが付与される。この問題への対応として、レビュー文書が肯定的な表現と否定的な表現の両方を含むとき、そのレビュー文書をデータセットから削除する。具体的には、フィルタリングには日本語評価極性辞書を用い、肯定語と否定語の両方がレビュー文中に出現した場合、当該データは訓練データセットには含めない。上記のフィルタリング処理を実施した後、基本的なマルチタスク学習モデルによって極性判定モデルを学習する。このモデルを MTL-Refined と呼ぶ。

評価実験では 2 つのデータセットを使用した。属性に対する感情分析のデータセットとして「楽天トラベルレビュー: アスペクト・センチメントタグ付きコーパス」を用いた。このデータセットには「立地」「部屋」「サービス」など 7 種類の属性に対する極性が付与されている。また、文書全体に対する感情分析のデータセットとして「楽天データセット」のうち楽天トラベルに投稿されたレビューの集合を用いた。利用者がレビュー投稿時に記載した 5 段階評価スコアをもとに極性ラベルを自動付与した。いずれのデータセットも「肯定」「否定」「中立」のラベル数が均等になるように 10,000 件をサンプリングした。また学習データ量の違いによる性能変化を検証するため、テストデータの量は変えずに訓練・検証データの量を元データの 100%、50%、25%、10%の 4 条件に設定した。モデルの性能は正解率および F 値により評価した。シングルタスク学習モデル(STL)をベースラインとし、提案手法の 3 種類のマルチタスク学習モデルと比較した。

実験の結果、訓練データ量が十分に大きい条件下では、マルチタスク学習モデルとシングルタスク学習モデルの間に大きな性能差は見られなかった。一方で、訓練データ量を 10%に制限した条件では、マルチタスク学習モデルがシングルタスク学習モデルを上回る性能を示した。特に、否定や中立といった分類が難しい極性ラベルに対して、MTL-Basic の正解率および F 値は STL と比べて最大 3 ポイント向上した。また、マルチタスク学習モデルでは学習過程における F 値や正解率の推移が滑らかであり、エポック数が少ない段階から高い F 値や正解率が得られる傾向が見られた。MTL-Shared は一部の属性に対しては有効であったが、全体としては肯定ラベルに強くバイアスされる(肯定を予測す

る割合が極端に多くなる)傾向があった。MTL-Refined は、中立ラベルに対する正解率を改善する効果が見られたものの、他のラベルに対する改善は限定的であった。さらに、各モデルが正解または不正解だった事例を分析し、モデルの特徴を考察した。マルチタスク学習モデルが短文や婉曲的な表現を含むレビューに対しても正しい極性を予測できたケースが確認された。

今後は、マルチタスク学習における 2 つのタスクの損失の重みの動的な調整、タスク間のパラメータ共有戦略の改善、注意機構の活用などを通じて、より柔軟かつ高性能なマルチタスク学習モデルの実現を目指す。さらに、属性に対する感情分析モデルの解釈性の向上や実用的応用への展開にも取り組む。

目次

図目次

表目次

第1章 はじめに

本章ではまず 1.1 節で本論文が研究の対象とする感情分析が必要とされる背景や感情分析のタスクの種類について紹介する。さらに感情分析のタスクにおける現在のデータセットに関する問題についても触れる。次に 1.2 節で本研究の目的とこの問題に対するアプローチについて説明する。最後に 1.3 節で本論文の構成を述べる。

1.1 背景

近年 EC サイトや SNS の普及に伴い、利用者が商品やサービスに対して自由に意見を発信する機会が多くなっている。商品を提供している企業や、商品の購入やサービスの利用を検討している人々にとって、こうしたレビューは利用者の生の声として極めて有用な情報源である。こうした背景から、近年ではレビュー文の中から人々の感情を自動的に抽出・分類する「感情分析（Sentiment Analysis)」が注目されている。

感情分析は大きく分けて 2 つのアプローチに分類される。一つは、文書全体に対して極性(ポジティブまたはネガティブ)を推定する「文書全体に対する感情分析」であり、全体的な印象をざっくり把握するのに適している。もう一つは、評価対象の特定の属性（「価格」「接客」「清潔さ」など）に対して極性を推定する「属性に対する感情分析（Aspect-Based Sentiment Analysis, ABSA)」である。属性に対する感情分析はより詳細な分析を可能にするため、マーケティング、推薦システム、カスタマーサポートなど多様な応用が期待されている。

しかし、属性に対する感情分析を機械学習モデルで実現するには、各属性に対して適切な極性ラベルを付与した教師データが必要となる。このようなデータセットを作るには、レビューを書いた人に属性毎の評価を答えさせるアンケートを実施するか、第三者がレビューを読んで属性毎の極性を推測してラベルを付与する必要がある。そのため大規模なデータセットを作成するには多大なコストと労力を必要とする。特に日本語においては、公開されている属性に対する感情分析のデータセットの種類が少ないことも、属性に対する感情分析を難しくしている原因である。

1.2 本研究の目的

本研究では、属性に対する感情分析の訓練データが不足している問題に対処

するために、文書全体に対する感情分析のタスクを補助タスクとして併用するマルチタスク学習モデルを提案する。文書全体の極性ラベルは、アノテーションの粒度が粗いために大量に取得しやすい。一方で、属性ごとの極性ラベル付与は細かい判断を要し、ラベル付与の工数が高くなりがちである。この性質の違いを活かし、2 種類のデータセットを併用してモデルを学習させることで、属性に対する感情分析の性能向上を狙う。

具体的には、Bidirectional Encoder Representations from Transformers (BERT)といった事前学習済み言語モデルをベースとし、文書全体に対する極性ラベルと属性に対する極性ラベルの両方を予測するマルチタスク学習を試みる。エンコーダを共有しつつ、各タスクに専用の出力ヘッドを設けることで、それぞれの目的に応じた予測を可能にしつつ、共有部分では文脈情報や語彙的知識の学習効率を高める。また、単なるエンコーダの共有だけでなく、中間層の共有方法など複数のバリエーションを検討し、属性に対する感情分析タスクにとって有効なマルチタスク学習モデルのアーキテクチャを模索する。

本研究の最終的な目的は、文書全体に対する感情分析から得られる汎用的な知識を活かし、少量の教師データしか利用できない状況でも高い精度で属性に対する感情分析を実現するモデルを構築することにある。

1.3 本論文の構成

本論文の構成は以下の通りである。2 章では、感情分析とマルチタスク学習に関する既存の研究を概観し、本研究の立ち位置と意義を明らかにする。3 章では、提案手法の全体像を提示する。具体的には、問題設定、ベースとなるマルチタスク学習モデル、特徴共有による拡張モデル、訓練データのフィルタリングについて述べる。4 章では、提案モデルの評価実験の設定と結果の詳細を述べ、ベースラインモデルと提案モデルの性能比較や事例分析を通して本手法の有効性を検証する。5 章では、本研究の成果をまとめ、今後の課題や展望について考察する。

第2章 関連研究

本章では、感情分析およびマルチタスク学習に関する先行研究を整理し、本研究の立ち位置と特色を明確にする。

まず 2.1 節では、感情分析の基本的な枠組みとして、文書全体に対する感情分析と属性に対する感情分析の 2 種類が存在することを述べ、それぞれの先行研究を概観する。2.2 節では、マルチタスク学習の概念と自然言語処理の分野における活用事例を紹介し、感情分析分野においても補助タスクを併用する形でのマルチタスク学習の適用が進んでいる現状を示す。これらの議論を踏まえ、2.3 節では本研究の特色を述べる。

2.1 感情分析に関する研究

感情分析は、テキストから書き手の主観的な感情や態度を抽出・分類する自然言語処理の代表的な応用タスクの一つである。レビュー文の分析、SNS 上の世論把握、チャットボットの応答制御など、実社会において多様な応用が進んでいる。初期の感情分析では感情語辞書を用いた手法やルールベースの手法が主流であり、文中の肯定的・否定的な語の出現に基づいて極性を判断するアプローチが採用されていた。しかし、これらの手法では文脈の情報を十分に捉えられず、否定表現や皮肉のような微妙な感情の扱いに限界があった。

その後、BoW (Bag-of-Words) や TF-IDF といった統計的特徴量をベースとする機械学習モデルが登場し、サポートベクターマシン (Support Vector Machine, SVM) やロジスティック回帰などによって感情分析の性能が大幅に向上した。2010 年代後半には、回帰型ニューラルネットワーク(Recurrent neural network、RNN) や Long Short-Term Memory(LSTM)といった系列モデルを用いた感情分析の研究が進み、感情の時間的推移といった文脈情報や文内の依存関係を考慮した分類が可能となった。さらにその後、事前学習済み言語モデル (Pre-trained Language Model, PLM) が登場した。特に、BERT [1]や RoBERTa [2]といったトランスフォーマーをベースとしたモデルは、双方向文脈の表現力に優れ、感情分析のタスクの精度を大きく押し上げた。

従来の感情分析の研究では、文書全体に対して単一の極性 (肯定/否定/中立) を予測する「文書全体に対する感情分析」が中心であった。このアプローチは、文書全体に一貫した感情が含まれている場合には有効であり、タスク設定も比較的単純で、訓練データ収集も容易である。しかし、実際のレビュー文や SNS の投稿には複数の評価対象 (例：価格、サービス、清潔さなど) が混在している。

たとえば「料理は美味しかったが、接客はひどい」のように、同一文内に相反する感情が存在することも多い。このような文書に対して単一の極性ラベルのみを推定するアプローチでは、詳細な分析が難しく、実用上の限界がある。

この問題を解決するために登場したのが属性に対する感情分析である。属性に対する感情分析では、テキスト中に現れる属性（サービス、バッテリー、デザインなど）ごとにその極性を個別に推定する。属性抽出（Aspect Term Extraction)、属性カテゴリ分類(Aspect Category Detection)、感情分類(Aspect Sentiment Classification) の三段階に分けられることが多い。個々の属性毎にユーザーの評価の極性を推定することで、ユーザーの発言に含まれる多次元的な評価を精密に捉えることが可能になる。

属性に対する感情分析は SemEval 2014 Task 4 をはじめとした国際的な評価型ワークショップを中心に通じて発展してきたが、その一方で、訓練データセットのラベル付けにおけるアノテーションコストの高さが常に課題であった。特に、複数の属性に対して極性ラベルを付与する作業は煩雑であり、大規模なデータセットの整備が難しい。公開されているデータセットとしては、SemEval のレストラン／ラップトップレビューや、MAMS (Jiang et al. 2019)、Twitter-ABSA などがあるが、その数と多様性は文書全体に対する感情分析のデータセットと比較して限られている。このようなデータセット構築の難しさを克服するため、データ拡張や転移学習、マルチタスク学習といった「補助的学習」のアプローチが提案されてきた。

なお、近年では、生成 AI の発展が感情分析の研究に対して大きなインパクトを与えている。GPT-3[3]や ChatGPT、T5[4]などの大規模言語モデルは、従来のラベル予測を前提とした構造とは異なり、プロンプトベースの自然言語指示に基づいて感情の表現や評価を生成する能力を備えている。たとえば、「この文における‘サービス’に対する筆者の態度を述べてください」といったプロンプトを入力するだけで回答が得られる。それでもセキュリティや推論速度などの難しさから、BERT などの事前学習済み言語モデルをベースとした従来型のアーキテクチャの有効性も根強く残っている。また、弱教師付きまたは自己教師あり学習の枠組みで、明示的なアノテーションを必要としない属性に対する感情分析も一部で実用化されつつある。

2.2 マルチタスク学習に関する研究

マルチタスク学習 (Multi-Task Learning: MTL) は、複数の関連タスクを同時に学習させることで、各タスクの性能向上や学習効率の改善を図る枠組みである。自然言語処理の分野においては、特にラベル付きデータの偏在やドメインご

とのデータ不足といった実用上の制約を補う手法として注目されてきた。

現在では、BERT をはじめとする事前学習済み言語モデルを共通エンコーダとして用い、その上にタスク固有の出力層を付加する設計が一般的である。このような構成により、文脈を考慮した文の抽象表現を得るためのエンコーダは複数のタスクで共有されつつ、それぞれの出力層は独立して最適化される。He らは、相互作用的なデコーダ機構を組み込んだマルチタスク学習構成を提案し、複雑なタスク間依存を動的に捉えるアーキテクチャを実現している[5]。

ただし、こうしたパラメータ共有には限界もある。異なるタスクが求める情報の粒度や注目範囲がずれている場合、無条件のパラメータ共有はかえって性能を劣化させる場合がある。そのため、層ごとに共有・分離を調整する部分共有（Partial Sharing）や、GradNorm [6]のように勾配ノルムを動的に補正する手法が提案されている。なお、文書全体に対する感情分析と特定の属性に対する感情分析のように、注目する文脈が異なるタスクを並行して学習させる場合、パラメータ共有が性能の悪化を招く傾向が顕著になる可能性がある。

さらにタスクの出力分布や損失スケールの違いに対応するため、学習スケジューリングや重み制御も重要な設計要素である。Jia らは、複数タスクを統一的な空間で扱う「ALL-IN-ONE」型の構成を提案し、異なる出力タスクを効率的に共学習させるアーキテクチャを構築している[7]。また彼らはマルチタスク学習の利点を過学習抑制の観点から言及している。単一タスクでの学習では、特定のパターンへの過度な適応が起こりやすいが、複数タスクの共学習によって多様な学習信号が導入され、モデルの汎化能力が高まる。

感情分析の分野では、文書全体に対する感情分析の極性ラベルは比較的収集しやすい一方で、属性に対する感情分析は人手アノテーションのコストが高く、大規模データセットの確保が難しい。こうした非対称性を解消する手法として、Zhang らは文書全体に対する感情分析と属性に対する感情分析を同時に扱い、擬似ラベルの生成と相互活用により粒度の異なる感情情報を統合する Dual-granularity Pseudo Labeling（DPL）を提案した[8]。

一方で、日本語における研究は属性に対する感情分析の研究は限定的である。とりわけ日本語のデータセットに対して属性に対する感情分析と文書全体に対する感情分析を明示的に統合するマルチタスク構成の事例は存在しない。張らは、日本語の属性に対する感情分析における属性カテゴリ検出に対して補助文を導入することで精度向上を実現した[9]。文書全体に対する感情分析の結果を補助文として活用することで属性に対する感情分析の性能と頑健性を向上させる枠組みは注目に値する。ただし、この研究は属性に対する感情分析単体に焦点を当てたものであり、文書全体に対する感情分析との統合学習を試みた研究ではない。

2.3 本研究の特色

本研究は、日本語における属性に対する感情分析と文書全体に対する感情分析の同時学習によるマルチタスク学習を検討し、その効果を分析する。特に、文書全体に対する感情分析と属性に対する感情分析を統合的に扱うマルチタスク学習の枠組みは、英語圏では一定の研究例が存在する一方で、日本語においては十分に検証されておらず、先行研究が限られている。このような状況を踏まえ、本研究は日本語を対象に属性に対する感情分析に対するマルチタスク学習の有効性を探究する。

本研究で用いたデータセットはいずれも楽天トラベルのレビューを基にしているが、文書全体に対する感情分析にはユーザーが入力した全体評価スコアを基にした極性ラベル、属性に対する感情分析には人手でアノテーションされた属性ごとの極性ラベルを用いている。両タスクは同一ドメインにありながら情報構造が異なるため、マルチタスク学習によって相補的な特徴抽出が期待される。

また本研究では、BERT ベースの共通エンコーダを軸としつつ、単純なタスク併用モデルから中間層共有型モデル、さらには訓練データのフィルタリングまで、複数のマルチタスク学習モデルを比較検証する。特に訓練データ量を制限した条件下において、補助タスクとしての文書全体に対する感情分析が属性に対する感情分析の性能向上に寄与するかを詳細に分析することで、リソース制約の大きい日本語の属性に対する感情分析についての実用的な知見を得る。

第3章 提案手法

本章では、属性に対する感情分析の性能向上を目的としたマルチタスク学習の手法について述べる。文書全体の感情分析のための大規模な訓練データが存在し、属性ごとの感情分析のための小規模な訓練データが存在するという状況を仮定し、その状況下において属性に対する感情分析の性能向上を図る。

まず 3.1 節では本研究が対象とするタスク設定と訓練データの前提条件について整理する。二種類の感情分析タスクの定義とその難易度の差、ならびに本研究における問題意識を明確にする。

3.2 節では、文書全体の感情分析と属性に対する感情分析を同時に学習する基本的なマルチタスク学習モデルを提案する。本モデルは BERT ベースのエンコーダを中心に構成され、文書全体の感情分析と属性に対する感情分析を並列に行う構造を持つ。

3.3 節では 3.2 節のモデルを拡張し、タスク間で共有する中間層を導入する。これにより両タスク間でより効果的に情報を伝達し合い、属性に対する感情分析の精度のさらなる向上を図る。

最後に 3.4 節では、文書全体の感情分析タスクの訓練データのフィルタリング手法について述べる。具体的には、文章全体の感情分析のデータセットから肯定的意見と否定的意見の両方を含むサンプルを訓練データから除外することで、属性に対する感情分析モデルに対する悪影響を軽減することを狙う。

3.1 問題設定

本研究では、レビュー文を対象にした感情分析タスクを扱う。主たるタスクは、文中の特定の属性（例：価格、品質、デザインなど）ごとに個別の極性ラベルを推論する「属性に対する感情分析」である。一方、文書全体に対して極性ラベルを推論する「文書全体に対する感情分析」は、主タスクの学習を支援する補助的な役割として位置付けている。補助タスクとして文書全体の感情分析を活用することで属性に対する感情分析の性能向上を図ることを目的としている。

両タスクの大きな違いのひとつは訓練データセットの取得容易性である。文書全体に対する極性ラベルは、レビュー投稿時の評価スコアを転用することで大量の事例を比較的容易に収集できる。これに対し属性に対する極性ラベルを作成するには、属性ごとの評価項目をアンケート等でユーザーに回答してもらうか、文中に記載された属性を抽出しそれぞれに対して極性ラベルを付与するアノテーションを行う必要がある。これらの作業は難易度が高く、十分な量のラ

ベル付きデータを確保することが困難であることが多い。したがって、文書全体に対する感情分析のラベル付きデータセットは豊富だが、属性に対する感情分析のラベル付きデータは少ないといった非対称な状況がしばしば発生する。本研究ではこのような状況下において、属性に対する感情分析モデルを学習することを問題設定とする。このような問題設定のもと、属性に対する感情分析の性能をいかにして高めるかが中心的な課題となる。

具体的には、両タスクに共通する特徴を活用するために、マルチタスク学習の枠組みを導入する。文書全体の極性に関する推論モデルを補助的に学習しながら、主目的である属性に対する感情分析の性能向上を目指す。本研究におけるマルチタスク学習とは、単一のエンコーダによって入力データであるレビュー文の特徴を抽出し、それを文書全体に対する感情分析タスクおよび属性に対する感情分析タスクの両方で共有しつつ、各タスク固有の予測器によって推論されたラベルを出力するという構成を指す。このような設計により、豊富な文書全体の感情分析データセットから学んだ情報を属性に対する感情分析にも間接的に活用できると期待される。

3.2 基本的なマルチタスク学習モデル

本節では、提案手法における基本的なマルチタスク学習モデルの構成について述べる。本モデルのアーキテクチャを図 3.1 に示す。このモデルは、文書全体に対する感情分析タスクと、属性に対する感情分析タスクを同時に扱うことを目的として設計されている。アーキテクチャは大きく三つの構成要素からなる。まず、入力テキストに対して BERT のエンコーダを適用し、テキストの文脈埋め込み表現を取得する。次に、エンコーダの出力を入力とし、文書全体に対する感情を判定する分類ヘッドと、属性に対する感情を判定する分類ヘッドを並列に置く。いずれの分類ヘッドも、Feed-forward Network (FFN)から構成され、文脈埋め込み表現から極性ラベルを推論する。

学習時の訓練データセットは文書全体の感情分析タスクのレビュー文および極性ラベルと、属性に対する感情分析タスクのレビュー文および極性ラベル両方が含まれるようにデータを構成する。バッチサイズや学習率などのハイパーパラメータの決定については 4.2 節にて後述する。文書全体の感情分析タスクと属性に対する感情分析タスクのそれぞれに対しクロスエントロピー損失を用いて学習を行い、最終的な損失は両タスクの損失の和とする。ここではタスク間の重要度を等価とみなし、単純な加算によって全体損失を計算している。学習中に動的に重みを調整する手法を導入する余地もあるが、本研究では最も単純な方法を採用する。

本モデルの主な利点としては、BERT エンコーダの出力をタスク間で共有することにより文書レベルと属性レベルの相補的な情報を活用できる点が挙げられる。さらに先行研究[7]によると、複数のタスクの損失を減少させる構成にすることは一方のタスクの過学習を防ぐ効果もある。また、モデルのパラメータ数を抑えながら複数の出力を得られる点も実用的なメリットである。

一方で、同一のエンコーダを複数タスクに共有する設計には限界も存在する。たとえば、文書全体の文脈把握と属性に対する局所的判断とでは必要とされる情報が異なるため、共有された特徴量が一方の性能を犠牲にするリスクがある。

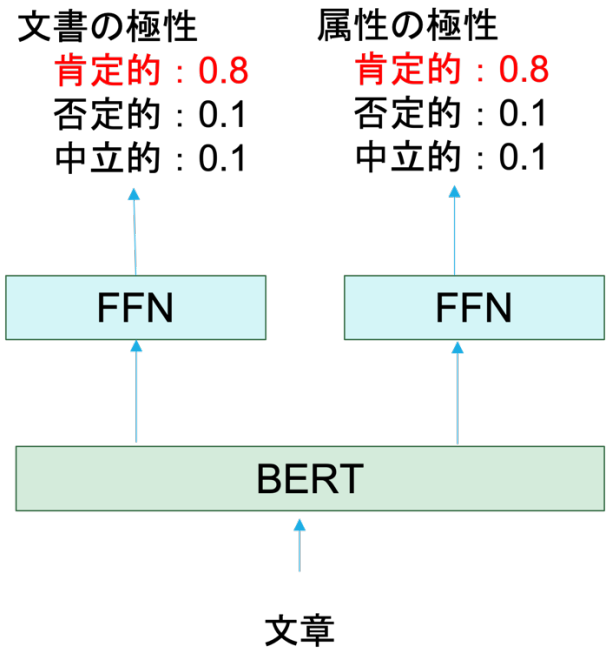

図 3.1 基本的なマルチタスク学習モデル

3.3 中間共有層を用いるマルチタスク学習モデル

前節では、BERT エンコーダの上に文書全体の感情分析および属性に対する感情分析それぞれに対応した専用の出力ヘッドを個別に接続する基本的なマルチタスク学習モデルを紹介した。本節ではそれを発展させ、エンコーダの出力と専用の出力ヘッドの間に中間的な共有層を挿入する構成を提案する。これを図 3.2 に示す。

この特徴共有型モデルでは、入力文に対し BERT エンコーダが文脈的なベクトル表現を生成した後、その出力を共通の全結合層(Feed Forward Network, FFN)に通す。この結合層は、文全体の意味や構造に関する特徴を強調しつつ、タスク固有の表現に変換しやすい形式へ整形する役割を担う。その後、この結合層の出

力を用いて、それぞれのタスク専用の出力ヘッドで極性ラベル分類を行う。

この構成により、単なるエンコーダの重み共有だけでは実現しにくい中間的な情報抽象化・強調が可能となる。特に属性に対する感情分析にとっては、文脈理解や表現の柔軟性が強化されることで、少量データでも有益な特徴が抽出されやすくなると考えられる。加えて、文書全体の感情分析と属性に対する感情分析の両方で意味的に類似した判断が求められる場面（例：否定表現、比較、皮肉など）において、共有結合層がその共通パターンを学習することが期待される。

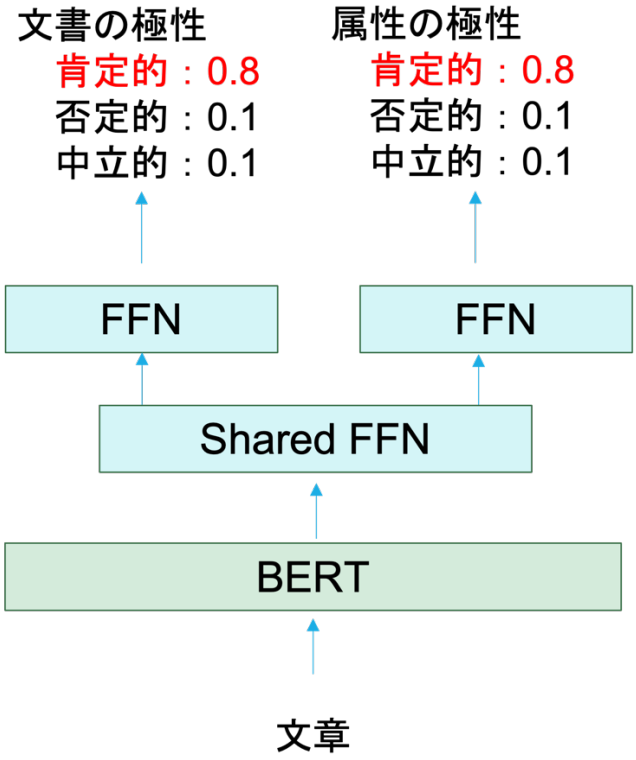

図 3.2 中間共有層を用いるマルチタスク学習モデル

3.4 訓練データのフィルタリング

マルチタスク学習において、補助タスクである文書全体の感情分析の訓練データセットの品質は、主タスクである属性に対する感情分析の性能にも大きな影響を与えると考えられる。具体的には、文書全体の感情分析の訓練データにおいて、2 つの極性が混在している事例や極性が不明瞭な事例は、属性に対する感情分析への知識転移に悪影響を及ぼす可能性がある。

たとえば、「料理は美味しい。接客はひどい。」といったレビュー文は料理に対しては肯定的、接客に対しては否定的な意見を示しているため、文章全体の感情は肯定的であるとも否定的であるとも判断できない。文書全体の感情分析の目

的はレビュー文に対しひとつの極性ラベルを推論することであるため、教師ラベルは「中立」になる。しかし、マルチタスク学習の過程でレビュー文中の肯定的な表現「美味しい」と否定的な表現「ひどい」が「中立」のクラスと関連付けられることにより、モデルが擬似相関を学習し、このことが主タスクのモデルの学習にも悪影響を及ぼす可能性がある。

この問題に対処するため、本研究では文書全体の感情分析の訓練データに対するフィルタリングを提案する。具体的には、日本語評価極性辞書:用言編(用言辞書)[10]を用いて各文に含まれる極性語を抽出し、肯定語と否定語が同時に出現しているレビュー文を検出する。表 3.1 は日本語評価極性辞書から極性語を一部抜粋したものであり、表 3.2 および表 3.3 は肯定語と否定語が同時に出現しているレビュー文の例である。こうしたレビュー文は文書全体の感情分析の極性が曖昧であると判断し、訓練データから除外する。最終的に、文章全体の感情分析タスクの訓練データはフィルタリング処理後の極性の曖昧さが少ない事例から構成する。

表 3.1 日本語評価極性辞書[10](抜粋)

否定語	肯定語
あきらめる	すご腕 と
あきる	すご腕 の
あきれる	すすどい
あきれる た	すばやい
あせる	すばらしい
あなどる	すべすべ
	すらり

表 3.2 肯定語と否定語が同時に使用されているレビューの例 1

レビュー文	部屋もきちんと掃除されてる感じで お風呂もゆったり家族で入れて満足でした。 【ご利用の宿泊プラン】 【素泊まりプラン】無料貸切天然露天風呂でゆったり・・・。遅いチェックインも可能です。洋室、和室選択可 気軽に洋室
抽出語	[('ネガ（評価）', '遅い'), ('ポジ（評価）', 'ゆったり'), ('ポジ（評価）', '満足')]

表 3.3 肯定語と否定語が同時に使用されているレビューの例 2

レビュー文	駅ちかで、設備も部屋もきれい、大満足です。今回は、楽天ポイントもついた上に、宿泊代も安く、すごいよかった！ひとつだけリクエストあるとすれば、朝ごはんに、お米あったらうれしいなあ。パンだけなので、米派としては、少し物足りない。。。
抽出語	[('ネガ（評価）', '物足りない'), ('ポジ（評価）', '大満足'), ('ポジ（評価）', '満足')]

第4章 評価実験

本章では、本研究で提案したマルチタスク学習モデルの有効性を検証するため、複数の条件下において感情分析タスクの評価実験を行う。

4.1 節では、実験に用いたデータセットの構成と特徴について述べる。属性に対する感情分析のデータセットとして、「楽天トラベルデータセット」と「ABSA タグ付きコーパス」の 2 つを用いる。

4.2 節では、実験環境、使用モデル、ハイパーパラメータ設定、評価指標など、実験の設定について説明する。

4.3 節では、楽天トラベルデータセットを用いた実験結果を報告し、その結果を考察する。

4.4 節では、ABSA タグ付きコーパスを用いた実験について報告する。訓練データ量を段階的に減少させる実験を行い、マルチタスク学習の頑健性と有効性を評価する。

4.5 節では、事例分析を行い、マルチタスク学習がどのような文に対して改善効果をもたらすのか、あるいは逆に性能を劣化させるのか考察する。

4.1 データセット

本研究では、文書全体に対する感情分析と属性に対する感情分析のマルチタスク学習を行うにあたり、それぞれに対応したデータセットを構築・使用する。構築したデータセットの一覧を表 4.1 に示す。いずれも楽天データセット[11]を基盤とする。同データセットは、楽天トラベルに投稿された宿泊施設に対するレビューから構成される。ただ、使用するレビュー文およびラベル付与の方法はタスクによって異なる。

まず、文書全体に対する感情分析タスクの正解ラベルとして用いたのは、楽天データセットにおいて利用者が投稿時に入力した宿泊施設への総合的な評価である。楽天トラベルではユーザーは 1〜5 の星の数によって総合評価を投稿する。本実験では、星 1・2 を「否定」、星 3 を「中立」、星 4・5 を「肯定」として極性ラベルを割り当てた。次に、楽天データセットから実験に使用するレビューを 10,000 件抽出した。この際、以下の 3 つの抽出方法で実験データを構築した。

- **ランダム**: 10,000 件のレビューをランダムに抽出する。
- **均衡**: 極性ラベル(肯定・中立・否定)それぞれのサンプルが同数となるようにレビューをランダムに抽出した。
- **フィルタリング**: 3.4 節のフィルタリングを行った後、肯定・中立・否定ラ

ベルのサンプルを同数ランダムに抽出した。
抽出後のデータは、訓練：検証：テストの割合が 7：2：1 になるよう分割した。

一方、属性に対する感情分析については、2 種類の極性ラベルを使用した。一つめは、楽天データセットにおいて利用者が投稿時に入力した宿泊施設への属性別の評価である。楽天データセットにおける宿泊施設の属性は、立地、部屋、食事、浴場、サービス、アメニティの 6 種類である。これらの属性についても総合的評価と同様に 5 段階の評価がされているので、星 1・2 を「否定」、星 3 を「中立」、星 4・5 を「肯定」とする極性ラベルを割り当てた。次に、実験に使用するレビューを属性毎に 10,000 件抽出した。文書全体に対する感情分析のデータセットと同様に、ランダムにレビューを抽出する「ランダム」と、極性ラベルのサンプル数が均等になるように抽出する「均衡」のデータセットを作成した。その後、10,000 件のサンプルを訓練：検証：テストの割合が 7：2：1 になるよう分割した。なお、4.3 節にて後述するように、このデータセットでは、レビュー本文と属性別スコアの内容が一致していない、またはレビュー文内に該当の属性の記述が存在しないにも関わらず星 5 の評価が付与されているなど、いくつかの問題点がある。

二つめの極性ラベルは楽天トラベル:レビューアスペクト・センチメントタグ付きコーパス[12] によるものである。以下、これを「ABSA タグ付きコーパス」と呼ぶ。このデータセットでは、楽天トラベルに投稿されたレビュー文に対し、それが言及している属性と、その属性に対する極性(肯定、否定)の 2 種類のラベルが付与されている。実験の際、極性ラベルが付与されていないデータについては中立のラベルを付与した。ABSA タグ付きコーパスは記述内容と感情の整合性が高く、属性ごとの多様な感情表現を含む高品質なデータである。当データセットの属性は立地、部屋、朝食、夕食、浴場、サービス、アメニティの 7 種類である。極性ラベルのサンプル数が均等になるように属性ごとに 10000 件を抽出し、訓練：検証：テストの割合が 7：2：1 になるよう分割した。さらに少ない訓練データ数におけるモデルの性能を調査するため、属性ごとに訓練・検証データの量をそれぞれ 50%、25%、10%へ削減したデータセットも作成した。なお、異なる量の訓練データから学習されたモデルの感情分析の性能を比較するため、訓練・検証データ量を減らしたデータセットでもテストデータの量は減らしていない。

楽天トラベルデータセットから作成した均衡データセットでは「肯定」「中立」「否定」のレビューが同数になるようにサンプリングした。しかし、ABSA コーパスでは極性ラベルの分布の偏りが著しいため、3 つの極性クラスのレビュー数を同じにすることができない場合があった。このとき、頻度の少ないラベルのレビュー文は全て選択し、残りのラベルのレビューは同じ数になるようにサンプ

リングした。例えば、全体のレビュー件数を 10000 件と設定したとき、否定ラベルのレビューが 2000 件しかない場合には、それを全てデータセットに含め、中立ラベルと肯定ラベルのレビューを 4000 件ずつランダムに選択した。

表 4.1 データセットの統計

略称	単位	属性一覧	抽出方法	訓練	検証	テスト
D-RD-Random	文書	-	ランダム	7000	2000	1000
D-RD-Balanced	文書	-	均衡	7000	2000	1000
D-RD-Filtering	文書	-	フィルタリング	7000	2000	1000
A-RD-Random	属性	立地, 部屋, 食事, 風呂, サービス, アメニティ	ランダム	7000	2000	1000
A-RD-Balanced	属性	立地, 部屋, 食事, 風呂, サービス, アメニティ	均衡	7000	2000	1000
A-ABSA-ALL	属性	立地, 部屋, 朝食, 夕食, 風呂, サービス, アメニティ	均衡	7000	2000	1000
A-ABSA-1/2	属性	立地, 部屋, 朝食, 夕食, 風呂, サービス, アメニティ	均衡	3500 (50%)	1000 (50%)	1000
A-ABSA-1/4	属性	立地, 部屋, 朝食, 夕食, 風呂, サービス, アメニティ	均衡	1750 (25%)	500 (25%)	1000
A-ABSA-1/10	属性	立地, 部屋, 朝食, 夕食, 風呂, サービス, アメニティ	均衡	700 (10%)	200 (10%)	1000

* データセットの略称における「RD」は楽天データセット[11]から、「ABSA」は ABSA タグ付きコーパス[12]からデータを取得したことを表す。

4.2 実験設定

本節では、本研究におけるマルチタスク学習モデルの評価実験に関する詳細

な設定について述べる。

4.2.1 学習条件

本実験では一般的なトランスフォーマーベースのモデルのファインチューニングに準拠した学習条件を採用した。ベースの BERT のモデルとして事前学習済み日本語モデルである cl-tohoku/bert-base-japanese-whole-word-masking を使用した。最大系列長は 512 トークンに設定した。学習には AdamW オプティマイザを用い、学習率は 2×10^{-5}, バッチサイズは 16、学習エポック数は最大 10 とし、各エポックごとに性能を検証した。過学習を防ぐために Early Stopping を導入し、性能が改善しない状態が 5 エポック続いた時点で学習を停止した。最適モデルは検証データにおける正解率に基づき自動的に保存されるよう設定した。本研究で実施した全ての実験は Google Colaboratory 上の NVIDIA T4 GPU 環境にて実行した。

4.2.2 比較手法

本研究では、提案するマルチタスク学習モデルの有効性を多角的に検証するため、以下の 4 種類のモデルを比較対象として設定した。すべてにおいて、事前学習済みの BERT を共通のバックボーンとし、構造的な違いが性能に及ぼす影響を評価した。

1. シングルタスク学習モデル（STL）
属性に対する感情分析タスクのみを行うモデル。

2. 基本的なマルチタスク学習モデル（MTL-Basic）
文書全体に対する感情分析と属性に対する感情分析を共通の BERT エンコーダから並列に学習するモデル（3.2 節参照）。

3. 中間層共有マルチタスク学習モデル（MTL-Shared）
MTL-Basic に共有の中間層を追加したモデル（3.3 節参照）。

4. 修正済み訓練データによるマルチタスク学習モデル（MTL-Refined）
MTL-Basic の構成に加え、曖昧な文を除外するフィルタリングを適用した訓練データを用いて学習するモデル（3.4 節参照）。

さらに、属性に対する感情分析タスクの訓練データの量を減少させた実験を実施した。4.1 にて述べた通り、属性ごとに訓練・検証データの量をそれぞれ 50%、25%、10%へ削減したデータセットを用いた。これにより訓練データの量が感情分析モデルの性能に及ぼす影響や、訓練データ量に対するマルチタスク学習の頑健性を評価する。

性能の比較には正解率と F 値を使用した。

4.3 楽天トラベルデータセットでの実験結果

4.3.1 実験結果と考察

初めに、楽天トラベルデータセットからランダムで抽出したレビュー文 1 万件を用いて、文書全体の感情と属性に対する感情を同時に予測するマルチタスク学習モデル(MTL-Basic)を訓練した。すなわち、D-RD-Random と A-RD-Random を用いて感情分析モデルを学習した。比較として、A-RD-Random のみからシングルタスク学習モデルを学習した。実験結果を表 4.2 に示す。どの属性についても、マルチタスク学習モデルはシングルタスク学習モデルと比べて正解率や F 値に大きな差は見られなかった。

実験 1 ではデータセット D-RD-Random における極性ラベルに大きな偏りが存在する。表 4.3 に示すように、楽天データセットでは肯定のラベルが付与されたレビューの数が中立や否定と比べてかなり多い。このような状況下では、どちらのモデルも「肯定」のラベルを予測することが多かった。このことが、マルチタスク学習によって属性に対する感情分析の正解率や F 値が改善しなかった原因と考えられる。

表 4.2 実験 1 の結果

	正解率		F 値	
	STL	MTL-Basic	STL	MTL-Basic
1. 立地	0.836	0.833	0.339	0.33
2. 部屋	0.754	0.755	0.521	0.489
3. 食事	0.747	0.743	0.58	0.492
4. 浴場	0.688	0.68	0.5	0.473
5. サービス	0.777	0.771	0.532	0.534
6. アメニティ	0.678	0.68	0.402	0.417

表 4.3　楽天データセットにおける極性ラベルの分布

	肯定	中立	否定
1. 立地	1,814,597	315,623	63,956
2. 部屋	1,651,708	407,214	135,254
3. 食事	1,623,847	412,892	157,437
4. 浴場	1,452,425	412,892	159,502
5. サービス	1,660,329	412,892	103,073
6. アメニティ	1,463,648	583,418	147,110

実験 1 の反省を踏まえ、実験 2 では肯定・中立・否定の各ラベルが同程度出現するようにデータセットを整備した。すなわち、D-RD-Balanced と A-RD-Balanced を用いて感情分析モデルをマルチタスク学習した。また、比較の対象となるシングルタスク学習モデルは A-RD-Balanced を用いて学習した。これによりモデルの予測が頻度の高いラベルに偏るという問題を回避しモデルの性能を明確に比較できると考えた。

実験 2 の結果を表 4.4 に示す。この表では感情分析の正解率及び F 値のマクロ平均を示すが、F 値を計算する過程で算出した精度と再現率を付録 A における表 A.1 に示す。「立地」「食事」「サービス」の属性については、MTL-Basic が STL を正解率、F 値ともに上回った。一方、属性が「部屋」のとき、MTL-Basic は SMT を正解率・F 値ともに下回った。一部の属性についてはマルチタスク学習によって性能の向上が見られたものの、全体的にはマルチタスク学習とシングルタスク学習とで大きな差は見られなかった。文書全体の感情分析タスクとのマルチタスク学習によって属性に対する感情分析タスクの性能が向上する可能性が示唆されたが、どの属性に対してもその有効性を発揮するためにはさらなる工夫が必要である。さらに、文書全体に対する感情分析と属性に対する感情分析とでの精度を比較したところ、文書全体に対する感情分析の方が明らかに高い精度を示した。比較結果を表 4.5 に示す。「正解率」「F 値」のいずれの指標においても、属性に対する予測は文書全体レベルの精度に及ばなかった。

表 4.4 実験 2 の結果

	正解率		F 値	
	STL	MTL-Basic	STL	MTL-Basic
1. 立地	0.538	0.547	0.52	0.529
2. 部屋	0.663	0.649	0.66	0.648
3. 食事	0.636	0.67	0.64	0.672
4. 浴場	0.613	0.611	0.611	0.611
5. サービス	0.674	0.678	0.672	0.671
6. アメニティ	0.555	0.566	0.56	0.549

表 4.5 実験 2 における ABSA と文書全体の精度比較

	ABSA		文書全体	
	正解率	F 値	正解率	F 値
1. 立地	0.547	0.529	0.747	0.612
2. 部屋	0.649	0.648	0.733	0.668
3. 食事	0.67	0.672	0.74	0.641
4. 浴場	0.611	0.611	0.75	0.603
5. サービス	0.678	0.671	0.704	0.638
6. アメニティ	0.566	0.549	0.74	0.609

4.3.2 実験 2 に対する誤り分析

提案手法の課題を明らかにするため、テストデータの一部に対する誤り分析を行う。分析対象としたのは、属性が「食事」のとき、マルチタスク学習モデルが正解しシングルタスク学習モデルが誤分類したケース、およびその逆のケース、並びに両モデルが誤ったケースの 3 つである。

マルチタスク学習モデルのみが正解した事例を表 4.6 に示す。これらの事例では、「夕食が残念だった」「ストローネが非常に美味しかった」など、食事に言及した意見が多く含まれていた。一方、シングルタスク学習モデルのみが正解した事例を表 4.7 に示す。これらの事例では、食事と関係のない記述が中心であった。属性に対する極性の根拠が曖昧な文が入力されたとき、マルチタスク学習では予測を誤ることが多いと言える。

両モデルが不正解となった事例を表 4.8 に示す。この事例では、「可もなく不可もなく」「期待しない方がよい」という否定的な意見が見られ、両モデルの予測ラベルも「否定」であるが、正解ラベルは「中立」である。このデータセットにおける極性ラベルはユーザーによって与えられた評点を基に決定しているが、その信頼性には疑問が残る。他のレビューについても評価スコアとレビュー文との整合性に欠けるものが確認された。たとえば、星 5 という高評価にもかかわらず、「髪が落ちていた」など否定的な表現が含まれていたり、逆に星 1 で「温泉がなかなかよい」など肯定的な内容が述べられていたりしているものもあった。このような乖離が生じる要因として、宿泊施設利用者の主観的な評価基準や入力ミスが考えられる。いずれにせよ、このようなレビューは学習においてノイズとして作用する可能性が高く、精度を下げる一因となっていると考えられる。

表 4.6 実験 2 において MTL-Basic のみが正解したケース

正解	STL	MTL-Basic	レビュー文
否定	中立	否定	部屋も廻りの景色も最高でした。一つ残念なのが夕食でした兎に角粗末な感じでした朝はバイキングで普通でしたが
否定	中立	否定	...種類を減らしてでも、あるいはラーメンとか蕎麦だけで もいいので美味しいと思える食事にして欲しい。 せっかくの宿泊が、ただ寝るだけに来たような感じで
肯定	否定	肯定	一応ほぼ全種類あったと思います。長崎を出る前に皿うどんが食べたかったので朝食はサラダとスープのみにしましたが、ミネストローネが本当に美味しかったです

表 4.7 実験 2 において STL のみが正解したケース

正解	STL	MTL-Basic	レビュー文
否定	否定	中立	とにかく部屋がきれいで清潔感があります。
肯定	肯定	否定	朝食がイタリアンだけなのは残念。しかも最初の皿にコー ルドミートがでてきては、....

表 4.8 実験 2 においていずれも不正解であるケース

正解	STL	MTL-Basic	レビュー文
中立	否定	否定	夕食付きのプランでしたが、「可もなく不可もなく」と いった感じで感動した料理は特にありませんでした。 朝食は他のかたも...、期待しない方が良いと思います。

4.4 ABSA タグ付きコーパスでの実験結果

本節では、属性に対する感情分析のデータセットとして ABSA タグ付きコーパスを用いた実験結果について報告する。4.1 節でも述べたが、4.3 節の実験とは異なり属性が 7 種類存在する。本実験では、STL、MTL-Basic、MTL-Shared、MTL-Refined の 4 つのモデルを比較する。また、訓練データ量を「全体」「1/2」「1/4」「1/10」としたときの実験を行う。実験に使用するデータセットを以下にまとめる。

- STL: A-ABSA-ALL, A-ABSA-1/2, A-ABSA-1/4, A-ABSA-1/10 のいずれか
- MTL-Basic: 文書全体に対する感情分析のデータセットとして D-RD-Balanced を、属性に対する感情分析のデータセットとして A-ABSA-ALL, A-ABSA-1/2, A-ABSA-1/4, A-ABSA-1/10 のいずれかを用いる。
- MTL-Shard: MTL-Basic と同じ。
- MTL-Refined:文書全体に対する感情分析のデータセットとして D-RD-Filtering を、属性に対する感情分析のデータセットとして A-ABSA-ALL, A-ABSA-1/2, A-ABSA-1/4, A-ABSA-1/10 のいずれかを用いる。

実験結果を表 4.9、4.10、4.11、4.12 に示す。これらの表における括弧内の数値は各指標の STL との差を示す。これらの表では感情分析の F 値のマクロ平均を示すが、F 値を計算する過程で算出した精度と再現率を付録 A における表 A.2、A.3、A.4、A.5 に示す。

まず全ての訓練データを用いた実験(表 4.9)では、シングルタスク学習モデルが全体的に高い性能を示し、マルチタスク学習による明確な利点は見られなかった。特に「立地」や「アメニティ」といった属性ではシングルタスク学習モデルが最も高い正解率を記録した。一部の属性でマルチタスク学習モデルが上回るものの、その差分は 1〜2 ポイント以内に留まり、有意な改善とは言い難い。平均正解率を比較すると、シングルタスクは 0.814、マルチタスクは 0.815 とほぼ同等であり、モデル間の性能差はごくわずかである。

次に、訓練データを 1/10 に大幅に減少させた実験(表 4.12)では、マルチタス

ク学習モデルの有効性が顕著に現れた。「立地」「部屋」「朝食」「浴室」「サービス」「アメニティ」についてマルチタスク学習モデルがシングルタスク学習モデルを上回り、全 7 属性のうち 6 属性についてマルチタスク学習の有効性が確認された。平均正解率でも、シングルタスクは 0.720 であるのに対し、マルチタスク学習モデル MTL-Basic は 0.750 と 3 ポイント以上の差を示した。これらの結果は低リソース環境でのマルチタスク学習の効果を裏付けている。

一方、訓練データを 1/2 や 1/4 に削減した実験(表 4.10 や 4.11)では、モデル間の性能差は属性ごとのばらつきが大きく、全体的な優劣は明確ではなかった。具体的には、「部屋」や「朝食」といった属性ではマルチタスク学習モデルがやや優勢である一方、「立地」や「夕食」などラベルの分布に偏りのある属性ではシングルタスク学習モデルが勝る場合もあった。これは各属性のテキスト特徴や感情表現の違いに起因すると考えられる。

MTL-Shared は、訓練データが十分にある条件（表 4.9）では STL および MTL-Basic と同等の正解率を示す一方で、訓練データを削減した条件下では全体的に性能が不安定であった。特に表 4.11 および表 4.12 においては、いくつかの属性で STL や MTL-Basic を下回る結果となり、平均正解率も他モデルよりやや劣る傾向が見られた。

MTL-Refined は、MTL-Basic と比較して目立った平均精度の向上は見られなかったが、訓練データ量を 1/10 にした表 4.12 では属性によっては STL を明確に上回るケース（立地、部屋など）もあり、訓練データのフィルタリングにより補助タスク由来のノイズが抑制された効果が一定程度あったと考えられる。

以上の結果より、マルチタスク学習モデル MTL-Basic および MTL-Refined は特に訓練データ量が少ない状況において性能向上に寄与することが示された。訓練データの量がどれほどあればシングルタスク学習でも十分な性能のモデルが得られるかの検証は十分ではないが、本実験の結果は今後の低リソース環境における感情分析モデルの設計の指針に重要な示唆を与えるものと考えられる。

表 4.9 ABSA タグ付きコーパスを用いた感情分析の実験結果（訓練データ量=全て）

no	項目	STL	MTL-Basic	MTL-Shared	MTL-Refined
1	立地	0.842	0.824 (-0.018)	0.812 (-0.03)	0.825 (-0.017)
2	部屋	0.793	0.813 (+0.02)	0.790 (-0.003)	0.802 (+0.009)
3	朝食	0.814	0.820 (+0.006)	0.791 (-0.023)	0.789 (-0.025)
4	夕食	0.820	0.815 (-0.005)	0.824 (+0.004)	0.817 (-0.003)
5	浴室	0.843	0.846 (+0.003)	0.835 (-0.008)	0.828 (-0.015)
6	サービス	0.807	0.823 (+0.016)	0.816 (+0.009)	0.805 (-0.002)
7	アメニティ	0.802	0.786 (-0.016)	0.785 (-0.017)	0.788 (-0.014)
	平均	**0.814**	**0.815 (+0.001)**	**0.810 (-0.004)**	**0.812 (-0.002)**

表 4.10 ABSA タグ付きコーパスを用いた感情分析の実験結果（訓練データ量 =1/2）

no	項目	STL	MTL-Basic	MTL-Shared	MTL-Refined
1	立地	0.811	0.807 (-0.004)	0.813 (+0.002)	0.802 (-0.009)
2	部屋	0.775	0.799 (+0.024)	0.788 (+0.013)	0.786 (+0.011)
3	朝食	0.800	0.821 (+0.021)	0.824 (+0.024)	0.821 (+0.021)
4	夕食	0.800	0.794 (-0.006)	0.797 (-0.003)	0.822 (+0.022)
5	浴室	0.815	0.818 (+0.003)	0.790 (-0.025)	0.817 (+0.002)
6	サービス	0.799	0.802 (+0.003)	0.762 (-0.037)	0.771 (-0.028)
7	アメニティ	0.778	0.773 (-0.005)	0.772 (-0.006)	0.763 (-0.015)
	平均	**0.799**	**0.799 (±0.000)**	**0.790 (-0.009)**	**0.793 (-0.006)**

表 4.11 ABSA タグ付きコーパスを用いた感情分析の実験結果（訓練データ量 =1/4）

no	項目	STL	MTL-Basic	MTL-Shared	MTL-Refined
1	立地	0.785	0.757 (-0.028)	0.747 (-0.038)	0.757 (-0.028)
2	部屋	0.741	0.770 (+0.029)	0.777 (+0.036)	0.756 (+0.015)
3	朝食	0.792	0.807 (+0.015)	0.802 (+0.010)	0.809 (+0.017)
4	夕食	0.765	0.786 (+0.021)	0.786 (+0.021)	0.814 (+0.049)
5	浴室	0.782	0.768 (-0.014)	0.735 (-0.047)	0.773 (-0.009)
6	サービス	0.771	0.770 (-0.001)	0.752 (-0.019)	0.753 (-0.018)
7	アメニティ	0.736	0.748 (+0.012)	0.757 (+0.021)	0.727 (-0.009)
	平均	**0.770**	**0.770 (±0.000)**	**0.758 (-0.012)**	**0.773 (+0.003)**

表 4.12 ABSA タグ付きコーパスを用いた感情分析の実験結果（訓練データ量=1/10）

no	項目	STL	MTL-Basic	MTL-Shared	MTL-Refined
1	立地	0.716	0.764 (+0.048)	0.698 (-0.018)	0.773 (+0.057)
2	部屋	0.727	0.787 (+0.060)	0.763 (+0.036)	0.769 (+0.042)
3	朝食	0.760	0.780 (+0.020)	0.757 (-0.003)	0.734 (-0.026)
4	夕食	0.749	0.741 (-0.008)	0.729 (-0.020)	0.719 (-0.030)
5	浴室	0.745	0.770 (+0.025)	0.745 (0.0)	0.766 (+0.021)
6	サービス	0.722	0.746 (+0.024)	0.743 (+0.021)	0.739 (+0.017)
7	アメニティ	0.676	0.713 (+0.037)	0.672 (-0.004)	0.690 (+0.014)
	平均	**0.720**	**0.750 (+0.030)**	**0.728 (+0.008)**	**0.742 (+0.022)**

さらに、学習時のエポック数を増やしたとき、感情分析モデルの検証データに対する正解率・F 値がどのように変化するかを調べた。図 4.1 と図 4.2 は、それぞれ、全ての訓練データ(A-ABSA-ALL)を用いたときの正解率、F 値の変化を示している。図 4.3 と図 4.4 は、それぞれ、訓練データ量を 1/10 にしたとき(A-ABSA-1/10)の正解率、F 値の変化を示している。

まず、図 4.1 と図 4.2 の結果から、十分な訓練データが存在する条件下では、エポック数が 1 でも正解率や F 値が高く、エポック数を増やしても大きく改善することはなかった。各エポック数において、シングルタスク学習モデルの方がマルチタスク学習モデルよりも正解率や F 値が高いが、その差はそれほど顕著ではなかった。

一方、訓練データ量を 1/10 にした条件(図 4.3・図 4.4)では、シングルタスク学習モデルの評価指標に振れ幅が大きく見られ学習の不安定性が確認された。対照的にマルチタスク学習モデルはエポック数が増えるについて評価指標が安定して向上していた。これは、補助タスクからの情報を活用することで安定したモデル学習が実現できているためと考えられる。さらに、検証用データセッ

トに対する正解率は最終的に両モデルで大きな差はなかったが、テストデータセットにおいてはマルチタスク学習モデルの方が一貫して高い正解率を示した。これらの結果は、マルチタスク学習が複数タスクの情報共有を通じてモデルの汎化性能を高めていること、特にデータが限られた環境において優位に働くことを示唆している。

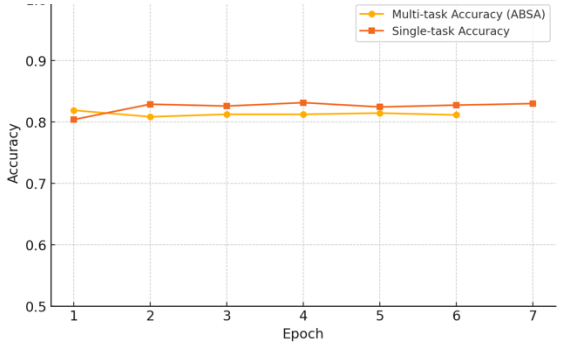

図 4.1 エポック数に対する検証データの正解率の推移（訓練データ量=全て）

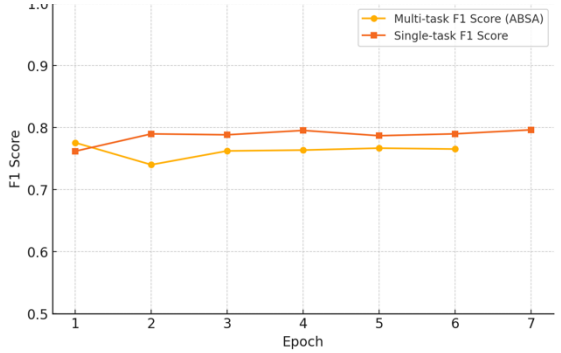

図 4.2 エポック数に対する検証データの F 値の推移（訓練データ量=全て）

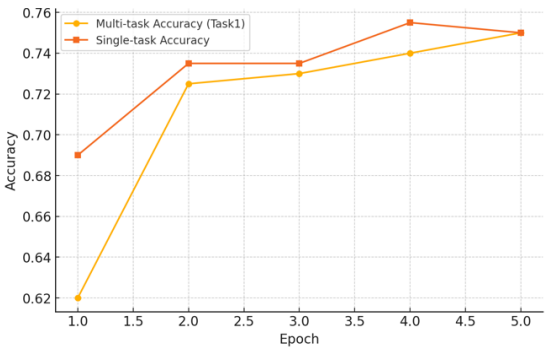

図 4.3 エポック数に対する検証データの正解率の推移（訓練データ量=1/10）

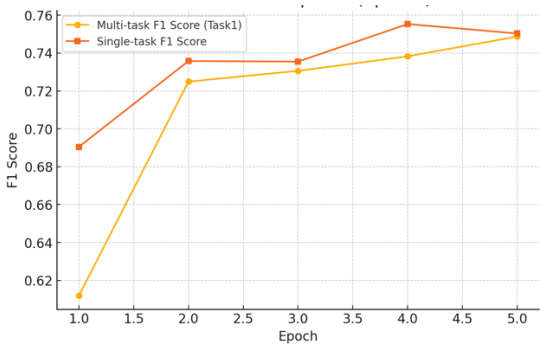

図 4.4 エポック数に対する検証データの F 値の推移（訓練データ量=1/10）

4.5 事例分析

本節では個々の極性判定の事例を分析することで提案手法のモデルの利点や欠点を明らかにする。事例分析に先立ち、訓練データ量を 1/10 にした実験設定において、4 つの分類モデルが正しい極性を予測できた事例数（正解数）を極性ラベル毎に調べた。表 4.13 は、7,000 件のテストデータを用いた実験において、極性ラベルごとに各モデルがどの程度正答していたかを比較したものである。この 7,000 件のテストデータは 7 つの各属性（立地、部屋、夕食、朝食、風呂、サービス、アメニティ）のサンプルから 1,000 件ずつ抽出して作成した。全体の傾向として、「否定」や「中立」といった極性ラベルに対する正解数がモデルによって大きく異なることが確認できる。

表 4.13 各モデルの極性ラベル毎の正解数(訓練データ量=1/10)

ラベル	STL	MTL-Basic	MTL-Shared	MTL-Refined
否定	1365	1438	12	1358
中立	1720	1795	230	1830
肯定	2010	2068	2058	2002

まずシングルタスク学習モデル STL は、否定・中立・肯定のいずれのラベルにおいても一定の正答数を記録しており、訓練データ数を 1/10 に減らしてもモデルが基本的な推論能力を備えていることが確認できる。ただ中立ラベルでは、MTL-Basic や MTL-Refined と比べてやや正答数が少なくなっている。

基本的なマルチタスク学習モデル MTL-Basic は 4 つのモデルの中で正解数の合計が最も多かった。特に否定や中立ラベルの正解数が顕著に多い。これは、

異なるタスクの文脈情報を学習の過程で取り込むことにより、レビュー文で表明されている意見の極性を識別する能力が強化されたためと考えられる。特定の極性ラベルのみを予測しやすいといったバイアスは見られず、肯定、中立、否定のいずれの特徴も学習できていると考えられる。また、データ件数がやや少ない否定ラベルにおいても、補助タスクの知識が補完されており、典型的なマルチタスク学習の利点が確認できた。

中間層を共有するマルチタスク学習モデル MTL-Shared では、出力ラベルの分布に極端な偏りが発生した。具体的には、肯定ラベルの正解件数はシングルタスク学習モデル STL と比べてわずかに多くなった一方、否定ラベルは正解件数 12 件、中立ラベルの正解件数は 230 件と、大幅に減少した。このような挙動は、共有層による表現の過剰な一般化が原因であると考えられる。層の共有によって、本来タスク固有であった特徴表現が失われ、より頻度が高く判別しやすい肯定ラベルに過剰に最適化された可能性が高い。本結果は、マルチタスク学習における層の共有が常に有効とは限らないことを示唆している。特にラベル分布に偏りが存在する場合には、層の共有によって逆に性能が大きく損なわれるリスクがある。今後、層の共有に際しては、タスクの固有の特徴表現が他のタスクの干渉によって失われることを抑制する設計や、タスク毎の特徴表現の選択的共有の導入などが必要になると考えられる。

訓練データのフィルタリングを行ったマルチタスク学習モデル MTL-Refined では、中立ラベルの正解件数が最も多く、否定および肯定ラベルの正解数がやや減少した。中立ラベルは明確な感情表現が含まれないレビュー文や感情表現が弱い文に割り当てられやすいが、肯定的意見と否定的意見の両方を含んで「中立」のラベルが付与された事例がフィルタリングによって削除されたことにより、このような文を正確に「中立」と分類できるようになったと考えられる。一方、肯定クラスや否定クラスの正解数はシングルタスク学習モデル STL と比べて大きな変化はないことから、訓練データのフィルタリングは肯定や否定の事例の分類に大きな影響を与えることなく、中立の事例の分類の再現率を向上することができた。

上記を踏まえて、具体的な事例の分析は、中間層を共有するマルチタスク学習モデルは否定や中立の正解数が大きく減少したことからこれを除外し、シングルタスク学習モデル(STL)・基本的なマルチタスク学習モデル(MTL-Basic)・訓練データのフィルタリングを行ったマルチタスク学習モデル(MTL-Refined)の 3 モデルに対して行う。以下では、これらの 3 つのモデルによる予測が一致しない事例を分析し、マルチタスク学習の利点と欠点を考察する。

(1) MTL-Basic と MTL-Refined が正解し、STL が不正解であるケース

表 4.14 は 2 種類のマルチタスク学習モデルが肯定と予測しシングルタスク学習モデルのみが誤って否定ラベルを予測した事例である。属性に対する評価表現が短文であるものや、肯定的意見が間接的に示されているような文が多い。たとえば、「ベッドの寝心地も良かったです（部屋)」といった短い文や、「ごはんがおいしいから好きと話したところ…（夕食)」のような文は極性が明示的に示されているわけではない。これらのケースでは、マルチタスク学習によって補助タスクの情報を利用することにより、曖昧な表現を理解する能力が向上したと考えられる。

また、STL のみが否定ラベルを肯定と誤って予測したケースでは、「部屋がカビ臭い（部屋)」「予約係の対応にガッカリしました（サービス)」「もう少し...があればと思いました（サービス)」など、婉曲的な否定表現が多く見られた。肯定ラベルに対する誤判定のケースと同様に、STL では曖昧な表現を捉えきれなかったと考えられる。

表 4.14　2 種類のマルチタスク学習モデルのみの予測が正しいレビュー文の例

レビュー文	属性	正解ラベル, MTL-Basic, MTL-Refined	STL
ベッドの寝心地も良かったです。	部屋	肯定	否定
担当してくれた仲居さんに「翠山亭はごはんがおいしいから好き」と話したところ、どうやらお米（米飯）のことだと思われてしまったようでこちらの言い方が悪かったなと後になって気付きました（笑）お米も美味しいですが、食事全般が美味しいから好き、ということを今更ですが付け足しておきます。	夕食	肯定	否定
向かいにはミニストップがあります。	立地	肯定	否定
温泉はよく、料理も美味しく接客も良くして頂きましたが、なにせ部屋がカビ臭い。	部屋	否定	肯定
部屋食でゆっくりしたかったのでなにが出てくるかわからない食事プランでしたが、子どもが食べれないものにも好き嫌いはアレルギーじゃないので追加注文しろとかスタッフの方々がいいのに予約係の対応にガッカリしました。	サービス	否定	肯定
スタッフがもう少し細やかな気配りがあればと思いました。	サービス	否定	肯定

(2) MTL-Basic のみが正解したケース

表 4.15 は STL が誤って肯定ラベルを否定と予測した中で、MTL-Basic のみが肯定と予測し正解している例を示している。これらのレビュー文では、属性に対する評価が明示されているものの、複数の属性に対しての評価が存在していることが多い。たとえば「シャワーも水圧しっかり、...設備は古いながらも

（風呂）」のように、肯定的評価が含まれていても、「古い」という否定的な評価語に影響されて否定と予測したと考えられる。

さらに否定ラベルを肯定と誤って予測したケースでは、「排水口の匂いが減点だが、...他は満足（風呂）」といったように、否定的な意見の後に肯定的な意見が続くことが多かった。このように、属性に対する否定評価が文の前半にありながらも、文末で他の属性に対する肯定的な意見を述べている場合、ユーザーが全体として肯定的な印象を持っているとモデルが認識したため、誤って「肯定」と予測したものと考えられる。

表 4.15 基本的なマルチタスク学習モデルのみが正しく予測したレビュー文の例

レビュー文	属性	正解ラベル, MTL-Basic	STL, MTL-Refined
シャワーも水圧しっかりだし、洗顔フォームあるし、設備は古いながらも不都合全くなしです。	風呂	肯定	否定
ロケーションもハード部分もリゾート感があって、期待度 MAX で入館したが、スタッフの教育が少々残念でした。	立地	肯定	否定
部屋着は子供用の貸し出しもあったので、パジャマ持って行かなくても良かったなぁと思いました。(サイズはちょっと分かりません)	アメニティ	肯定	否定
ただお部屋の排水口の匂いがしたのと空気清浄機の音の煩さや、最上階の温泉のシャワーブースの排水のハケの悪さが減点ぐらいで、あとはかなり満足でした。	風呂	否定	肯定
部屋、ユニットバスは手狭感が有りますが、飲み物のサービス、朝食は良いと思います。	部屋	否定	肯定
繁華街にあるので決して環境的に良いところとは言えないが、大浴場・朝食の素晴しさはそれを十分にカバーしていると思う。	立地	否定	肯定

(3) STL のみが正解したケース

表 4.16 は STL のみが正解し、2 種類のマルチタスク学習モデルが誤判定したケースの例である。

まず肯定のレビュー文をマルチタスク学習モデルが否定と誤判定したケースに着目したが、誤判定の要因を特定することはできなかった。ただレビュー文が極めて短文であるものや、対象の属性に対する明示的な評価が見られずかつ文全体としてやや否定的なトーンを帯びているといった特徴が見られた。

また否定のレビュー文をマルチタスク学習モデルが肯定と誤判定したケースでは、「カメムシが出て不快だった」「寺隣接」といったように、対象の属性と関連していると判断するのが難しいレビュー文が見られた。

表 4.16 シングルタスク学習モデルのみが正しく予測したレビュー文の例

レビュー文	属性	正解ラベル, STL	MTL-Basic, MTL-Refined
朝食にも飛騨牛が出たのには驚きました。	朝食	肯定	否定
食事は美味しかったのですが、お部屋がお値段の割に、部屋の壁の剥がれなど古さがとても残念でした。	朝食	肯定	否定
【風呂】★★★★	風呂	肯定	否定
寺隣接なので、墓所等気になる人は川沿いの部屋がよいかも・・？	立地	否定	肯定
(^^ゞ季節柄カメムシが多く、部屋にも沢山いたのでまずはその退治から始まりましたが、大騒ぎしながらの退治もまた、想い出に残る時間でした。	部屋	否定	肯定

(4) MTL-Refined のみが正解した中立ラベルのケース

表 4.13 において MTL-Refined が他のモデルよりも中立ラベルの正解率が高かったことに着目し、正解ラベルが「中立」で MTL-Refined のみが正しく予測できた事例を抽出した。その一部を表 4.17 に示す。文全体としては肯定的または否定的な印象を与えるものの、それが対象の属性に対する評価を表すとは明確に言えないレビューが多く見られた。

MTL-Refined は、補助タスクとして用いた文書全体の感情分析タスクについて、極性表現を含まない中立的な文を選択的に訓練データとして用いている。これにより属性に対する明示的な評価が存在しないケースにおいては意見を述べていない（＝中立と判定する）という判断をする傾向が強くなった可能性がある。

表 4.17 修正済み訓練データによるマルチタスク学習モデルのみが正解した中立ラベルのレビュー文の例

レビュー文	属性	正解ラベル, MTL-Refined	STL, MTL-Basic
特に本館改装後は客室も快適で日を追うごサービス面も向上されていると感じます。	立地	中立	否定
今回の旅行は、非常に楽しく、よい思い出になり、ありがとうございました。	朝食	中立	否定
基本、市役所や繁華街も近く、むらさき川の辺で公園もすぐでロケーションも良く、朝食も普通に美味しかったし、部屋も広めででしたし、一人のフロントマンを除けばホテルの方も感じが良く、オススメのホテルです。	夕食	中立	否定
ホテル自体の場所はなかなかわかりにくかったですが、到着してみてその絶景に驚きました！	部屋	中立	肯定
日常の通りをひと山越えたら、そこはまさに別世界。	サービス	中立	肯定
少し気にした方が良いです。	アメニティ	中立	肯定

第5章 おわりに

本章では本論文を総括する。5.1 節では本研究の成果をまとめる。5.2 節では本研究の今後の展望について述べる。

5.1 本研究のまとめ

本研究では、属性に対する感情分析における訓練データの不足という問題に対処するため、文書全体に対する感情分析を補助タスクとして活用するマルチタスク学習手法を提案した。ベースモデルには BERT を用い、文書全体に対する感情分析と属性に対する感情分析の両タスクの分類モデルを同時に学習させた。マルチタスク学習のアーキテクチャとして、BERT に対してタスク毎の出力層を追加する基本的なモデルと、BERT の次にタスク間で共通する特徴を学習する共有層を追加したモデルを提案した。さらに、文書全体に対する感情分析のデータセットから、複数の属性に対する感情を含み、単一の極性を割り当てることが必ずしも妥当ではない事例を削除するフィルタリング手法を提案した。

評価実験では、文書全体に対する感情分析と属性に対する感情分析の両方のデータセットを用いた学習を行い、属性に対する感情分析単体タスクでの学習（シングルタスク学習モデル）と比較した。その結果、マルチタスク学習モデルは特に訓練データ量が制限される条件下で安定して優れた性能を示し、文書全体に対する感情分析から得られる文脈的特徴が属性に対する感情分析の予測精度向上に貢献することが示唆された。

一方で本研究の手法には限界も確認された。まず、訓練データが十分に存在する状況においては、シングルタスク学習の方が属性によっては高い正解率や F 値を示すことがあり、マルチタスク学習によってモデルの性能が必ずしも向上するわけではないことがわかった。特に「立地」や「夕食」などの属性では、マルチタスク学習の効果が限定的であった。

また、中間層を共有するマルチタスク学習モデルは、文章全体の感情分析タスクのデータセットの影響を過度に受ける傾向があり、肯定クラスへの過剰なバイアスが確認された。これは、補助タスクの情報が属性に対する感情分析タスクに良い影響ではなく悪い影響を与えたためと考えられる。加えて、文書全体のレビュー文の前処理として肯定的・否定的意見の両方を含むレビュー文書を除外するフィルタリング処理を施したが、これによる属性に対する感情分析への影響は限定的であり、補助タスクの訓練データの品質を高めるためのより洗練された手法の開発が必要であることがわかった。

5.2 今後の課題

今後の課題として、まず損失関数の設計が挙げられる。本研究では、属性に対する感情分析と文書全体に対する感情分析の損失を単純に加算する構成を採用したが、両タスクの難易度や学習進度の差を考慮すると、重みの導入による 2 つのタスクの損失のバランス調整が効果的である可能性がある。

また、パラメータ共有の方式についても再検討が必要である。今回は中間層を共有する構成としたものの、予測が肯定クラスに極度に偏ったため、共有方法を再検討する必要がある。加えて、アテンションの可視化や重み分布の比較といったモデル内部の分析は未着手であったが、マルチタスク学習がどのような文脈処理に有効に働いたのかを解明するうえで、今後優先的に取り組むべき課題である。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019). 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 (NAACL-HLT), pp.4171-4186.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS 2020), 33, pp.1877-1901.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), pp.1-67.

[5] Ruidan He, Wee Sun Lee, Hwee Tou Ng, Daniel Dahlmeier. An Interactive Multi-Task Learning Network for End-to-End. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp.504-515,2019.

[6] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. (2018). GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. Proceedings of the 35th

International Conference on Machine Learning (ICML 2018), pp. 794-803.

[7] Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, Edward Gehringer. ALL-IN-ONE: Multi-Task Learning BERT models for Evaluating Peer Assessments. The 14th International Conference on Educational Data Mining,pp.525-532, 2021.

[8] Yiming Zhang, Min Zhang, Sai Wu, and Junbo Zhao. (2022). Towards Unifying the Label Space for Aspect- and Sentence-based Sentiment Analysis. Findings of the Association for Computational Linguistics: ACL 2022, pp. 20-30

[9] 張 懿陽, 竹下 昌志, ラファウ・ジェプカ, 荒木 健治. 補助文自動生成を用いた BERT による日本語アスペクトベース感情分析におけるアスペクトカテゴリ検出の精度向上. 言語処理学会 第 29 回年次大会 発表論文集, pp.572-576, 2023.

[10] 日本語評価極性辞書,
https://www.cl.ecei.tohoku.ac.jp/Open_Resources-Japanese_Sentiment_Polarity_Dictionary.html

[11] 楽天データセット,
https://www.nii.ac.jp/dsc/idr/rakuten/

[12] 楽天トラベルレビュー: アスペクト・センチメントタグ付きコーパス
https://irdb.nii.ac.jp/02613/0004754908

付録

A. 実験結果の詳細

4.3 節で述べた楽天トラベルデータセットを用いた実験において、肯定・中立・否定のサンプルを同数用意する設定の実験結果を表 4.4 に示した。F 値算出の過程で計算した精度と再現率を表 A.1 に示す。表 4.4 の正解率と F 値も再掲する。

表 A.1　表 4.4 の実験結果の詳細

		STL	MTL-Basic
立地	正解率	0.538	**0.547**
	F 値	0.52	**0.529**
	精度	0.52	**0.543**
	再現率	0.538	**0.547**
部屋	正解率	**0.663**	0.649
	F 値	**0.66**	0.648
	精度	**0.661**	0.647
	再現率	**0.663**	0.649
食事	正解率	0.636	**0.67**
	F 値	0.64	**0.672**
	精度	**0.678**	0.675
	再現率	0.636	**0.67**
浴場	正解率	**0.613**	0.611
	F 値	0.611	0.611
	精度	**0.613**	0.611
	再現率	**0.613**	0.611
サービス	正解率	0.674	**0.678**
	F 値	**0.672**	0.671
	精度	**0.674**	0.67
	再現率	0.674	**0.678**
アメニティ	正解率	0.538	**0.566**
	F 値	**0.56**	0.549
	精度	**0.571**	0.549
	再現率	0.555	**0.566**

4.4 節で述べた ABSA タグ付きコーパスを用いた実験において、訓練データ量を「全体」「1/2」「1/4」「1/10」にしたときの実験結果を表 4.9、4.10、4.11、

4.12 にそれぞれ示した。F 値算出の過程で計算した精度と再現率を表 A.2、A.3、A.4、A.5 にそれぞれ示す。元の表に掲載した正解率と F 値も再掲する。

表 A.2 表 4.9 の実験結果の詳細(訓練データ量=全て)

		STL	MTL-Basic	MTL-Shared	MTL-Refined
立地	正解率	**0.842**	0.824	0.812	0.825
	F 値	**0.823**	0.800	0.773	0.788
	精度	**0.815**	0.799	0.772	0.793
	再現率	**0.833**	0.802	0.773	0.786
部屋	正解率	0.793	**0.813**	0.790	0.802
	F 値	0.795	**0.812**	0.790	0.801
	精度	0.797	**0.814**	0.792	0.804
	再現率	0.800	**0.815**	0.795	0.815
朝食	正解率	0.814	**0.820**	0.791	0.789
	F 値	0.812	**0.818**	0.787	0.785
	精度	0.815	**0.821**	0.795	0.799
	再現率	0.819	**0.824**	0.795	0.796
夕食	正解率	0.820	0.815	**0.824**	0.817
	F 値	0.808	0.809	**0.812**	0.806
	精度	0.812	0.805	0.820	**0.823**
	再現率	0.804	**0.818**	0.806	0.794
浴場	正解率	0.843	**0.846**	0.835	0.828
	F 値	0.840	**0.844**	0.833	0.826
	精度	0.838	**0.841**	0.834	0.827
	再現率	0.843	**0.848**	0.831	0.825
サービス	正解率	0.807	**0.823**	0.816	0.805
	F 値	0.807	**0.821**	0.813	0.803
	精度	0.808	**0.823**	0.816	0.805
	再現率	0.807	**0.823**	0.816	0.805
アメニティ	正解率	**0.802**	0.786	0.785	0.788
	F 値	**0.801**	0.781	0.784	0.787
	精度	**0.801**	0.789	0.790	0.794
	再現率	**0.802**	0.786	0.785	0.788

表 A.3 表 4.10 の実験結果の詳細(訓練データ量=1/2)

		STL	MTL-Basic	MTL-Shared	MTL-Refined
立地	正解率	0.811	0.807	**0.813**	0.802
	F 値	0.777	0.770	**0.791**	0.766
	精度	0.772	0.749	**0.774**	0.747
	再現率	0.787	0.810	**0.814**	0.800
部屋	正解率	0.775	**0.799**	0.788	0.786
	F 値	0.772	**0.798**	0.790	0.788
	精度	0.776	**0.799**	0.794	0.795
	再現率	0.791	**0.809**	0.795	0.791
朝食	正解率	0.800	0.821	**0.824**	0.821
	F 値	0.800	0.820	**0.823**	0.820
	精度	0.798	0.819	**0.823**	0.819
	再現率	0.806	0.824	**0.828**	0.821
夕食	正解率	0.800	0.794	0.797	**0.822**
	F 値	0.793	0.790	0.791	**0.814**
	精度	0.788	0.784	0.783	**0.807**
	再現率	0.808	0.806	0.814	**0.824**
浴場	正解率	0.815	**0.818**	0.790	0.817
	F 値	0.816	0.816	0.786	**0.817**
	精度	0.818	0.811	0.785	**0.819**
	再現率	0.818	0.825	0.801	**0.827**
サービス	正解率	0.799	**0.802**	0.762	0.771
	F 値	**0.799**	0.601	0.762	0.579
	精度	**0.801**	0.602	0.763	0.582
	再現率	**0.799**	0.601	0.762	0.578
アメニティ	正解率	**0.778**	0.773	0.772	0.763
	F 値	**0.776**	0.770	0.769	0.763
	精度	**0.779**	0.772	0.773	0.763
	再現率	**0.778**	0.773	0.772	0.763

表 A.4 表 4.11 の実験結果の詳細(訓練データ量=1/4)

		STL	MTL-Basic	MTL-Shared	MTL-Refined
立地	正解率	**0.785**	0.757	0.747	0.757
	F 値	**0.752**	0.715	0.704	0.712
	精度	**0.732**	0.702	0.697	0.700
	再現率	**0.814**	0.782	0.791	0.773
部屋	正解率	0.741	0.770	**0.777**	0.756
	F 値	0.740	0.773	**0.779**	0.756
	精度	0.745	0.772	**0.778**	0.759
	再現率	0.761	0.775	**0.784**	0.762
朝食	正解率	0.792	0.807	0.802	**0.809**
	F 値	0.790	0.806	0.799	**0.807**
	精度	0.791	0.804	0.803	**0.811**
	再現率	0.799	**0.808**	0.796	0.805
夕食	正解率	0.765	0.786	0.786	**0.814**
	F 値	0.755	0.778	0.779	**0.808**
	精度	0.760	0.778	0.773	**0.801**
	再現率	0.779	0.779	0.801	**0.821**
浴場	正解率	**0.782**	0.768	0.735	0.773
	F 値	**0.778**	0.763	0.734	0.771
	精度	**0.774**	0.766	0.731	0.770
	再現率	0.791	0.783	0.754	**0.792**
サービス	正解率	**0.771**	0.770	0.752	0.753
	F 値	**0.769**	0.767	0.741	0.751
	精度	**0.772**	0.770	0.754	0.751
	再現率	**0.771**	0.770	0.752	0.753
アメニティ	正解率	0.736	0.748	**0.757**	0.727
	F 値	0.729	0.749	**0.753**	0.726
	精度	0.739	0.755	**0.759**	0.727
	再現率	0.736	0.748	**0.757**	0.727

表 A.5 表 4.12 の実験結果の詳細(訓練データ量=1/10)

		STL	MTL-Basic	MTL-Shared	MTL-Refined
立地	正解率	0.716	0.764	0.698	**0.773**
	F 値	0.674	0.721	0.651	**0.731**
	精度	0.668	0.703	0.651	**0.717**
	再現率	0.763	**0.781**	0.733	0.762
部屋	正解率	0.727	**0.787**	0.763	0.769
	F 値	0.729	**0.788**	0.765	0.771
	精度	0.730	**0.786**	0.767	0.776
	再現率	0.747	**0.798**	0.779	0.778
朝食	正解率	0.760	**0.780**	0.757	0.734
	F 値	0.756	**0.777**	0.753	0.732
	精度	0.759	**0.780**	0.753	0.743
	再現率	0.756	**0.786**	0.757	0.747
夕食	正解率	**0.749**	0.741	0.729	0.719
	F 値	**0.741**	0.733	0.720	0.700
	精度	0.735	**0.743**	0.718	0.726
	再現率	0.760	**0.762**	0.744	0.697
浴場	正解率	0.745	**0.770**	0.745	0.766
	F 値	0.741	**0.769**	0.746	0.764
	精度	0.738	**0.767**	0.756	**0.767**
	再現率	0.750	**0.779**	0.758	0.775
サービス	正解率	0.722	**0.746**	0.743	0.739
	F 値	0.723	**0.741**	0.739	0.737
	精度	0.724	**0.751**	0.741	0.746
	再現率	0.722	**0.746**	0.743	0.739
アメニティ	正解率	0.676	**0.713**	0.672	0.690
	F 値	0.672	**0.714**	0.673	0.689
	精度	0.670	**0.717**	0.676	0.689
	再現率	0.676	**0.713**	0.672	0.690