

# 修士学位論文

題 目

アンサンブル学習を利用した  
深層言語モデルによる  
マルチラベル分類手法の改良

主 査 森 直樹 教授

副 査 藤本 典幸 教授

副 査 黄瀬 浩一 教授

令和 3 年（ 2022 年）度 修了

(No. 2210104025 ) 楠本 祐暉

大阪府立大学大学院工学研究科  
電気・情報系専攻 知能情報工学分野

## 目次

<b>1</b>	<b>はじめに</b>	<b>1</b>
<b>2</b>	<b>関連研究</b>	<b>3</b>
2.1	マルチラベル分類 (Multi Label Classification, MLC) . . . . .	3
2.2	Sentiment Analysis (SA) . . . . .	4
2.3	Aspect Based Sentiment Analysis (ABSA) . . . . .	5
2.3.1	三浦らが取り組む ABSA . . . . .	5
<b>3</b>	<b>データセット</b>	<b>7</b>
3.1	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 . .	7
3.2	データのクレンジング . . . . .	8
<b>4</b>	<b>要素技術</b>	<b>13</b>
4.1	アンサンブル学習 (Ensemble Learning) . . . . .	13
4.2	分散表現 . . . . .	13
4.3	Transformer . . . . .	14
4.4	BERT . . . . .	16
<b>5</b>	<b>MLC 手法を改良するためのアンサンブル学習を利用した深層言語モデルの提案</b>	<b>20</b>
5.1	Entity Classifiers と Attribute Classifiers . . . . .	20
5.2	Entity Classifiers の作成 . . . . .	21
5.3	Attribute Classifiers の作成 . . . . .	23
5.4	Multi pre-trained models Transformer (Mpm+T) . . . . .	24
5.4.1	新規性 1: アスペクトの分類に特化した小分類器の導入 . . . . .	26
5.4.2	新規性 2: Transformer Encoder への入力分散表現の限定 . . . . .	27
<b>6</b>	<b>ベースモデルの数値実験</b>	<b>29</b>
6.1	実験条件 . . . . .	29
6.2	実験結果と考察 . . . . .	30
<b>7</b>	<b>Mpm+T の有効性を検証する数値実験</b>	<b>31</b>

7.1	比較手法	31
7.2	実験条件	31
7.3	実験結果と考察	32
7.4	新たな評価指標の定義とテストデータによる考察	33
7.4.1	完全正解 (Completely Accuracy) の定義	33
7.4.2	テストデータによる考察	34
<b>8</b>	<b>提案手法 2</b>	<b>39</b>
8.1	3 値分類 Attribute Classifiers の作成	39
8.2	Multi-pretrained models Transformer2 (Mpm+T2)	41
<b>9</b>	<b>ベースモデルの 3-class Attribute Classifiers の数値実験</b>	<b>43</b>
9.1	実験条件	43
9.2	実験結果と考察	44
<b>10</b>	<b>Mpm+T2 の有効性を検証する数値実験</b>	<b>45</b>
10.1	実験条件	45
10.2	実験結果と考察	46
10.3	テストデータによる考察	46
<b>11</b>	<b>まとめと今後の課題</b>	<b>47</b>
11.1	まとめ	47
11.2	今後の課題	48
	<b>参考文献</b>	<b>49</b>

## 図目次

2.1	SA の分析フロー図 (文献 <sup>[1]</sup> Fig 1 より) . . . . .	4
2.2	アスペクトベースセンチメント分析ネットワークモデル (文献 <sup>[2]</sup> Fig 1 より) . . . . .	6
3.1	各アスペクトどうしの相関関係 . . . . .	12
4.1	スタッキングアルゴリズムを用いた深層モデルの概要図 . . . . .	14
4.2	Transformer の概要図 (文献 <sup>[3]</sup> Fig 1 より) . . . . .	17
4.3	Scaled Dot-Product Attention の概要図 (文献 <sup>[3]</sup> Fig 2 より) . . . . .	18
4.4	Multi-Head Attention の概要図 (文献 <sup>[3]</sup> Fig 2 より) . . . . .	19
4.5	BERT の概要図 (文献 <sup>[4]</sup> Fig 1 より) . . . . .	19
5.1	Entity Classifiers と Attribute Classifiers のモデル概要図 . . . . .	21
5.2	入力文書に対する Attribute Classifiers の出力例 . . . . .	23
5.3	入力文書に対する Attribute Classifiers の出力例 . . . . .	25
5.4	提案モデル Multi pre-trained models Transformer (Mpm+T) の概要図 . . . . .	26
5.5	Mpm+T におけるデータの流れ . . . . .	27
7.1	三浦らのモデル概要図 . . . . .	32
8.1	3-class Attribute Classifiers のモデル概要図 . . . . .	40
8.2	入力文書に対する 3-class Attribute Classifiers の出力例 . . . . .	41
8.3	提案モデル Multi pre-trained models Transformer2 (Mpm+T2) の概要図 . . . . .	42

## 表目次

3.1	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のデータ具体例	7
3.2	例文に対するエンティティ、アトリビュート、そしてアスペクト	7
3.3	ニュートラルデータと1つのカテゴリにポジネガが両方付与されたデータの例	9
3.4	ラベル数ごとのデータ数	9
3.5	各アスペクトに含まれる正例の数	11
5.1	Entity Classifiers の作成に用いたデータの例と、分類器が注目すべきエンティティ	22
5.2	朝食アスペクトの次元の圧縮例	22
5.3	Attribute Classifiers の作成に用いたデータの例（朝食と風呂データのみを抜粋）	24
6.1	Entity Classifiers と Attribute Classifiers の作成に用いたデータ数と、それぞれのアスペクトにおける正例と負例の数	29
6.2	Entity classifiers と Attribute Classifiers の実験条件	30
6.3	Entity Classifiers と Attribute Classifiers の正解率と標準偏差	30
7.1	実験で用いたデータ数と、それぞれのアスペクトの正例と負例の数	33
7.2	Mpm+T と三浦らのモデルの実験条件	34
7.3	各 MLC モデルの評価指標の平均と標準偏差	35
7.4	6000 件のテストデータに含まれるシングルラベルデータとマルチラベルデータの数と、マルチラベルデータに含まれるそれぞれのラベルごとのデータ数	35
7.5	ラベル数が6の場合の文書と、正解ラベルと予測ラベル	36
7.6	完全正解データを占めるマルチラベルデータと シングルラベルデータのデータ数と割合	37
7.7	それぞれのラベル数のマルチラベルデータにおける完全正解数とその割合	37
7.8	それぞれのモデルにおける各アスペクトの F1	38
8.1	3-class Attribute Classifiers の作成に用いたデータの例 （「夕食」と「施設」データを抜粋）	42

---

9.1	30000 件のデータを占めるポジティブ, ネガティブ, ニュートラルラベル データの数 . . . . .	43
9.2	実験条件 . . . . .	44
9.3	3-class Attribute classifiers と Attribute Cassifiers の正解率と標準偏差 . . .	44
10.1	Mpm+T2 の実験条件 . . . . .	45
10.2	Mpm+T2 と Mpm+T の評価指標の平均と標準偏差 . . . . .	46

## 1 はじめに

近年 SNS や EC サイト (electronic commerce site) の発達により日々大量のテキストデータが生成されている。それらを有効に活用するために、さまざまなシーンでテキスト分類タスクの需要が高まっている。テキスト分類タスクとは、事前に定義されたカテゴリまたはラベルをテキストに割り当てる処理である。特にビジネスシーンにおいては、作業の省人化とデータの有効活用を目的として、各テキストを自動的にカテゴリに分類することと、その精度を向上させることは大きな課題となっている。テキスト分類では1つのテキストに1つのラベルのみを割り当てる場合もあるが、データを有効活用するためにはテキストには2つ以上のラベルを割り当てる必要がある。このようにテキストに複数のラベルを割り当てるタスクはマルチラベル分類 (Multi Label Classification, MLC) と呼ばれ、自然言語処理 (Natural Language Processing, NLP) の分野では感情分析 (Sentiment Analysis, SA)<sup>[1]</sup> におけるタスクとして注目を集めている。その中でもアスペクトベースセンチメント分析 (Aspect-Based Sentiment Analysis, ABSA)<sup>[5]</sup> は文書から様々な観点を抽出し、その観点到感情極性を付与する研究である。本研究の目的は NLP におけるテキスト分類問題における、MLC の分類精度を向上させることである。

分類対象となるテキストデータはアスペクトベースで作成された日本語文章のカスタマーレビューであり、単一または複数のラベルが付与されている。ABSA における MLC の分類精度向上は大きな課題であり、多くの先行研究が存在する。それらの中でも三浦らのモデル<sup>[2]</sup> は英語の文章データに対して高い識別精度を示しているが、日本語の文章データに対する識別精度<sup>[6]</sup> には改善の余地がある。

そこで本研究では三浦らのモデルにアンサンブル学習を導入することで日本語レビューデータに対する MLC の改良手法の有効性を検証をする。モデルの構築にはアンサンブル学習におけるスタッキングアルゴリズム (Stacking Algorithm)<sup>[7]</sup> を用いる。近年スタッキングアルゴリズムと深層モデルを組み合わせた研究<sup>[8][9][10]</sup> は多数存在し、その有効性が報告されている。また、MLC を Binary Relevance (BR) 法<sup>[11]</sup> と Pairwise 法<sup>[12]</sup> を用いてをシングルラベル分類 (Single Label Classification, SLC) に落とし込むことで、識別精度が向上するこ

とが吉村ら<sup>[13]</sup>によって報告されている。

上述の手法を組み合わせることに加えて, 三浦らのアスペクトセンチメント分析ネットワークの BERT に Transformer Encoder を接続するという構造に倣ったモデルを構築する。つまりモデルの詳細な構築としては, 日本語レビューデータに対する MLC を SLC に落とし込むために BR 法と Pairwise 法に基づいて作成した BERT ベースのモデルをアンサンブルして, それらの出力を特徴量として Transformer Encoder に入力することで最終的なマルチラベルを出力する, というものになっている。これを提案手法としてその有効性を検証する。

以下に本稿の構成を示す。2 章で関連研究の紹介, 3 章で使用するデータセットについて説明して, 4 章では要素技術を紹介する。5 章, 6 章, 7 章では提案手法を説明して, 数値実験の結果やその考察を述べる。8 章, 9 章, 10 章では 7 章で確認した提案手法の有効性を踏まえて, 識別精度を更に改善するために提案手法の一部を変更したモデルを説明して, 数値実験の結果やその考察を述べる。11 章でまとめと今後の課題について述べる。



## 2 関連研究

本章では、本研究の提案手法に関する研究について述べる。

### 2.1 マルチラベル分類 (Multi Label Classification, MLC)

機械学習における重要な問題としてマルチラベル分類 (Multi Label Classification, MLC) がある。MLC とは、1 つの対象に対して複数の事前に定義されたカテゴリまたはラベルを割り当てる分類問題である。この他の分類問題として、対象物を 2 値の集合に分類する 2 値分類や、2 値以上の集合に分類するマルチクラス分類があり、これらは MLC と区別してシングルラベル分類 (Single Label Classification, SLC) と呼ばれる。MLC の具体的な応用先として、テキスト分類、感情分類、画像分類などがある。MLC を解く手法については大きく 2 つのアプローチに分けられることが吉村ら<sup>[13]</sup>によって報告されている。1 つ目は分類器を拡張することで MLC をそのまま扱う手法であり、決定木のアルゴリズムを MLC に適用できるように拡張した Quinlan<sup>[14]</sup>の手法などがある。2 つ目は MLC を SLC に変形させてシングルラベル分類器を用いてラベル推定をする手法であり、その中でも Binary Relevance (BR) 法<sup>[11]</sup>と Label Powersets (LP) 法<sup>[15]</sup>が主流である。BR 法とは各ラベルごとに 1 対残り (One-vs-Rest) で学習をして 2 値分類器を作成することで、MLC を SLC に落とし込んで解く手法である。BR 法の関連手法として全クラスから任意の 2 クラスを選択して 1 対 1 (One-vs-One) の学習をして 2 値分類器を作成する Pairwise 法<sup>[12]</sup>がある。LP 法は候補ラベルとして考えられる全ての組み合わせ集合の中から 1 つの組み合わせを選び出すように分類器を学習して、ラベルを予測する手法である。吉田ら<sup>[13]</sup>は、BR 法の利点は計算量の少なさである一方で、ラベル間の相関関係を利用した予測が出来ない欠点があると述べている。対して、LP 法はラベル間の相関関係を利用した予測が出来る利点があるが、計算量がラベル数に対して指数関数的に増大する欠点がある。つまり、BR 法と LP 法の間にはトレードオフ関係があるとされている。3 章に示す図 3.1 より、ほとんどのアスペクトどうしの相関は強くないことが確認できる。したがって本研究では相関関係を考慮しない BR 法と、その関連手法であるペアワイズ法を採択した。

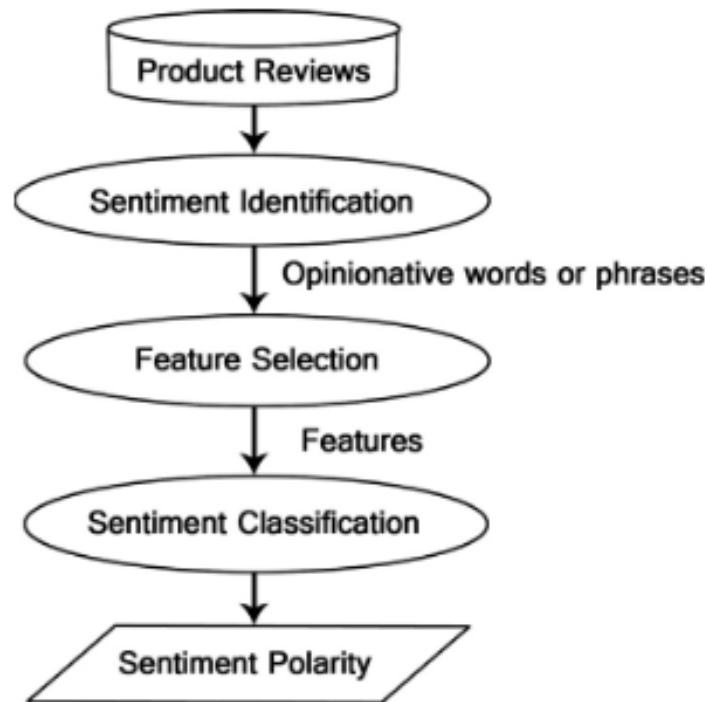


図 2.1: SA の分析フロー図 ( 文献<sup>[1]</sup> Fig 1 より )

## 2.2 Sentiment Analysis (SA)

Sentiment Analysis (SA)<sup>[1]</sup> は、あるエンティティに対する人々の意見、態度、感情を計算機的に研究する分野である。Liu らによるとエンティティ<sup>[16]</sup>とは、文章中における製品、サービス、組織、個人、問題、イベント、トピックなどを指しており、SA における文書分析には重要な指標になるとされている。図 2.1 に Medhat ら<sup>[1]</sup>による、SA の一般的な分析フロー図を示す。SA の分析フローは、生成されたレビュー文章の感情識別をして、エンティティとなる言葉やフレーズに関連した特徴量を選択する。そして、それらの特徴量から文章がどのような感情を表しているかを特定して、その極性を分類することが目標となる。論文では分析対象として商品レビューが採択されているが、Medhat らによると SA は製品レビューだけでなく、株式市場<sup>[17][18]</sup>、ニュース記事<sup>[19]</sup>、政治討論<sup>[20]</sup>にも適用することが報告されている。

## 2.3 Aspect Based Sentiment Analysis (ABSA)

Aspect-Based Sentiment Analysis (ABSA)<sup>[5]</sup> は近年の SA においてアスペクトに基づいて分析する研究分野である. Liu<sup>[5]</sup> は従来の SA におけるテキスト分類では、文書のエンティティに関する感情の詳細が得られないという問題があったため、アスペクトを利用する必要があると提唱した. アスペクトとは文章におけるアトリビュート（肯定的または否定的）とエンティティ（文章内容）で構成される. アスペクト情報を利用するということは、言語構造（文書、段落、文、節、句）を見るのではなく、フレーズとそれに関する感情を利用するということである.

アスペクトに基づいて文書を分析する例を説明する. 「サービスはあまり良くないが、私はこのレストランが大好きだ」という文章がある. この文章はレストランというエンティティについては肯定的だが、サービスというエンティティについては否定的である. このように全体として肯定的と感じる文章でも、アスペクトベースの分析をすることでより詳細な解釈が可能となる.

ABSA における重要な研究ワークショップとして、Semantic Evaluation (SemEval)<sup>1</sup> がある. SemEval は NLP の国際的な評価会である. 意味分析 (Semantic Analysis) や SA の様々な問題を解くための高品質なアノテーションデータセットの作成を支援し、研究ワークショップを毎年開催している. ABSA はテキスト分類におけるアプローチの必要性が注目を集め、2014 年の SemEval2014<sup>[21]</sup> で初めて導入されたタスクであり、その後も継続的にさまざまなアスペクトベースのデータセットによる識別性能の評価がなされている.

### 2.3.1 三浦らに取り組む ABSA

本研究では三浦らのモデルの構造に倣って MLC モデルを構築した. そこで、5 章で詳述する本研究の新規性と比較するために三浦らのモデル概要図を示して研究の概要とデータの流れを説明する. 図 2.2 に先行研究である三浦らが提案したアスペクトベースセンチメント分析ネットワークモデル<sup>[2]</sup>を示す. 図 2.2 のアスペクトカテゴリ分類ネットでは、文書を入力として文埋め込み層

---

<sup>1</sup><https://semEval.github.io/>

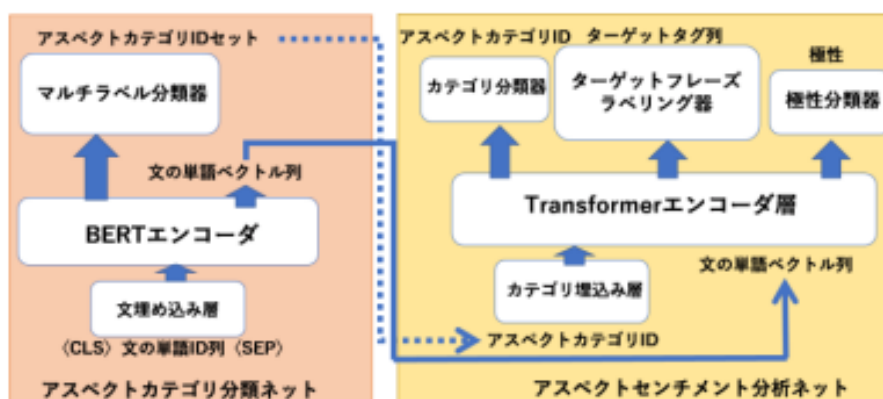


図 2.2: アスペクトベースセンチメント分析ネットワークモデル ( 文献<sup>[2]</sup> Fig 1 より )

で文書トークンを生成して, BERT エンコーダでそれぞれのトークンの分散表現を出力する. そして, それらの分散表現の先頭トークンである CLS トークンをマルチラベル分類器に入力してラベルを推定する. アスペクトセンチメント分析ネットでは, アスペクトカテゴリ分類ネットで得た分散表現とマルチラベル分類結果から推定ラベルと関連度が高いフレーズ ( ターゲットフレーズ ) や極性を分類する.

このモデルは「SemEval-2016 Task 5 : Aspect Based Sentiment Analysis」<sup>[22]</sup> で公開されたレストランのレビューから成るレストランドメインデータセット<sup>2</sup>を用いて, MLC やターゲットフレーズと呼ばれるアスペクト情報の抽出をする. 注目すべきは, マルチラベルデータに対する識別性能の高さであり, 従来手法<sup>[23]</sup>と比較して F1 値は 8.5% 向上していることが報告されている.

また, 三浦らは<sup>[6]</sup>は日本語データセットを用いた SA における MLC に取り組んでいる. この研究では 2018 年に TIS インテックグループの TIS 株式会社によって機械学習で SA をするためのものとして公開された「chABSA-dataset」<sup>3</sup>を用いている. MLC モデルの識別精度は約 70% であり, 区別が難しいアスペクトが含まれているため精度が高くなかったと報告されている.

<sup>2</sup><https://github.com/NUSTM/ACOS/tree/main/data/Restaurant-ACOS>

<sup>3</sup><https://github.com/chakki-works/chABSA-dataset>

### 3 データセット

#### 3.1 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

表 3.1: 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のデータ具体例

		マルチラベルデータ	シングルラベルデータ
文章		部屋も綺麗で, 対応もよく, 朝食もおいしいのでとても満足しています.	一部の方が指摘した通り, 廊下がタバコ臭いのが気になりました.
朝食	ポジティブ	1	0
	ネガティブ	0	0
夕食	ポジティブ	0	0
	ネガティブ	0	0
風呂	ポジティブ	0	0
	ネガティブ	0	0
サービス	ポジティブ	1	0
	ネガティブ	0	1
立地	ポジティブ	0	0
	ネガティブ	0	0
施設	ポジティブ	0	0
	ネガティブ	0	0
部屋	ポジティブ	1	0
	ネガティブ	0	0

表 3.2: 例文に対するエンティティ, アトリビュート, そしてアスペクト

	エンティティ	アトリビュート	アスペクト
朝食は美味しかったし, 部屋も綺麗でした.	朝食	良かった	朝食_ポジティブ
	部屋	綺麗	部屋_ポジティブ

楽天グループ株式会社が事業等を通じて取得したデータから構築して, 国立情報学研究所を通じて研究者に提供しているデータセット「楽天データセット (コレクション)」<sup>[24]</sup>に含まれている「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」<sup>[25]</sup>がある. 表 3.1 にデータの具体例を

示す. 76623 件の楽天トラベルの日本語レビュー文書とそれぞれの文書について, 「朝食, 夕食, 風呂, サービス, 施設, 立地, 部屋」の 7 項目のカテゴリが付与されており, それぞれに感情極性としてポジティブまたはネガティブのタグが付与されている. つまり, 全てのデータに付与されるクラス数は表 3.1 のように 14 個であり, 該当するクラスには正例として 1 のラベルが, 該当しないクラスには負例として 0 のラベルが付与されている. このデータセットの文書にラベルが付与される工程を ABSA の観点から説明する. 表 3.2 に, 例文に対するエンティティとアトリビュート, それらによって特定されるアスペクトを示す. 表 3.2 に示す通り, 予め割り当てられた 7 個のアスペクトに関するエンティティと, そのエンティティの感情極性を表すアトリビュートを抽出することで, 「朝食ポジティブ」と「部屋ポジティブ」のアスペクトに属していると判断する.

14 クラスのうち, ラベル 1 が 1 つしか付与されていないデータをシングルラベルデータとする. また, 14 クラスのうち, ラベル 1 が 2 つ以上付与されているデータをマルチラベルデータとする.

### 3.2 データのクレンジング

全 76623 件のデータから, ノイズを含むデータや日本語以外の言語で書かれたデータと, 文書のトークン数が 101 以上のデータを除外した. この処理によってデータ数は 73461 件となった. この処理後のデータセットから更に 2 種類のデータを除外した. それら 2 種類のデータは 14 のいずれのカテゴリにも属さないニュートラルデータ 23428 件と, 1 つのカテゴリにポジティブとネガティブの両方が付与されたポジネガ両ラベルデータ 1858 件である. 表 3.2 にそれらのデータの具体例を示す. ニュートラルデータを除外した理由は, ABSA は文書のエンティティとアトリビュートによって構成されるアスペクトを利用する SA であるためである. 表 3.2 に示すニュートラルデータの文章のエンティティは「風呂, 客室 (施設, 部屋)」であるが, それらにアトリビュート (肯定的: ポジティブ, または否定的: ネガティブ) のような感情極性は含まれていない. 次に表 3.3 に示すポジネガ両ラベルデータを除外した理由を示す. 最大の理由はデータ件数の少なさである. また, 後に紹介するアトリ

表 3.3: ニュートラルデータと 1 つのカテゴリにポジネガが両方付与されたデータの例

		ニュートラルラベルデータ	ポジネガ両ラベルデータ
文章		展望風呂付客室に 宿泊しました.	朝食の品数がそこまで多くない ですが味と素材が大満足です.
朝食	ポジティブ	0	1
	ネガティブ	0	1
夕食	ポジティブ	0	0
	ネガティブ	0	0
風呂	ポジティブ	0	0
	ネガティブ	0	0
サービス	ポジティブ	0	0
	ネガティブ	0	0
立地	ポジティブ	0	0
	ネガティブ	0	0
施設	ポジティブ	0	0
	ネガティブ	0	0
部屋	ポジティブ	0	0
	ネガティブ	0	0

表 3.4: ラベル数ごとのデータ数

ラベル数	1	2	3	4	5	6	7
データ数	31731	11139	2523	592	167	42	2159

ビュートに関する 2 値分類器の作成時において、ポジネガが両方付与されているデータは扱いが難しかったためである。

最終的にモデルの学習や評価に使用するデータ数は 48353 件となり、最小ラベル数は 1 個で、最大ラベル数は 7 個となった。データに付与されているラベル数の平均は 1.66 個である。このデータのシングルデータ数は 31731 件であり、マルチラベルデータ数は 16622 件である。

表 3.4 にラベル数ごとのデータ数を示す。ラベル数が多い程データ数が少いことが確認できるがラベル数が 7 のデータはその傾向に該当しない。これはラベル数が 7 の場合、2159 件のデータのうち 2112 件が「朝食, 夕食, 風呂, サービス, 施設, 立地, 部屋」の全てについてポジティブであるか、全てについてネガティブとなっている。文書の例を一部紹介する。「総じてレベルが高く、落ち着くホテルです。」「ハイクラスと呼ばれる宿でこんなことは初めてでしたし、非常に悲しかったです。」のように、エンティティや文書全体に対してポジティブあるいはネガティブのアトリビュートがかかっているデータが多い。表 3.5 に全 48353 件のデータのうち、各アスペクトに含まれる正例の数を示す。いずれのアスペクトにおいても正例の数は全データ数の 3 割にも満たないことと、ポジティブラベルデータの方がネガティブラベルデータより多いことがわかる。正例の数が最も多いのは「サービスポジティブ」で 12772 件であり、最も少ないのは「立地ネガティブ」の 857 件である。

図 3.1 に各アスペクトどうしの相関関係を示す。ほとんどのアスペクトは互いの相関が強くないことが確認できる。一方で「朝食ポジティブと夕食ポジティブ」の相関は 0.61, 「朝食ネガティブと夕食ネガティブ」の相関は 0.47 と、他のアスペクトに比べて強いことが確認できる。しかし全体的な相関は強くないため、本研究ではアスペクト間の相関関係をモデルの学習に利用していない。



表 3.5: 各アスペクトに含まれる正例の数

		正例の数
朝食	ポジティブ	11268
	ネガティブ	2438
夕食	ポジティブ	8423
	ネガティブ	1744
風呂	ポジティブ	6922
	ネガティブ	2092
サービス	ポジティブ	12772
	ネガティブ	5208
立地	ポジティブ	5417
	ネガティブ	857
施設	ポジティブ	8024
	ネガティブ	5134
部屋	ポジティブ	7505
	ネガティブ	2342

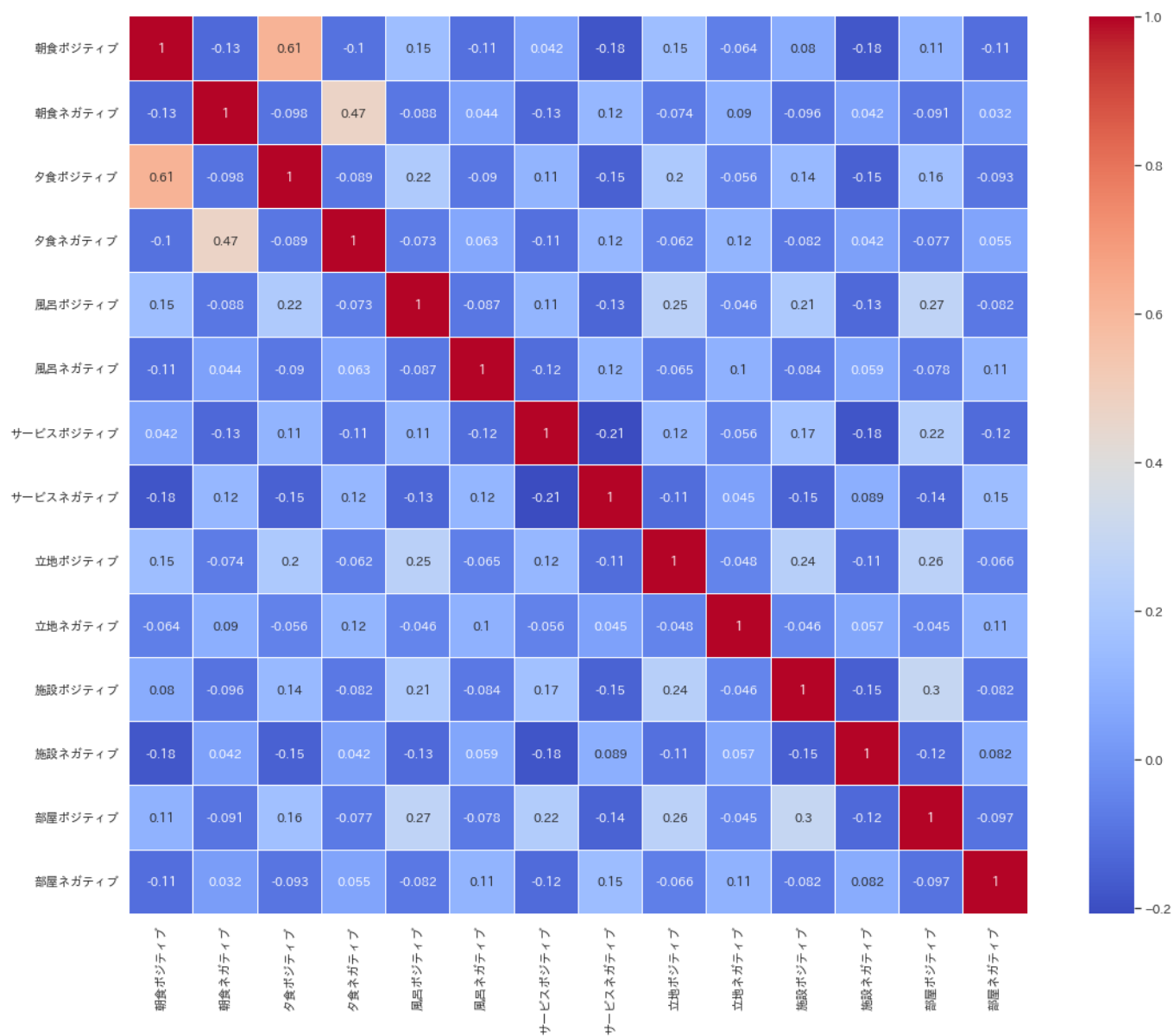


図 3.1: 各アスペクトどうしの相関関係

## 4 要素技術

### 4.1 アンサンブル学習 (Ensemble Learning)

アンサンブル学習 (Ensemble Learning)<sup>[26]</sup> は教師ありの機械学習の分野において、ベースモデルと呼ばれる複数の分類器を組み合わせて推論の精度を高める手法である。多くの代表的なアルゴリズムとしてブースティングアルゴリズム (Boosting Algorithm)<sup>[27]</sup> やバギングアルゴリズム (Bagging Algorithm)<sup>[28]</sup> が存在するが、その中でも本研究で用いているスタッキングアルゴリズム (Stacking Algorithm)<sup>[7]</sup> について述べる。図 4.1 にスタッキングアルゴリズムを用いた深層モデルの概要図を示す。学習のプロセスは大きく 2 つの段階がある。まず初めに、複数のベースモデルを構築して、データを入力としてそれぞれのモデルで推論結果を出力する。次に、出力される予測結果を受け取り、それらをメタ学習器<sup>[29]</sup> と呼ばれるモデルへの入力特徴量として、最終的な予測値を出力する。

つまりスタッキングアルゴリズムとは、複数のベースモデルに、それらの推論結果を学習する 1 つのメタ学習器を結合することでタスクの精度を向上させる手法である。

### 4.2 分散表現

NLP において、言語モデルに深層学習が用いられる以前は単語を one-hot ベクトルで表現する手法として局所表現が主流であった。しかし、単語の語彙数とベクトルの次元数が等しいため、語彙数の増加に応じてベクトルの次元数も増大して、スパースになるという問題があった。その問題点を解決するために考案された手法が分散表現である。

分散表現は各概念をベクトルの単一次元ではなく複数次元の実数で表す。単語の分散表現は“類似した文脈で使用される単語は類似した意味をもつ”という分布仮説を基盤としている。単語を実数値密ベクトルで表現することにより、単語間の意味的な関係性をベクトル空間上での類似度として定義できるという利点がある。

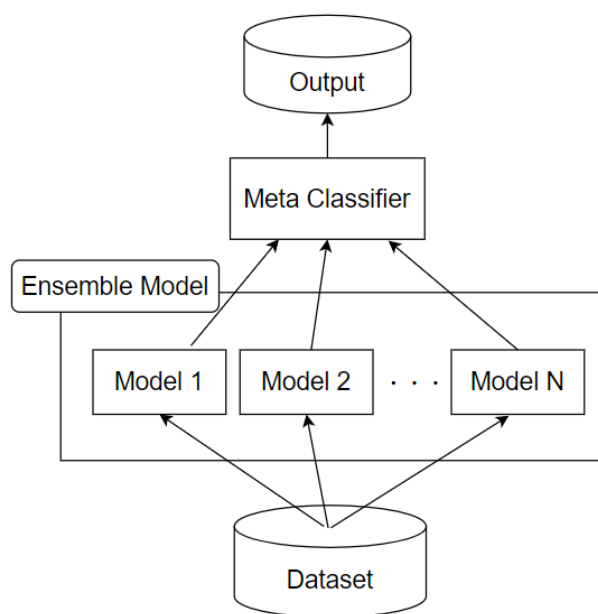


図 4.1: スタッキングアルゴリズムを用いた深層モデルの概要図

### 4.3 Transformer

Transformer<sup>[3]</sup> は Recurrent Neural Network (RNN) や Convolutional Neural Network (CNN) を用いずに, Attention のみを利用した Encoder-Decoder モデルである. 図 4.2 に Transformer の概要図を示す. Transformer は RNN の, 並列処理が不可能なため学習に時間がかかるという問題を, Attention のみを使用して並列計算を可能とすることで解決している. Transformer は, 入力された原言語の単語列  $(x_1, \dots, x_n)$  をエンコードする Transformer Encoder と, 目的言語の単語列  $(y_1, \dots, y_n)$  を生成する Transformer Decoder で構成される. Encoder と Decoder は, 図 4.2 にように, それぞれ Encoder layer と Decoder layer を  $N$  層ずつ積み重ねたものである.

Transformer は RNN のように回帰構造を持たないため, 入力単語の位置情報は Positional Encoding で考慮する. Positional Encoding の行列を  $PE_{pos}$  とすると,

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d}) \quad (4.1)$$

$$\text{PE}_{(\text{pos}, 2i+1)} = \cos(\text{pos}/10000^{2i/d}) \quad (4.2)$$

となる. なお,  $\text{pos}$  は単語の位置,  $d$  は単語埋め込みの次元数,  $i$  は埋め込み成分の次元である. 式 4.1, 4.2 によって計算された  $\text{PE}_{\text{pos}}$  を単語の埋め込み行列に加算したものが Encoder および Decoder の入力となる.

Encoder の各層は 2 つのサブレイヤを持っている. 1 つ目は Multi-Head Self-Attention 機構であり, 2 つ目は単純な Position-Wise Feed-Forward Network である. Position-Wise Feed-Forward Network を  $\text{FFN}(x)$  とすると,

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (4.3)$$

となる. Decoder では, Encoder における 2 つのサブレイヤに加えて, 3 つ目のサブレイヤとして Multi-Head Attention を採用している.

Attention は Query と出力ペアである Key と Value のセットからなる. ここで, Query と Key と Value は全てベクトルである. 出力は Value の加重合計で計算され, このときの重みは Query と対応する Key から計算される値で各 Value と関係のある値である.

図 4.3, 4.4 に Scaled Dot-Product Attention と Multi-Head Attention を示す. それぞれの図における  $Q, K, V$  は Query, Key, Value である. Scaled Dot-Product Attention は  $d_k$  次元の Query と Key と,  $d_v$  次元の Value を入力とする. Query と全ての Key のドット積を計算し,  $\sqrt{d_k}$  で各値を割り, ソフトマックス関数を適用することで Value に対する重みを得る. Scaled Dot-Product Attention を  $\text{Attention}(Q, K, V)$  とすると,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (4.4)$$

となる. なお,  $Q, K, V$  は Query, Key, Value である.

Multi-Head Attention では異なる表現空間の情報を組み合わせて使用する. これは  $d_{\text{model}}$  次元の Key と Value と Query を持つ 1 つの Attention 関数を使用するよりも, 複数の線型写像を組み合わせた方が効果的であるためである. Multi-Head Attention を  $\text{MultiHead}(Q, K, V)$  とすると,

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (4.5)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (4.6)$$

となる. なお,  $QW_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}} \times d_k}$  である.

## 4.4 BERT

BERT (Bidirectional Encoder Representations from Transformers) <sup>[4]</sup> は Transformer によって構成される双方向のエンコーダである. 2018 年 10 月に Google が発表した言語モデルであり, 文章分類, 質問応答, 固有表現抽出等の多様なタスクで公開当時の最高性能を達成するといった大きな成果が報告されている. 図 4.5 に BERT の概要図を示す. BERT は ELMo (Embeddings from Language Models) <sup>[30]</sup> をもとにしたモデルであり, 双方向 LSTM を Transformer に置き換えている. 教師なし学習による事前学習をする際, Transformer によって, 全ての層で左右両方の文脈を加味した学習が可能となる.

BERT には事前学習と fine-tuning の 2 つの学習ステップが存在する. 事前学習では, 周囲の単語からある単語を予測する MLM (Masked Language Model) と, 入力された 2 文が連続した文章であるかを予測する NSP (Next Sentence Prediction) によってモデルを学習する. このような事前学習によって得られたモデルを fine-tuning することで, 様々なタスクに対応することができる.

さらに, 以前はモデル毎に語彙を 1 から学習させるため, 非常に多くの時間とコストがかかっていたが, BERT ではオープンソースで公開されている文脈を既に学習させた Pre-Trained BERT モデル を使用することで短時間で学習ができる.

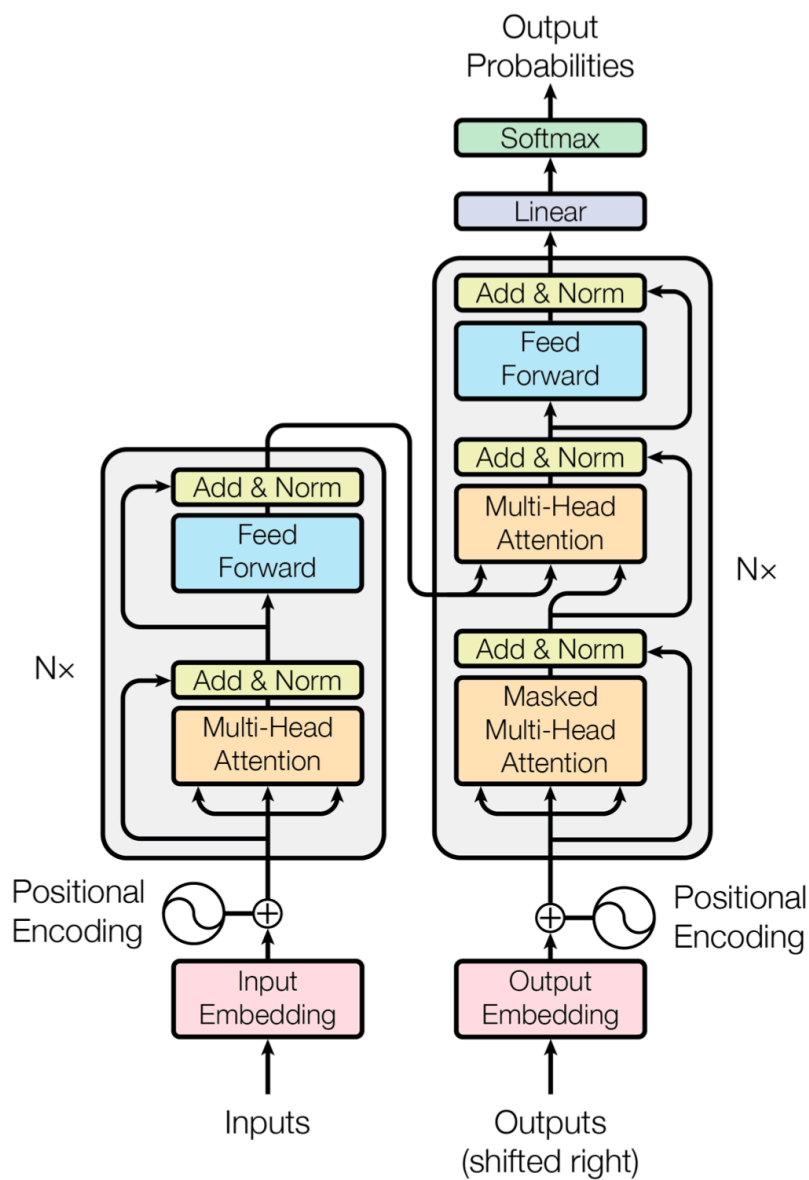


図 4.2: Transformer の概要図 (文献<sup>[3]</sup> Fig 1 より)

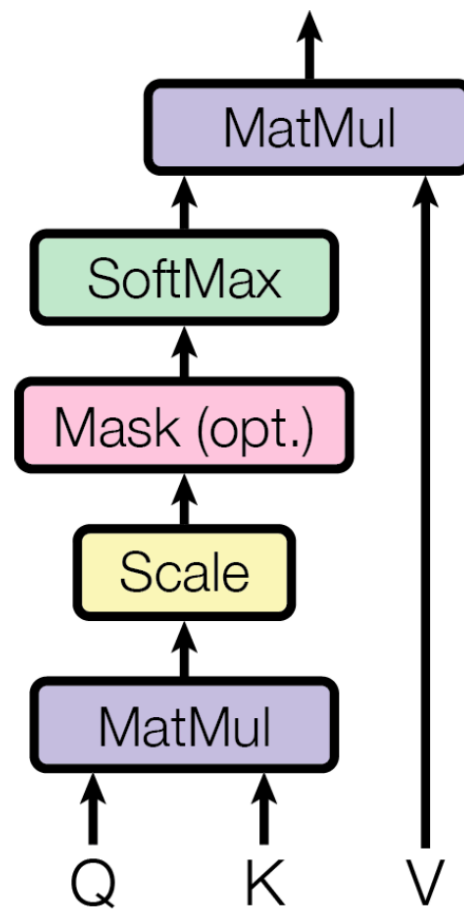


図 4.3: Scaled Dot-Product Attention の概要図 (文献<sup>[3]</sup> Fig 2 より)



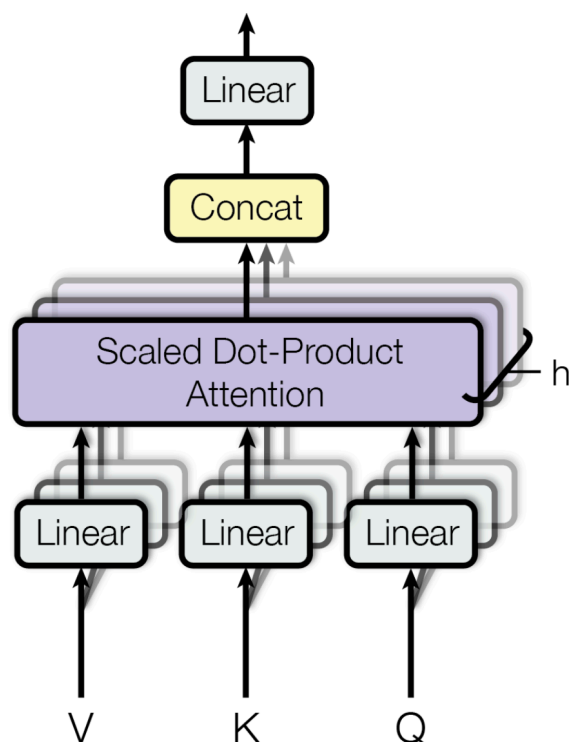


図 4.4: Multi-Head Attention の概要図 (文献<sup>[3]</sup> Fig 2 より)

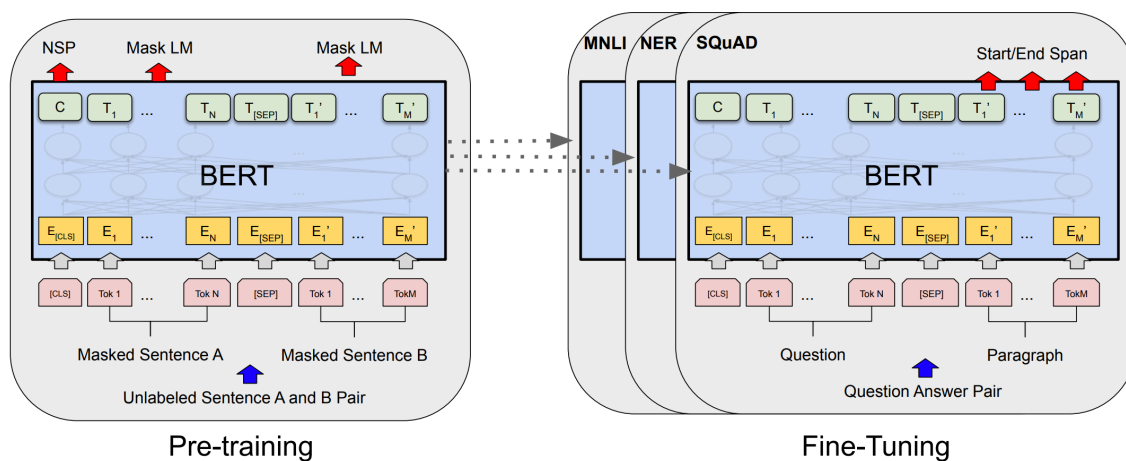


図 4.5: BERT の概要図 (文献<sup>[4]</sup> Fig 1 より)

## 5 MLC 手法を改良するためのアンサンブル学習を利用した深層言語モデルの提案

ABSA で扱われるデータセットの多くは英語文章データである. ABSA を扱う研究ワークショップである SemEval<sup>[21][31][22]</sup> でも日本語データは扱われていない. 三浦ら<sup>[6]</sup> は日本語データセットを用いた SA における MLC に取り組んでいるが, MLC モデルの識別精度は約 70% であり, 改善の余地がある. そこで, 本研究では英語文章のマルチラベルデータセットに対して, 98.6% と高い識別性能を示した三浦らのアスペクトベースセンチメント分析ネットワークモデル<sup>[2]</sup> に, アンサンブル学習のスタッキングアルゴリズム<sup>[7]</sup> を用いた深層学習と, MLC を解く手法として BR 法と Pairwise 法によるベースモデルを導入することで改良モデルを構築する. そして日本語カスタマーレビューデータに対する識別精度を向上することを目的とする.

本章ではアスペクトベースデータの MLC モデルとして, アンサンブル学習を利用した深層言語モデルの提案とアンサンブル学習のために構築したベースモデルについて述べる.

### 5.1 Entity Classifiers と Attribute Classifiers

MLC タスクにおいて 1 つの文書から複数のラベルに分類する情報を持つ分散表現を抽出することは容易ではない. そこで, ABSA におけるアスペクトの構成に着目した. アスペクトは文書の書かれたエンティティとアトリビュートによって定義されることから, エンティティとアトリビュートの分類に特化した小分類器を作成することで MLC を SLC に落とし込む.

エンティティに対応する 2 値分類器として Entity Classifiers を, アトリビュートに対応する 2 値分類器として Attribute Classifiers を提案する. 図 5.1 に Entity Classifiers と Attribute Classifiers のモデル概要図を示す. 入力文書から文書トークンを生成して, BERT に入力することで文書の分散表現を出力する. そして, 分散表現の先頭トークンである CLS トークンのみを Multi Layer Perceptron (MLP) に入力して, 2 次元の予測ラベルを出力する.

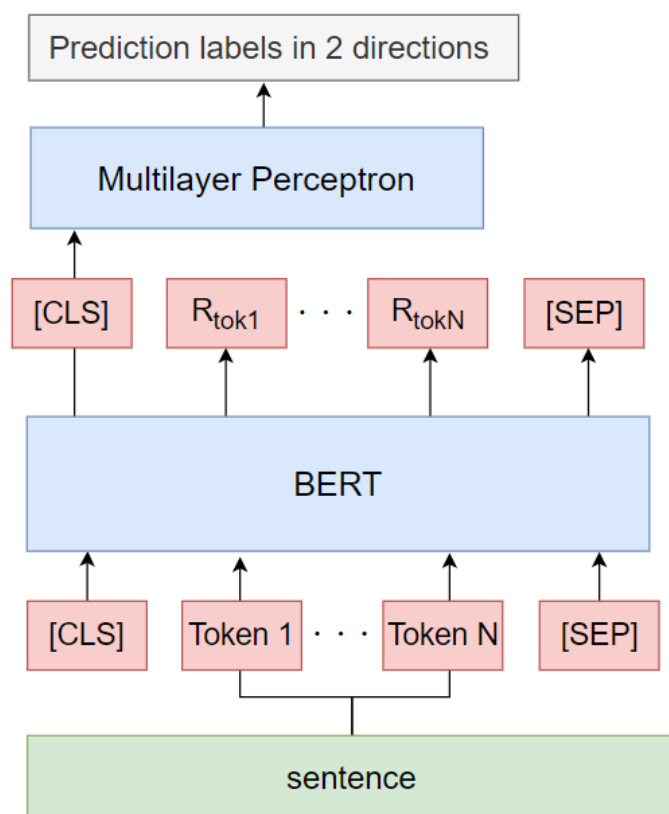


図 5.1: Entity Classifiers と Attribute Classifiers のモデル概要図

## 5.2 Entity Classifiers の作成

Entity Classifiers は BR 法に基づいて「朝食, 夕食, 風呂, サービス, 施設, 立地, 部屋」のアスペクトごとに作成した 7 個の 2 値分類器である. 文書トークンを入力することで文書の分散表現を出力する. 表 5.1 にの作成に用いたデータの例として入力文書と, 入力情報ではないが, 分類器が注目すべきエンティティを示す. Entity Classifiers が取り組むタスクは文書のエンティティからその文書が属するアスペクトを推定することである. そこで訓練用のデータセットを作成した. 「あるアスペクトのポジティブまたはネガティブに属していれば, そのアスペクトに属している」という解釈の下で, 各アスペクトのポジティブとネガティブの 2 次元のラベルを 1 次元に圧縮した. 表 5.2 に朝食アスペクト次元の圧縮例を示す. このように, あるアスペクトに属するならラベル 1 を, 属さないならラベル 0 を割り当てることで 30000 件のデータセッ

表 5.1: Entity Classifiers の作成に用いたデータの例と, 分類器が注目すべきエンティティ

文章	温泉も気持ちよく朝のバイキングの 名古屋メシはサイコーでした.	エンティティ
朝食	1	朝のバイキング
夕食	0	-
風呂	1	温泉
サービス	0	-
立地	0	-
施設	0	-
部屋	0	-

表 5.2: 朝食アスペクトの次元の圧縮例

朝食			朝食
ポジティブ	ネガティブ		
1	0	→	1
0	1	→	1
0	0	→	0

トを作成した.

Entity Classifiers が取り組むタスクを例を用いて説明する. 図 5.2 に入力文書に対する Entity Classifiers の出力例を示す.

朝食の Entity Classifier は朝食に属しているか否かを分類する. この例では朝食とサービスの Entity Classifier は入力文書が朝食およびサービスに属していることを指すラベル 1 が付与される確率が高くなり, 属さないことを指すラベル 0 が付与される確率が低くなるような分散表現を出力する. 一方で朝食とサービス以外の Entity Classifier は入力文書がそれぞれのアスペクトに属していることを指すラベル 1 が付与される確率が低くなり, 属さない事を指すラベル 0 が付与される確率が高くなるような分散表現を出力する.

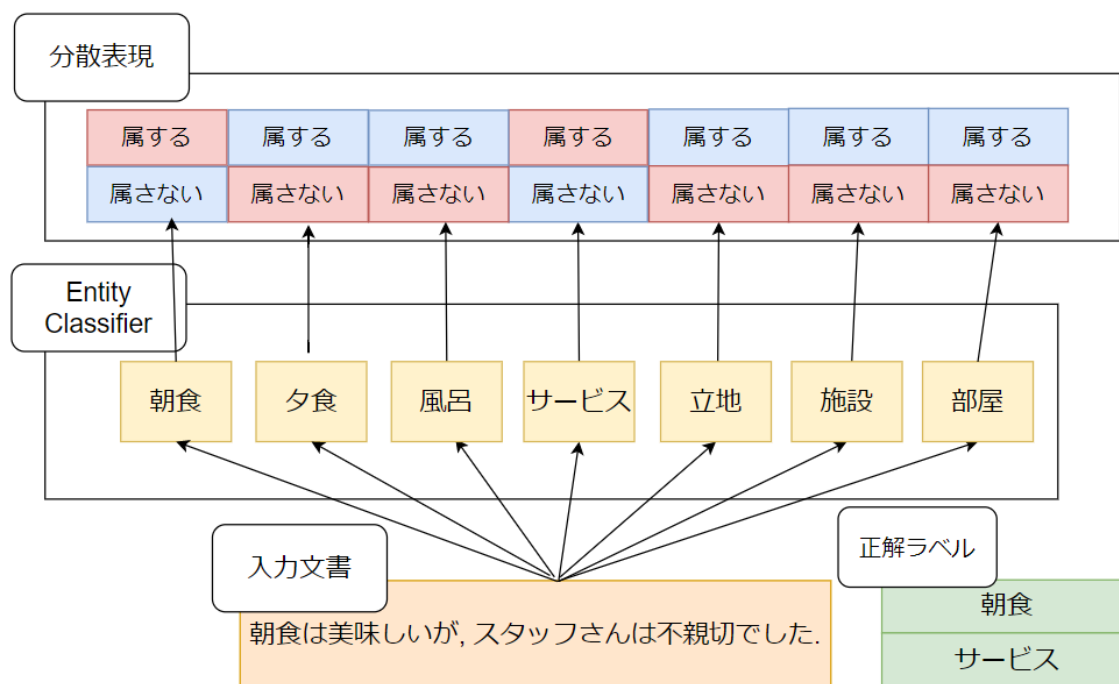


図 5.2: 入力文書に対する Attribute Classifiers の出力例

### 5.3 Attribute Classifiers の作成

Attribute Classifiers は Pairwise 法に基づいて、「朝食, 夕食, 風呂, サービス, 施設, 立地, 部屋」のアスペクトごとの感情極性としてポジティブまたはネガティブを分類する 7 個の 2 値分類器である. 表 5.3 に作成に用いたデータの例を示す. ポジティブならラベル 1 が, ネガティブならばラベル 0 が付与されている. Attribute Classifiers が取り組むタスクは, 文書のアトリビュートからあるアスペクトに関してポジティブであるかネガティブであるかどうかを 2 値分類することである. そのため, データセットから各アスペクトに対して, ポジティブまたはネガティブのラベルが付与されているデータのみを抽出して訓練用のデータセットを作成した.

Attribute Classifiers のタスクを例を用いて説明する. 図 5.3 に入力文書に対する Attribute Classifiers の出力例を示す. 朝食の Attribute Classifier は入力文書が朝食に関してポジティブかネガティブかの分類をするため, 朝食に関してポジティブラベル 1 が付与される確率が高く, ネガティブラベルの 0 が付与される確率が低くなるような分散表現を出力する. サービスの Attribute Classifier

表 5.3: Attribute Classifiers の作成に用いたデータの例 (朝食と風呂データのみを抜粋)

朝食の 文章データ	おかずの品数も多く, ある意味朝食を楽しみに 泊まっていると言ってもイイです.	チキンナゲット, ポテトが固くて 美味しくなかったのが残念です.
ラベル	1 (positive)	0 (negative)
風呂の 文章データ	コンセントも枕元にもあるし, 鏡の前にもあるしで, 使いやすい部屋でした.	大浴場や食事会場が部屋から 若干遠いかなと思いました.
ラベル	1	0

は入力文書がポジティブかネガティブかの 2 値分類をするため, サービスに関してネガティブラベル 0 が付与される確率が高くなり, ポジティブラベルの 1 付与される確率が低くなるような分散表現を出力する.

この例では, 朝食とサービス以外の分類器はポジティブまたはネガティブに分類される確率が大きく偏る分散表現は出力されないことが望ましい. しかし, 偏りが出たとしても Entity Classifiers で, 朝食とサービス以外のアスペクトに属していないことを示す分散表現を出力することによって, Transformer Encoder で確率を制御することが可能であると考えられる. つまり, Entity Classifiers は Attribute Classifiers の出力の正誤を制御する役割を持つ.

このように, アスペクトに関してポジティブであるかネガティブであるかの 2 値分類に特化させて, 該当するラベルの識別精度を強化することが, 個別にデータセットを作成して Attribute Classifiers を構築した目的である.

## 5.4 Multi pre-trained models Transformer (Mpm+T)

本研究では, MLC 手法を改良するためにアンサンブル学習を利用した深層言語モデルである, Multi pre-trained models Transformer (Mpm+T) モデルを提案する. 図 5.4 に本研究で作成した Mpm+T のモデル概要図を示す.

Mpm+T はアンサンブル学習におけるスタッキングアルゴリズム<sup>[7]</sup>を用いて構築している. 7 個の Entity Classifiers と 7 個の Attribute Classifiers は 4 章の図 4.1 のスタッキングアルゴリズムと深層モデルを組み合わせたネットワーク

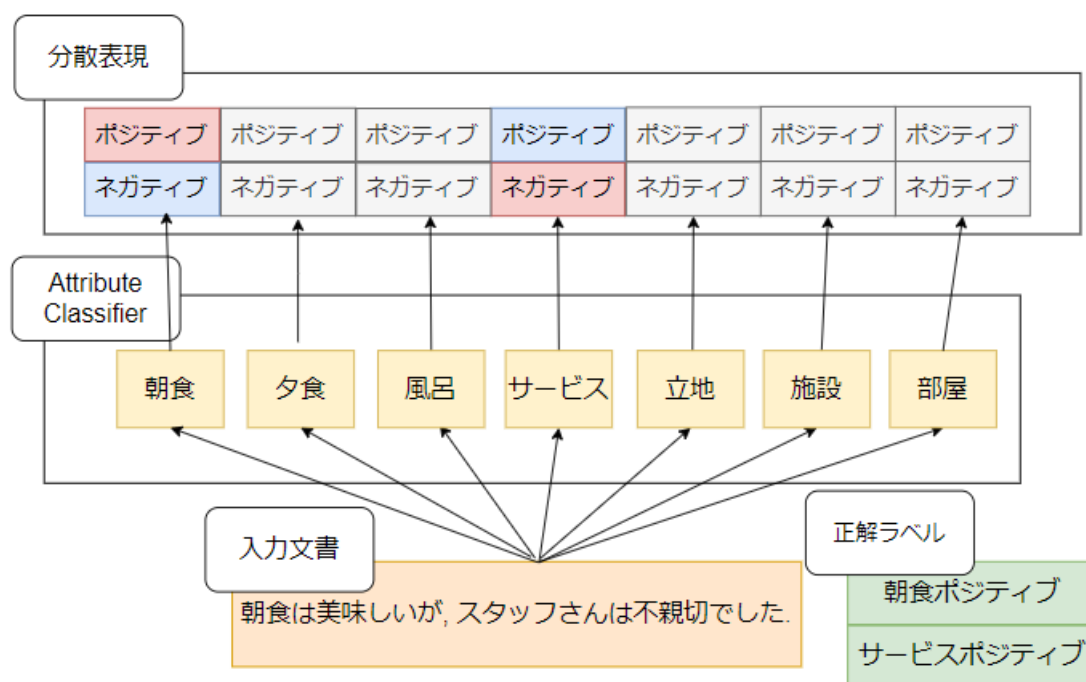


図 5.3: 入力文書に対する Attribute Classifiers の出力例

モデルにおけるベースモデル (Ensemble model) に該当し, Transformer Encoder はメタ学習器 (Meta Classifier)<sup>[29]</sup> に該当する. つまり, それぞれのアスペクトに対応する 14 個のベースモデルにメタ学習器である Transformer Encoder を接続したモデルである. また, 図 5.5 に Mpm+T におけるデータの流れを示す. モデルにおける CLS は BERT で使用されているタグの分散表現ベクトルである.

概要図に基づいてデータの流れを説明する. まず, はじめに入力文書から文書トークンを生成する. 次にそれらのトークンをベースモデルである Entity Classifiers と Attribute Classifiers に入力して, それぞれから文書の分散表現の先頭トークンである CLS トークンを出力する. それぞれの CLS トークンに対して, 図 5.5 に示すように Positional Embeddings を付与して Transformer Encoder へ入力する. そして, MLC 情報を持った CLS トークンを出力して, 最後に MLP に入力することで, 14 次元の予測ラベルを出力する.

本提案モデルの全体的な構造と, BERT<sup>[4]</sup> から得られる分散表現の内積計算機構として Transformer<sup>[3]</sup> を使用することは, 先行研究<sup>[2]</sup>である三浦らの提案

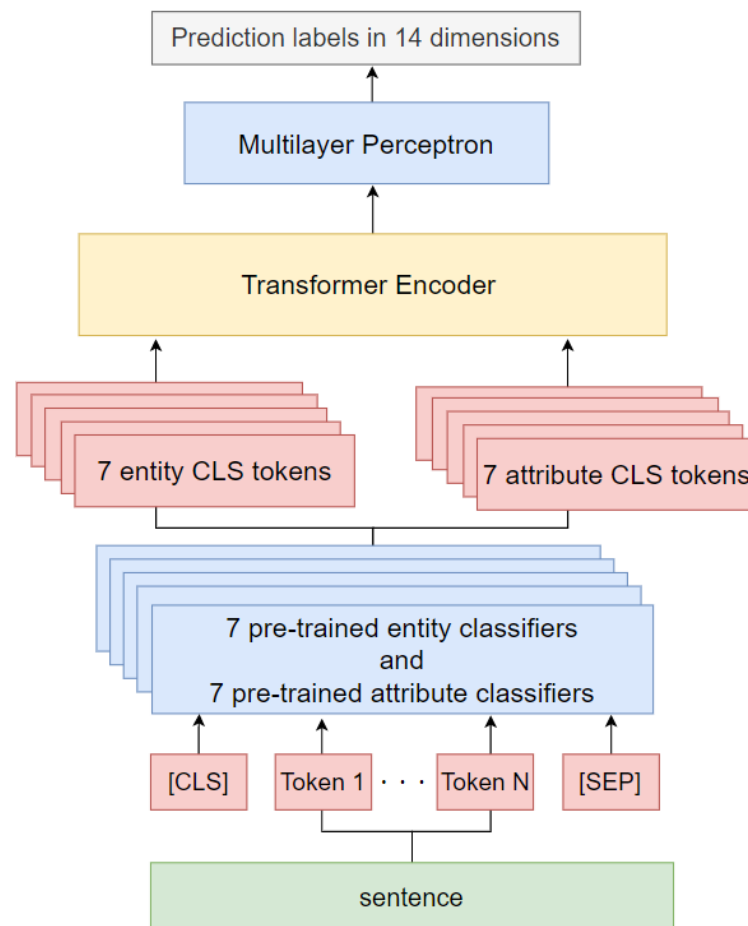


図 5.4: 提案モデル Multi pre-trained models  
Transformer (Mpm+T) の概要図

モデルであるアスペクトベースセンチメント分析ネットワークに倣っている。以下に本モデルと三浦らのモデルを比較した場合の 2 つの新規性を挙げる。

#### 5.4.1 新規性 1 : アスペクトの分類に特化した小分類器の導入

1 点目は, ABSA における MLC を SLC に落とし込んで解くために, BR 法で作成した Entity Classifiers と Pairwise 法で作成した Attribute Classifiers をアンサンブル的に導入したことである。

三浦らのモデルは単一の BERT で MLC モデルを作成して Transformer Encoder



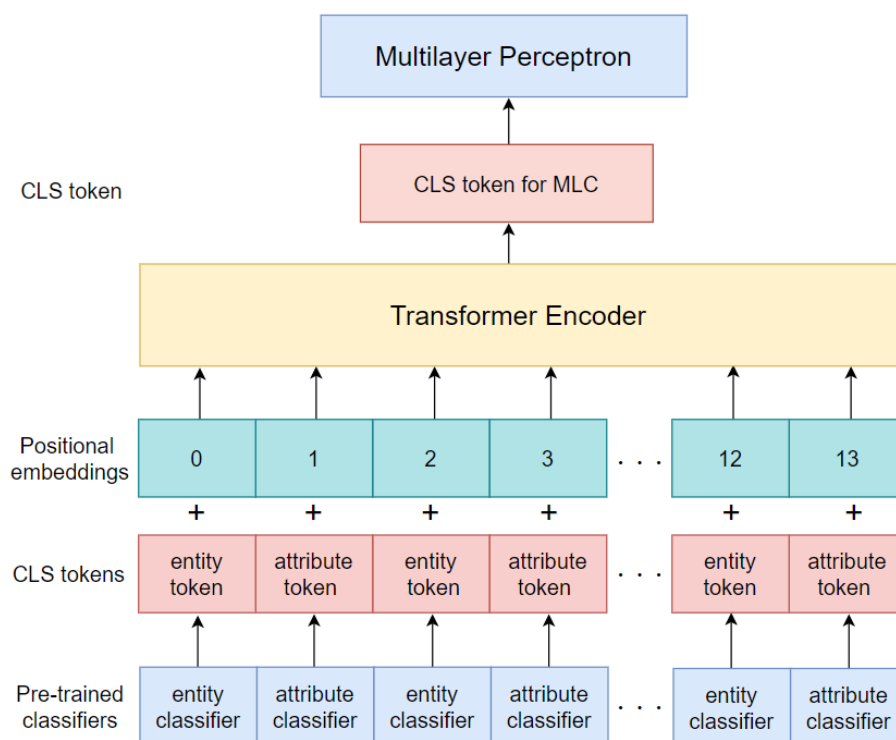


図 5.5: Mpm+T におけるデータの流れ

を接続している。三浦らのモデルは Transformer Encoder をマルチタスクモデルとすることで, MLC とエンティティとなるワード、フレーズの抽出に取り組んでいる。本研究は 4 章の図 4.1 に示すスタッキングアルゴリズムを用いた深層言語モデルを構築することによって, MLC の日本語カスタマーレビューデータに対する分類精度を改善することに焦点を当てている。そのため, Transformer Encoder をシングルタスクモデル (メタ学習器) として, 単一の BERT ではなく複数の BERT ベースのベースモデルを導入した。ベースモデルは各アスペクトの分類情報を持つ CLS トークンを出力して, それらをアンサンブルして Transformer Encoder に入力する。

#### 5.4.2 新規性 2 : Transformer Encoder への入力分散表現の限定

2 点目は, Transformer Encoder での計算量削減を目的として, ベースモデルから出力される文書の分散表現のうち, CLS トークンのみを入力とすることである。

提案手法には欠点がある。それはアスペクトの量に比例してベースモデルが多くなり、それらを作成するコストや Transformer Encoder の計算量が大きくなることである。そこで、本研究の目的である MLC の精度向上に必要となる情報のみを特徴量として Transformer Encoder への入力とすることで計算量の削減をした。

一般的な BERT や Transformer によるテキスト分類タスクでは文書トークンの分散表現を全て用いる。これは注目すべきトークンの分散表現に重みをかける学習をさせるためである。しかし Mpm+T では、Transformer Encoder の Attention 機構によって、各アスペクトに対応する分類情報を持つ CLS トークンの中の、注目すべきトークンに重みをかける学習をさせる。

表 6.1: Entity Classifiers と Attribute Classifiers の作成に用いたデータ数と、それぞれのアスペクトにおける正例と負例の数

ラベル	Entity Classifiers			Attribute Classifiers		
	データ数	正例	負例	データ数	正例	負例
朝食	30000	8494	21506	11924	10125	1799
夕食	30000	6456	23544	10167	8423	1744
風呂	30000	5638	24362	9014	6922	2092
サービス	30000	11278	18722	17980	12722	5258
立地	30000	3865	26135	6274	5417	857
施設	30000	8152	21848	13158	8025	5133
部屋	30000	6033	23967	9847	7505	2342

## 6 ベースモデルの数値実験

ベースモデルである Entity Classifiers と Attribute Classifiers の識別精度を検証するために、訓練と検証データの比率を 4:1 として 5 分割交差検証で精度を確かめた。ベースモデルの実験条件と、実験結果と考察について述べる。

### 6.1 実験条件

BR 法に基づいた BERT ベースの小分類器である Entity Classifiers と、Pairwise 法に基づいた小分類器である Attribute Classifiers の識別精度を検証するために、訓練と検証データの比率を 4:1 として 5 分割交差検証で精度を確かめた。表に 7.1 に Entity Classifiers と Attribute Classifiers の作成に用いたデータ数と、それぞれのアスペクトにおける正例と負例の数を示す。Entity Classifiers のデータはいずれのアスペクトにおいても負例（そのアスペクトに該当するエンティティを含まないという意味）が多く、Attribute Classifiers のデータは正例（ポジティブなアトリビュートを含むという意味）が多いことが確認できる。また、表 7.2 に Entity Classifiers と Attribute Classifier の実験条件を示す。

表 6.2: Entity classifiers と Attribute Classifiers の実験条件

モデル	Entity Classifier	Attribute Classifier
入力次元数	768	768
出力次元数	2	2
最適化関数	Adam	Adam
バッチサイズ	40	40
損失関数	Cross-entropy Loss	Cross-entropy Loss
学習率	0.0001	0.0001
エポック数	20	12

表 6.3: Entity Classifiers と Attribute Classifiers の正解率と標準偏差

	朝食	夕食	風呂	サービス	立地	施設	部屋
Entity Classifiers	0.916 $\pm$ 0.002	0.923 $\pm$ 0.006	0.943 $\pm$ 0.006	0.888 $\pm$ 0.005	0.951 $\pm$ 0.001	0.868 $\pm$ 0.01	0.919 $\pm$ 0.001
Attribute Classifiers	0.936 $\pm$ 0.001	0.922 $\pm$ 0.005	0.917 $\pm$ 0.001	0.926 $\pm$ 0.001	0.925 $\pm$ 0.003	0.908 $\pm$ 0.094	0.905 $\pm$ 0.007

## 6.2 実験結果と考察

表 7.2 に正解率の平均と標準偏差を示す. ほとんどのアスペクトにおいて小分類器は高い精度で 2 値分類ができている. しかし, Entity Classifiers の「サービス」と「施設」の正解率は他のアスペクトに比べて低くなっている. これは Entity Classifiers の予測が 0 (そのアスペクトに属さない) に寄ってしまう傾向があることと, 「サービス」と「施設」のアスペクトにおける正例と負例の割合に起因する. 表 7.1 の Entity Classifiers の各アスペクトに付与されている正例の 30000 件のデータに対する割合の平均は 0.237 となっている一方で, サービスと施設はそれぞれ 0.372, 0.272 と平均を上回る割合である. そのため, 予測が 0 に寄る傾向がある中で正例の多さが分類を難しくしたと考えられる. これらのモデルを Mpm+T におけるベースモデルとして扱う.

## 7 Mpm+T の有効性を検証する数値実験

Mpm+T の識別精度を検証するために, 訓練と検証データの比率を 4:1 として 5 分割交差検証で精度を確かめた. また, 2 つの比較手法を用意して同様に 5 分割交差検証した. 比較手法を紹介して, 各手法の実験条件と, 実験結果と考察について述べる.

### 7.1 比較手法

提案手法に対する比較手法として, 2 種類を挙げる.

1 つ目は従来手法として BERT+MLP を選択した.

2 つ目に三浦らのアスペクトベースセンチメント分析ネットワークの模倣モデル<sup>[2]</sup>を選択した. モデル全てを再現するのではなく, 2 章の図 2.2 の MLC タスクを解く部分のみを模倣した. 本研究ではこれを三浦らのモデルとする. 図 8.1 に三浦らのモデル概要図を示す.

モデルにおけるデータの基本的な流れは 5 章で述べた Mpm+T と同様である. 異なるのは, Mpm+T におけるベースモデルが単一の BERT であるという点と, Transformer Encoder に全ての文書トークンの分散表現を入力する点である.

### 7.2 実験条件

表に実験に用いたデータ数と正例と負例の数を示す. 各アスペクトは負例数が多いが, 正例数はポジティブの方がネガティブよりも多いことが確認できる.

表 7.2 に Mpm+T と三浦らの手法の実験条件を示す. BERT+MLP は BERT の入力次元数が 768, 出力次元数が 14, バッチサイズが 40, 損失関数は Cross-entropy Loss, 学習率は  $7.6 \times 10^{-6}$ , エポック数は 20 である.

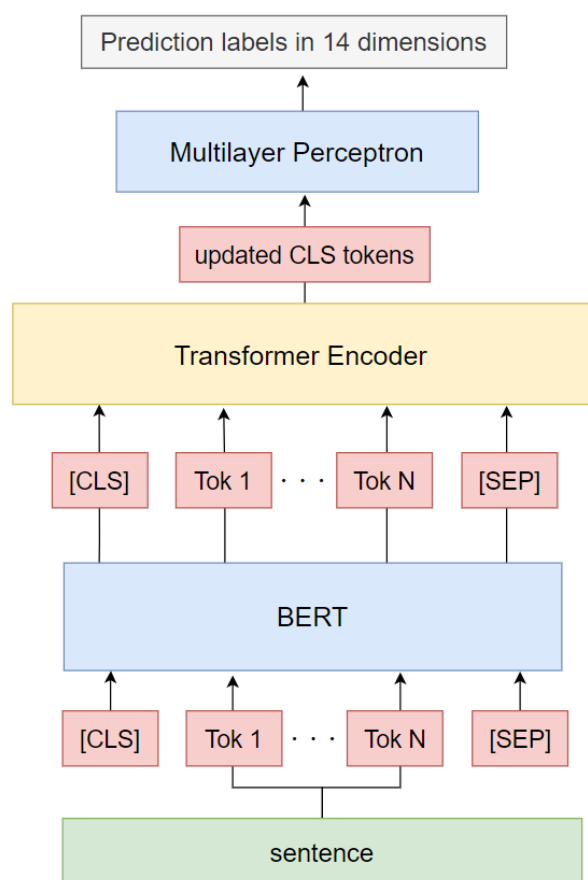


図 7.1: 三浦らのモデル概要図

### 7.3 実験結果と考察

表 7.3 に 5 分割交差検証の検証データ 6000 件でのモデルの評価指標の平均と標準偏差を示す. 関連研究<sup>[2]</sup>と同様に各モデルの micro-F1 を比較すると, Mpm+T は三浦らのモデルより 8.5% 向上した. Mpm+T と三浦らのモデルを比較して MLC の精度が向上した最大の要因は, アンサンブル学習の導入によりベースモデルを構築したことであると考えられる. しかし, この結果のみでは提案手法の有効性を確認できないと考えたため, テストデータを用意して更なる考察をする.

表 7.1: 実験で用いたデータ数と, それぞれのアスペクトの正例と負例の数

データ数		30000	
ラベル		正例	負例
朝食	ポジティブ	6979	23021
	ネガティブ	1515	28485
夕食	ポジティブ	5293	24707
	ネガティブ	1163	28837
風呂	ポジティブ	4258	25742
	ネガティブ	1380	28620
サービス	ポジティブ	7843	22157
	ネガティブ	3435	26565
立地	ポジティブ	3307	26693
	ネガティブ	558	29442
施設	ポジティブ	4892	25108
	ネガティブ	3260	26740
部屋	ポジティブ	4509	25491
	ネガティブ	1524	28476

## 7.4 新たな評価指標の定義とテストデータによる考察

### 7.4.1 完全正解 (Completely Accuracy) の定義

MLC の評価指標として一般的に用いられる micro-F1 は, 全予測ラベルのうち正しく予測できたラベルの割合を示す. 本研究で扱うデータセットのようにラベル数が多く, 負例の割合が高い場合では手法の有効性を述べる評価指標としては不十分であると考えた. そこで, 本研究で用いる評価指標として完全正解 (Completely Accuracy) を定義する. 完全正解とは, 1 つの正解データとそれに対する予測データの 14 個のラベルが全て一致した場合を指す評価指標である.

表 7.2: Mpm+T と三浦らのモデルの実験条件

モデル	Mpm+T	三浦らのモデル
ベースモデル (BERT) の入力次元数	768	768
ベースモデルの入力トークン数	100	100
ベースモデルの出力次元数	768	768
ベースモデルの出力トークン数	14	100
Transformer Encoder の入力次元数	768	768
Transformer Encoder の入力トークン数	14	100
Transformer Encoder の出力次元数	14	14
Transformer Encoder のヘッド数	4	4
Transformer Encoder のレイヤ数	3	2
バッチサイズ	4	40
最適化関数	Adam	Adam
損失関数	Binary Cross-entropy Loss	Binary Crossentropy Loss
学習率	$7.6 \times 10^{-6}$	$5.7 \times 10^{-5}$
エポック数	15	20

#### 7.4.2 テストデータによる考察

モデルの詳細な分析をするために、テストデータ 6000 件を用いて各モデルの評価をする。表 7.4 に 6000 件のテストデータに含まれるシングルラベルデータとマルチラベルデータの数と、マルチラベルデータに含まれるそれぞれのラベルごとのデータ数を示す。

表 7.5 にテストデータ 6000 件において、各モデルの完全正解したマルチラベルデータとシングルラベルデータの数と割合を示す。

Mpm+T の完全正解データ数はマルチラベルとシングルラベル共に三浦らのモデルを上回っていることが確認できる。特にマルチラベルデータの完全正解データ数の向上が著しく、データ数の割合としては BERT+MLP より約 0.2、三浦らのモデルより約 0.1 向上している。このことから Mpm+T は比較手法よりも MLC に有効なモデルであるといえる。続いてマルチラベルの完全正解



表 7.3: 各 MLC モデルの評価指標の平均と標準偏差

Evaluation indices	Precision	Recall	micro-F1
Mpm+T	0.846 $\pm$ 0.019	0.872 $\pm$ 0.028	0.858 $\pm$ 0.017
BERT+MLP	0.683 $\pm$ 0.005	0.772 $\pm$ 0.009	0.724 $\pm$ 0.005
三浦らのモデル	0.741 $\pm$ 0.015	0.804 $\pm$ 0.006	0.773 $\pm$ 0.013

表 7.4: 6000 件のテストデータに含まれるシングルラベルデータとマルチラベルデータの数と, マルチラベルデータに含まれるそれぞれのラベルごとのデータ数

	シングルラベルデータ	マルチラベルデータ					
ラベル数	1	2	3	4	5	6	7
データ数	3940	1364	320	73	20	6	294

データについて考察する. 表 7.6 それぞれのラベル数のマルチラベルデータにおける完全正解数とその割合を示す. ほとんどのラベル数において, Mpm+T が最も完全正解数が多いことが確認できる. しかし, 完全正解割合が最も高いのは三浦らのモデルのラベル数の場合の 0.707 であった. ラベル数の 7 のデータは, 3 章で述べたようにデータセットの約 97% が全てのアスペクトがポジティブまたはネガティブとなっているため, ラベル数の多さとは対照的に予測が容易であったと考えられる. 他のラベル数は増加と共に完全正解数が減少しており, ラベル数が 6 の場合ではいずれのモデルでも正解数は 0 であった. 表 7.7 にラベル数が 6 の場合の文書と, 正解ラベルと予測ラベルを示す. 左の文書では「部屋ポジティブ」のラベルを負例と予測して誤っている. 右の文書では「風呂ポジティブ」のラベルを正例と予測して誤っている. この例文のように, ラベル数 6 の文書は人の判断によってもラベル付与の判断が難しいと感じるようなデータであった. また, データセットにおけるデータ数の少なさも分類を難しくしたと考えられる.

最後に, 各アスペクトごとの F1 について考察する. 表 7.7 に各アスペクトごとの F1 を示す. 結果, 全てのアスペクトにおいて三浦らのモデル以上の精度でラベルを分類できている事が確認できた. 最も分類精度が向上したアスペクトは, 「立地ネガティブ」であり, 三浦らのモデルと比較して 0.562 から 0.708 に向上した. 他に分類精度が 0.5 以上向上したアスペクトは「サービス

表 7.5: ラベル数が 6 の場合の文書と、正解ラベルと予測ラベル

		文書			
		本当に至れり尽くせり, お世話になりました.		金沢市内だと一泊素泊まりで 12000 円ほどするホテルがあること を思えば, 大満足の一泊でした.	
		正解ラベル	予測ラベル	正解ラベル	予測ラベル
朝食	ポジティブ	1	1	1	1
	ネガティブ	0	0	0	0
夕食	ポジティブ	1	1	1	1
	ネガティブ	0	0	0	0
風呂	ポジティブ	1	1	0	1
	ネガティブ	0	0	0	0
サービス	ポジティブ	1	1	1	1
	ネガティブ	0	0	0	0
立地	ポジティブ	0	0	1	1
	ネガティブ	0	0	0	0
施設	ポジティブ	1	1	1	1
	ネガティブ	0	0	0	0
部屋	ポジティブ	1	0	1	1
	ネガティブ	0	0	0	0

ネガティブ」, 「立地ポジティブ」, 「設備ポジティブ」, 「設備ネガティブ」, 「部屋ネガティブ」である. 全体として分類精度が向上したものはネガティブに関するアスペクトが多いことが確認できる. しかし, 分類精度としては高いとはいえない. これは 3 章の表 3.5 に示したように, ネガティブのアスペクトはいずれもポジティブのものよりも正例の数が少なく, 予測が負例に寄る傾向があるためだと考えられる.

上述の 5 分割交差検証結果とテストデータによる分析結果から, 提案手法として ABSA にアンサンブル学習を導入する有効性を確認した.

表 7.6: 完全正解データを占めるマルチラベルデータと  
シングルラベルデータのデータ数と割合

データの種類	完全正解	
	マルチラベル データ	シングルラベル データ
Mpm+T	1298 (0.630)	3013 (0.764)
BERT+MLP	862 (0.418)	2326 (0.641)
三浦らのモデル	1097 (0.533)	2787 (0.706)

表 7.7: それぞれのラベル数のマルチラベルデータにおける完全正解数とその割合

ラベル数	マルチラベルデータ					
	2	3	4	5	6	7
Mpm+T	879 (0.644)	173 (0.541)	35 (0.479)	10 (0.500)	0	203 (0.690)
BERT+MLP	600 (0.440)	83 (0.259)	18 (0.247)	3 (0.150)	0	158 (0.537)
三浦らのモデル	737 (0.540)	136 (0.425)	30 (0.411)	5 (0.250)	0	208 (0.707)

表 7.8: それぞれのモデルにおける各アスペクトの F1

		Mpm+T	BERT+MLP	三浦らのモデル
朝食	ポジティブ	0.901	0.848	0.885
	ネガティブ	0.695	0.562	0.691
夕食	ポジティブ	0.878	0.803	0.858
	ネガティブ	0.702	0.394	0.680
風呂	ポジティブ	0.895	0.843	0.895
	ネガティブ	0.707	0.559	0.703
サービス	ポジティブ	0.881	0.823	0.865
	ネガティブ	0.824	0.656	0.746
立地	ポジティブ	0.879	0.799	0.815
	ネガティブ	0.708	0.467	0.562
施設	ポジティブ	0.828	0.728	0.765
	ネガティブ	0.764	0.636	0.709
部屋	ポジティブ	0.874	0.768	0.838
	ネガティブ	0.726	0.474	0.654

## 8 提案手法 2

7 章で ABSA における MLC にアンサンブル学習を導入する手法の有効性を確認した。しかし、ネガティブに関するアスペクトの分類精度が高くないことが問題である。そこで、アトリビュートの識別精度を向上させるためにベースモデルである Attribute Classifiers に焦点を当てる。

3 章で示したように、データセット 48353 件に付与されているラベル数の平均は 1.66 個であり、データの 14 個のラベルのほとんどを負例が占める。これは、ほとんどのデータは 13 個あるいは 12 個のアスペクトに属していないことを表している。つまり、ほとんどのデータは 13 個あるいは 12 個のアスペクトに関してポジティブでもなくネガティブでもないため、ニュートラルであると言える。そのため、Mpm+T を学習する際は、Attribute Classifiers はそれぞれのアスペクトには該当しない文書を入力とする場合が多い。そのような文書に対して出力する分散表現をより正確にすることが MLC の精度向上に寄与すると仮説を立てた。そこで、ポジティブラベル (1)、ネガティブラベル (0) に対してニュートラルラベル (2) を追加することで、ラベル数を 2 クラス 3 クラスに拡張したデータセットを作成して、アトリビュートに関して 3 値分類するベースモデルを導入する。これは入力文書から 3 値分類情報を持つ分散表現を出力することで、MLC の精度を更に向上させることを目的としている。そして、Mpm+T の Attribute Classifiers のみを 3 値分類器に置換するモデルを提案する。

### 8.1 3 値分類 Attribute Classifiers の作成

アトリビュートに関するデータのラベル数をポジティブラベル (1)、ネガティブラベル (0) の 2 クラス から、ニュートラルラベル (2) を加えて 3 クラスに拡張したデータセットを作成して、アトリビュートに関して 3 値分類する 3-class Attribute Classifiers を提案する。

図 8.1 に 3-class Attribute Classifiers のモデル概要図を示す。データの入力などの流れは Entity Classifiers および Attribute Classifiers と同様である。しかし出力次元数が異なり、3 次元となっている。

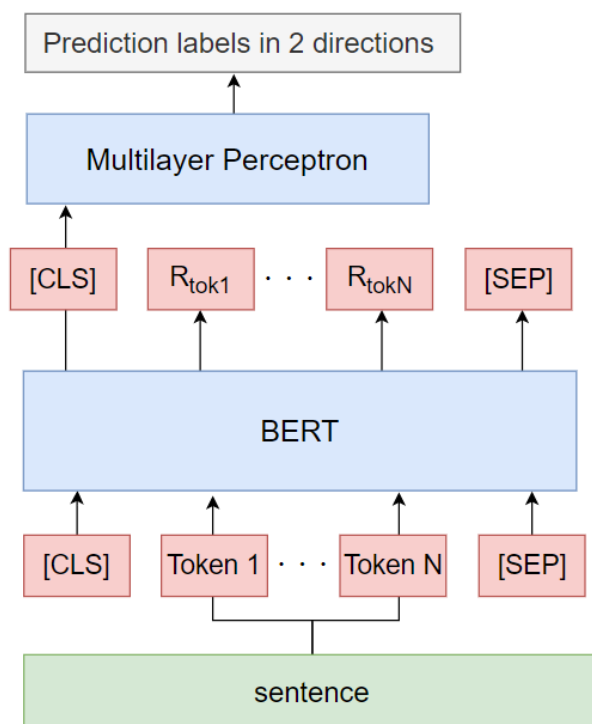


図 8.1: 3-class Attribute Classifiers のモデル概要図

本提案の最大の目的は、アトリビュートをのクラス数を拡張することでベースモデルが出力する情報量を増加させることである。図 9.2 に入力文書に対する 3-class Attribute Classifiers の出力例を示す。また、学習に必要なデータ数の確保も目的の 1 つである。6 章の表 6.1 に示す Attribute Classifiers の作成に用いたデータ数を参照すると、最大件数のサービスの 17980 件と最小件数の立地の 6274 件で 12000 件近くの差がある。このようなデータ数の偏りを解消することも本提案の目的の 1 つである。

表 8.1 に 3-class Attribute Classifiers の作成に用いたデータの例として「夕食」と「施設」データを抜粋して示す。表 8.1 のようにポジティブ (1) またはネガティブ (0) またはニュートラルのラベル (2) を付与したデータセットを「朝食, 夕食, 風呂, サービス, 施設, 立地, 部屋」の 7 個のアスペクトごとに 30000 件作成した。

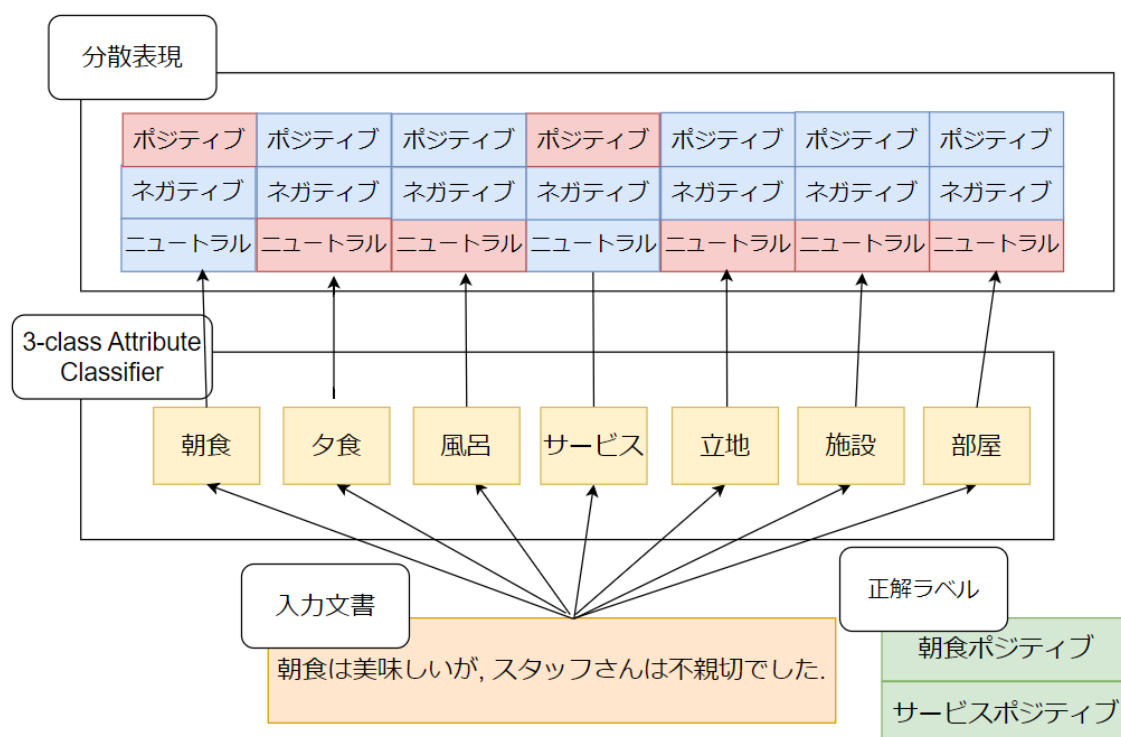


図 8.2: 入力文書に対する 3-class Attribute Classifiers の出力例

## 8.2 Multi-pretrained models Transformer2 (Mpm+T2)

Mpm+T における Attribute Classifiers を全て 3-class Attribute Classifiers に置換したモデルとして Multi-pretrained models Transformer (Mpm+T2) を提案する. 図 8.2 に Mpm+T2 のモデル概要図を示す.

Mpm+T2 の入力と出力は Mpm+T と同様であり, データの流れも図 5.3 と同様である. ただし, 置換したのは Attribute Classifiers のみであり, Entity Classifiers の変更はない. Mpm+T2 もアンサンブル学習におけるスタッキングアルゴリズムを用いて構築している. 7 個の Entity Classifiers と 7 個の 3-class Attribute Classifiers は 4 章の図 4.1 のスタッキングアルゴリズムと深層モデルを組み合わせたネットワークモデルにおけるベースモデルに該当し, Transformer Encoder はメタ学習器に該当する.

表 8.1: 3-class Attribute Classifiers の作成に用いたデータの例  
(「夕食」と「施設」データを抜粋)

夕食の 文書データ	夕食はホテル内の居酒屋で済ませましたが 金額も良心的で良かったです.	お部屋も和洋室で広く, ガス台, レンジもあったり親も子供も大喜び.
ラベル	1 (positive)	2 (neutral)
施設の 文書データ	洗面化粧台の電球が切れていましたが, すぐに対応して頂きました.	ステーキは美味しかったです.
ラベル	0 (negative)	2

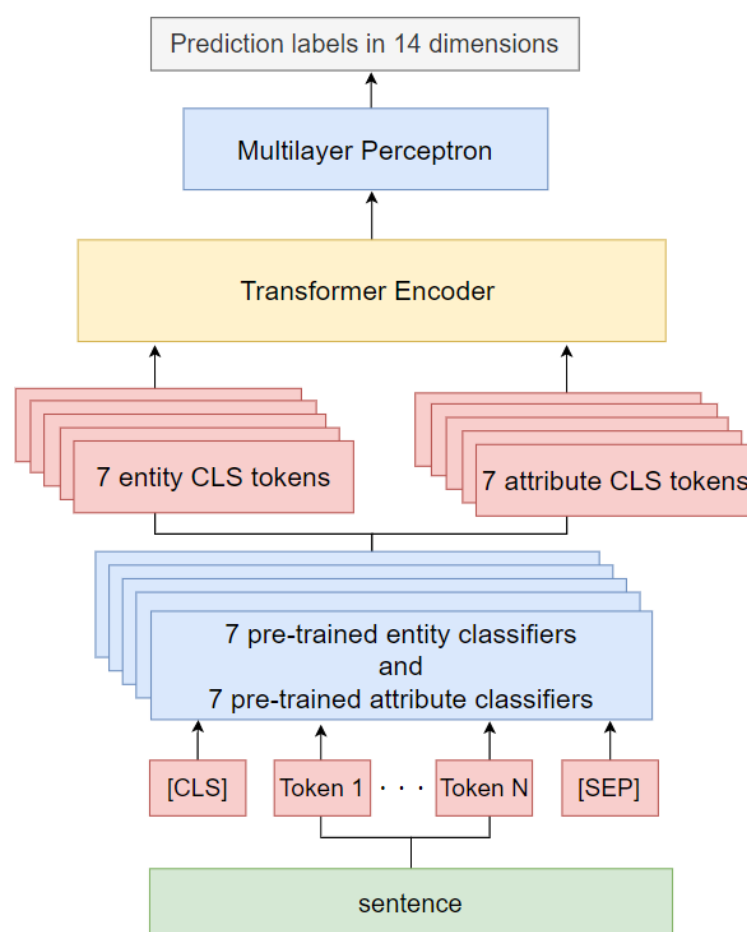


図 8.3: 提案モデル Multi pre-trained models  
Transformer2 (Mpm+T2) の概要図



表 9.1: 30000 件のデータを占めるポジティブ, ネガティブ, ニュートラルラベルデータの数

	ポジティブラベル	ネガティブラベル	ニュートラルラベル
朝食	6979	1525	21506
夕食	5293	1163	23544
風呂	4258	1380	24362
サービス	7843	3435	18722
立地	3307	558	26135
施設	4892	3260	21848
部屋	4509	1524	23967

## 9 ベースモデルの 3-class Attribute Classifiers の数値実験

Mpm+T2 のベースモデルの 1 つである 3-class Attribute Classifiers の識別精度を検証するために, 訓練と検証データの比率を 4:1 として 5 分割交差検証で精度を確かめた. 3-class Attribute Classifiers の実験条件と, 実験結果と考察について述べる.

### 9.1 実験条件

表 10.1 に実験に用いた 30000 件のデータを占める, ポジティブ, ネガティブ, ニュートラルラベルデータの数を示す. いずれのアスペクトでもニュートラルラベルが多い. また, ポジティブラベルの方がネガティブよりも多い事が確認できる. 最もポジティブラベルの割合が高いのは「サービス」の 6979 件であり, 最もその割合が低いのは「立地」の 3307 件である. また, 最もネガティブラベルの割合が高いのは「サービス」の 3435 件であり, 最もその割合が低いのは「立地」の 558 件である. 最後に, 最もニュートラルラベルの割合が高いのは「立地」の 26135 件であり, 最もその割合が低いのは「サービス」の 18722 件である. つまり, ポジティブまたはネガティブラベルが付与されているデータの割合が最も高いのは「サービス」であり, それらのラベルが付与されているデータの割合が最も低いのは「立地」である. 表 10.2 に実験条件

表 9.2: 実験条件

モデル	朝食	夕食	風呂	サービス	立地	施設	部屋
BERT の入力次元数	768						
BERT の出力次元数	3						
最適化関数	Adam						
損失関数	Cross-entropy Loss						
学習率	$1.0 \times 10^{-4}$	$1.0 \times 10^{-5}$	$9.0 \times 10^{-6}$	$1.0 \times 10^{-3}$	$6 \times 10^{-5}$	$1 \times 10^{-4}$	$3 \times 10^{-5}$
エポック数	20						

表 9.3: 3-class Attribute classifiers と Attribute Classifiers の正解率と標準偏差

	朝食	夕食	風呂	サービス	立地	施設	部屋
3-class Attribute Classifiers	$0.827 \pm 0.019$	$0.805 \pm 0.023$	$0.848 \pm 0.025$	$0.810 \pm 0.004$	$0.850 \pm 0.020$	$0.830 \pm 0.007$	$0.813 \pm 0.020$
Attribute Classifiers	$0.936 \pm 0.001$	$0.922 \pm 0.005$	$0.917 \pm 0.001$	$0.926 \pm 0.001$	$0.925 \pm 0.003$	$0.908 \pm 0.094$	$0.905 \pm 0.007$

を示す。データのラベルが不均衡であるため、損失関数のパラメータについて調整した。表 10.1 の各データセットにおける各クラスのデータ数の逆数を重みとして掛けて、損失への寄与率をデータ数に反比例するように調整した。

## 9.2 実験結果と考察

表 10.3 に 5 分割交差検証の検証データ 6000 件での 3-class Attribute Classifiers の正解率の平均と標準偏差に加えて、7 章の表 7.3 より Attribute Classifiers の正解率の平均と標準偏差を示す。

いずれのアスペクトにおいても識別精度は低下していることが確認できる。3-class Attribute Classifiers の精度が最も高いアスペクトは「立地」の 0.850 であり、最も低いのは「夕食」の 0.805 である。精度の低下が最も小さいアスペクトは「風呂」の 0.69 であり、最も大きいのは「夕食」の 0.117 である。精度が 0.100 以上低下しているアスペクトは「朝食, 夕食, サービス」である。全体として精度が低下した原因はニュートラルラベルに予測が寄りにくくなるように損失関数の重みを調整したことで、ニュートラルラベルの正解率が低下したためであると考えられる。作成した 3-class Attribute Classifiers と 7 章で作成した Entity Classifiers を Mpm+T2 におけるベースモデルとして扱う。

表 10.1: Mpm+T2 の実験条件

モデル	Mpm+T2
ベースモデル (BERT) の入力次元数	768
ベースモデルの入力トークン数	100
ベースモデルの出力次元数	768
ベースモデルの出力トークン数	14
Transformer Encoder の入力次元数	768
Transformer Encoder の入力トークン数	14
Transformer Encoder の出力次元数	14
Transformer Encoder のヘッド数	4
Transformer Encoder のレイヤ数	3
最適化関数	Adam
損失関数	Binary Cross-entropy Loss
学習率	$1.0 \times 10^{-6}$
エポック数	15

## 10 Mpm+T2 の有効性を検証する数値実験

Mpm+T2 の識別精度を検証するために, 訓練と検証データの比率を 4:1 として 5 分割交差検証で精度を確かめた. 検証実験によって確認する Mpm+T2 の有効性とはベースモデルの一部を 3 値分類器に置換したことである. そのため, 比較手法として Mpm+T を選択した. 実験条件と, 実験結果と考察について述べる.

### 10.1 実験条件

実験に用いたデータ数と正例と負例の数は 8 章の表 8.1 と同様である. 表 10.1 に実験条件を示す.

表 10.2: Mpm+T2 と Mpm+T の評価指標の平均と標準偏差

Evaluation indices	Precision	Recall	micro-F1
Mpm+T2	$0.821 \pm 0.007$	$0.839 \pm 0.010$	$0.812 \pm 0.012$
Mpm+T	$0.846 \pm 0.019$	$0.872 \pm 0.028$	$0.858 \pm 0.017$

## 10.2 実験結果と考察

表 10.2 に 5 分割交差検証の検証データ 6000 件での Mpm+T2 の評価指標の平均と標準偏差と, 表 7.3 より Mpm+T の結果を抜粋して示す. micro-F1 値を比較すると Mpm+T2 の精度は Mpm+T より 0.46 低下していることが確認される. 低下した理由として, Mpm+T のベースモデルの Attribute Classifiers を 3-class Attribute Classifiers に置換したことが考えられる. 有効性を十分に検証するためにテストデータを用いて Mpm+T2 の評価をする.

## 10.3 テストデータによる考察

## 11 まとめと今後の課題

### 11.1 まとめ

本研究では, アンサンブル学習を利用した深層言語モデルによるマルチラベル分類の改良手法について述べた. 手法として, Multi-pretrained models Transformer (Mpm+T) を提案して, 先行研究<sup>[2]</sup>の三浦らのモデルを模倣して識別精度を比較して分析した結果, ABSA にアンサンブル学習を利用する有効性を確認した.

手法としてアンサンブル学習のスタッキングアルゴリズムを用いており, アスペクトに基づいた小分類器である Entity Classifiers と Attribute Classifiers をベースモデルとして BERT をベースに作成した. 入力文書に対してベースモデルから分散表現を出力して, それらの特徴量ベクトルとして学習するメタ分類器として Transformer Encoder を利用した. BERT に Transformer Encoder を接続するというモデルの構築は先行研究の三浦らのアスペクトベースセンチメント分析ネットワークに倣っている.

先行研究と比較した新規性として, アスペクトに基づいた小分類器を導入することと, 小分類器から得られる CLS トークンのみを Transformer Encoder に入力することが挙げられる. 30000 件のデータで 5 分割交差検証した結果, Mpm+T と先行研究の模倣モデルを比較して, 約 77% に対して約 86% となり, 識別精度の向上を確認した. また, 評価指標として完全正解を定義して, 6000 件のテストデータを用いて各モデルの評価をした. 結果, マルチラベルデータにおける完全正解数は Mpm+T は三浦らのモデルより 200 件多く向上した. ABSA にアンサンブル学習を導入する有効性を確認して, Multi-pretrained models Transformer2 (Mpm+T2) を提案した. これは Mpm+T の改良を目的としてベースモデルの Attribute Classifiers を, ポジティブまたはネガティブまたはニュートラルラベルの 3 値分類をする 3-class Attribute Classifiers に置換したモデルである. 30000 件のデータで 5 分割交差検証した結果, Mpm+T と比較して, 約 86% に対して約 81% となり, 識別精度の向上は確認できなかった. また, 6000 件のテストデータを用いてモデルの評価をした.

## 11.2 今後の課題

今後の課題として, クレンジングで除いたデータなどを活用してモデルの汎用性を向上させることがある. また, それぞれの小分類器がアスペクト情報を抽出できていることを確認する必要がある.

## 参考文献

- [1] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.
- [2] 三浦義栄, 渥美雅保. 事前学習言語モデルを用いたアスペクトベースセンチメント分析ニューラルネットワーク. 人工知能学会全国大会論文集 第35回 (2021), pp. 2Yin507–2Yin507. 一般社団法人 人工知能学会, 2021.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [5] B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.
- [6] 三浦義栄, 赤井龍一, 渥美雅保. 文中の複数アスペクトのセンチメント分析のための自己注意ニューラルネットワーク. 人工知能学会全国大会論文集 第34回 (2020), pp. 3Rin441–3Rin441. 一般社団法人 人工知能学会, 2020.
- [7] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.
- [8] 増田和輝, 二宮順一, 斎藤武久. アンサンブル学習ニューラルネットワークを用いた富山湾沿岸波浪推算モデルの開発. 土木学会論文集 B2 (海岸工学), 76(2):I.235–I.240, 2020.
- [9] 箱石健太, 荒木健, 一言正之. スタッキングの適用による河川水位の時系列予測. AI・データサイエンス論文集, 1(J1):453–458, 2020.

- [10] 池田欽一. ディープラーニング株価予測のアンサンブル学習による精度向上についての分析. 2021.
- [11] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12:191–202, 2018.
- [12] B. Lauser and A. Hotho. Automatic multi-label subject indexing in a multilingual environment. In *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003 Trondheim, Norway, August 17-22, 2003 Proceedings 7*, pp. 140–151. Springer, 2003.
- [13] 吉村皐亮, 馬場雪乃, 鹿島久嗣. 相互依存モデルによるマルチラベル分類. 人工知能学会全国大会論文集 第 31 回 (2017), pp. 2K21in1–2K21in1. 一般社団法人 人工知能学会, 2017.
- [14] S. L. Salzberg. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993, 1994.
- [15] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.
- [16] B. Liu. Sentiment analysis: Mining opinions, sentiments, and emotions. cambridge: Cup. *Go to original source*, 2015.
- [17] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, 41:89–97, 2013.
- [18] M. Hagenau, M. Liebmann, and D. Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision support systems*, 55(3):685–697, 2013.



- [19] T. Xu, Q. Peng, and Y. Cheng. Identifying the semantic orientation of terms using s-hal for sentiment analysis. *Knowledge-Based Systems*, 35:279–289, 2012.
- [20] I. Maks and P. Vossen. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision support systems*, 53(4):680–688, 2012.
- [21] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 27–35, Dublin, Ireland, Aug. 2014. Association for Computational Linguistics.
- [22] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 19–30, San Diego, California, June 2016. Association for Computational Linguistics.
- [23] C. Sun, L. Huang, and X. Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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. 380–385, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [24] 楽天グループ株式会社. 楽天データセット（コレクション）, aug 2010.
- [25] 楽天グループ株式会社. 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, feb 2021.
- [26] O. Sagi and L. Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

- [27] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine learning*, 42:287–320, 2001.
- [28] L. Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.
- [29] 星野厚, ホシノアツシ. 人間と ai 予測サービスの多様性に基づくアンサンブル学習システムの研究. 2022.
- [30] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations, 2018.
- [31] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 486–495, Denver, Colorado, June 2015. Association for Computational Linguistics.