豊田工業大学 課題研究報告書

文書・文間及びカテゴリ間の関係を考慮したレーティング予測

平成28年 2月

工学部 先端工学基礎学科
知能数理研究室
12056 外山洋太

目次

表目次

図目次

1　序論

本研究は多カテゴリにおけるレーティング予測に関する研究である。以下に、研究背景と目的、及び、提案手法の概略、本研究の貢献を示す。最後に論文全体の構成を示す。

1.1　背景

企業がマーケティングのために行う商品の評判分析において、商品レビューに対するレーティング予測は重要な要素技術のひとつである。何万件という大量のレビューデータを人手で処理することは難しく、計算機による自動化が望まれる。その中で商品を複数のカテゴリにおいてレーティング予測をする問題がある。カテゴリとは、宿泊施設のレビューを例にすると、サービス、立地、食事等のレーティングが付けられる各項目のことである。この問題に関する従来手法 [1] は、文間の関係性を考慮しておらず、カテゴリ間については考慮しているものの複雑な関係性を捉えられていない。

近年、そのレーティング予測において、ニューラルネットワークを用いた手法 [3, 4, 5] が提案されており、従来の手法を上回る正答率を達成している。ニューラルネットワークをレーティング予測に用いる利点はまず層の数を増やすことによって入力の複雑な繋がりを考慮できることである。例えば、文毎の素性を入力とすれば文間の関係を考慮することができる。さらに、多カテゴリのレーティング予測においてはカテゴリ間の関係を考慮した予測が実現できる。しかし、レーティング予測に関する多くの研究は 1 つのカテゴリにおける二値または多値分類問題を対象としている。

文や文書の意味表現の学習手法として、単語と文書の分散表現を同時に学習するパラグラフベクトル [2] がある。これはレーティング予測において優れた性能を示している。しかし、文書全体にパラグラフベクトルを用いた場合、レーティングの予測時に文の位置関係を考慮できない。

1.2　目的

複数カテゴリにおけるレーティング予測について、文書及び文間の関係とカテゴリ間の関係を同時に考慮したものの実現を目的とする。これにより、提案手法が従来手法 [1] より高い正答率を達成することを目指す。また、実際に文書・文間の関係の考慮がレーティング予測の正答率向上に有効であるか検証する。

1.3　提案手法

提案手法はパラグラフベクトルとニューラルネットワークを用いたレーティング予測の手法である。提案手法では、まずパラグラフベクトル [2] によって各レビューの文書ベクトルと文ベクトルを生成する。次に、各レビューの文ベクトルをその位置によって重み付け平均する。これにより、文の大まかな位置関係を保持したままレビュー間の文ベクトルの数を固定にする。最後に、その文書ベクトルと重み付け平均された文ベクトルをニューラルネットワークによる分類器によって分類しレーティング予測を行う。

1.4　貢献

本研究は、多カテゴリにおけるレーティング予測について、従来手法 [1] より高い正答率を示す手法を提案した。その手法は、特にレーティング不可能という意味を持ったレーティング値を予測することにおいて従来手法より優れていることが分かった。また、レーティング予測において文書・文ベクトルを同時に用いることが有効であることを示した。これは同時に、文書ベクトルと文ベクトルがいくらか異なる特徴を捉えていることを示す。さらに、文ベクトルの位置関係の考慮がレーティング予測に重要であることを示した。

1.5　構成

本論文は 6 節からなる。1 節は本節であり、本研究の背景、目的、研究分野への貢献、論文の全体の構成について述べる。2 節では、本研究の従来手法 [1] と、提案手法がレーティング予測のために用いるいくつかの手法、研究について述べる。3 節では、提案手法の基礎となる考えとその具体的なアルゴリズムについて説明する。4 節では、提案手法と従来手法、及び、3 つの比較手法を用いた実験の実験設定と結果について説明する。5 節では、実験結果とそれによって明らかとなった提案手法の性質について考察する。また、明らかとなった提案手法の問題点からその改善方法について議論する。6 節では、まとめと今後の予定について述べる。

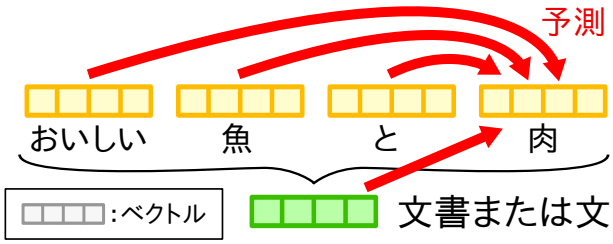

図 1: パラグラフベクトルの学習の概略

2 関連研究

多カテゴリにおけるレーティング予測に関する従来研究、及び、提案手法に関連した研究、技術について述べる。はじめに提案手法で用いているパラグラフベクトルと深層学習について述べる。次に、レーティング予測について、多カテゴリのものを対象とした従来手法とニューラルネットワークを用いた手法について述べる。最後に、その他本研究で利用した技術について述べる。

2.1 パラグラフベクトル

パラグラフベクトルは、文や文書といった大きな単位の言語表現の意味表現を学習する手法である。これは、Continuous BOW (CBOW) または Skip-gram[6] という単語の意味表現の学習手法を応用した手法である。ここでは CBOW を応用した Distributed Memory model of Paragraph Vectors (PV-DM) について説明する。PV-DM は BOW と異なり、単語の並び順を考慮した文や文書の分散表現を生成することができる。

以下に具体的なアルゴリズムを示す。ここでは文書の意味表現を学習する場合について考える。学習の概略を図 1 に示す。まず、意味表現を学習する対象となる文書に含まれる単語を初めから一つずつ読んでいく。その際、現在の単語及びその周辺の単語、現在の文書について、式 1 に示す目的関数 L を最大化するように各パラメータの学習を行う。

$$L = \sum_d L_d \tag{1}$$

$$L_d = \frac{1}{T}\sum_{t=k}^{T} \log p(w_t|w_{t-k},...,w_{t-1}),$$

$$p(w_t|w_{t-k},...,w_{t-1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}},$$

$$y = b + Uh(w_{t-k},...,w_{t-1},d;W,D)$$

ここで、L_d は文書 d の目的関数である。w_i は単語、W は全ての単語の分散表現を表す行列、D は全ての文書の分散表現を表す行列である。k はウィンドウサイズ、T は現在の文書に含まれる単語数である。ある単語の周辺を表す区間をウィンドウという。p は softmax 関数により正規化された、文脈から現在の単語が導かれることの尤度である。p を構成する y は現在の単語とウィンドウ内の単語及び現在の文書から導出される。$h(w_{t-k}, ..., w_{t-1}, d; W, D)$ は引数となるベクトルを平均したベクトルまたは結合したベクトルを返す関数である。

PV-DM によって得られたパラグラフベクトルはレーティング予測において BOW 等に比べ高い正答率を示すことが示されている。しかし、文書全体にパラグラフベクトルを用いる場合、文同士の位置関係が予測時に考慮できない。

2.2 深層学習

深層学習とは、多層のニューラルネットワークを用いた機械学習の手法の総称である [11]。以下には、その内ニューラルネットワークの正則化を行うための 2 つの手法、ドロップアウトと重み減衰について述べる。また、提案手法と直接関係がないが、関連するニューラルネットワークによるレーティング予測の手法 [3, 4, 5] で用いられているため、畳み込みニューラルネットワークについても説明する。

ドロップアウトとは、ニューラルネットワークにおける層のニューロンの数を一時的に減らすことによって正則化を行う方法である。ある層に対してドロップアウトを行うには、その層が持つニューロンの出力値を確率的に 0 とする。これを各学習回で行うことで、ニューラルネットワーク全体を学習する。

次に、重み減衰について説明する。重み減衰とは、ニューラルネットワークの各重みをその大きさに応じて学習回毎に減少させる正則化の手法である。重み減衰を行うためには、ニューラルネットワークの最小化すべき目的関数に対して各重みの 2 ノルムを足し合わせる。式 2 に、重み減衰を適用した目的関数 L' 示す。

$$L' = L + \frac{\lambda}{2} \sum_{\mathbf{w}} |\mathbf{w}|^2 \tag{2}$$

ここで、L はニューラルネットワークの重み減衰を適用していない目的関数である。$\mathbf{w}$ はニューラルネットワークの各層における重みである。λ は重みの減衰率である。式 2 により、ニューラルネット

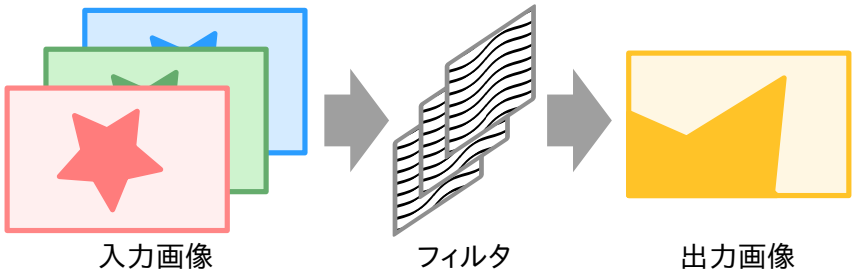

図 2: 畳み込み層の概略

ワークのある層の重み $\mathbf{w}$ に対する更新式は式 3 のようになる。

$$\mathbf{w} \leftarrow \mathbf{w} - \frac{\partial L}{\partial \mathbf{w}} - \lambda \mathbf{w} \tag{3}$$

ここで、$a \leftarrow b$ は変数 a の値をそのときの式 b の値で更新することを示す。ただし、全結合層のバイアス等では一般に重み減衰を行わない。これはそのような重みの値は場合によって大きい値を取る必要があるためである。

最後に、畳み込みニューラルネットワークについて説明する。畳み込みニューラルネットワークとは、畳み込み層とプーリング層を用いたニューラルネットワークである。一般に、畳み込み層とプーリング層は交互に配置される。畳み込みニューラルネットワークは入力の局所的な特徴を抽出することができる。また、畳み込みニューラルネットワークは元々画像認識に応用されていた手法であり、その入力は複数のチャネルを持つことがある。チャネルとは、画像でいう RGB の各色のことである。例として、畳み込みニューラルネットワークの入力を RGB 画像とした場合、それは（チャネル）×（画像の幅）×（画像の高さ）の 3 次元行列で表される。以下では、複数チャネルを持つ画像に対して畳み込みニューラルネットワークを適用する場合の畳み込み層とプーリング層について説明する。

畳み込み層とは、前の層の各ニューロンがそれと位置の近い次の層のニューロンとしか結合しないように全結合層を単純化した層である。具体的にはある $H \times H$ の大きさを持つフィルタを考え、それを畳み込み層の入力値に適用して値を出力する。図 2 にその概略を示す。 式 4 に、畳み込み層のフィルタ f によって処理された出力値 u_{fij} を示す。

$$u_{fij} = \sum_{c=0}^{C-1} \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} z_{c,i+p,j+q}^{(l-1)} h_{fcpq} + b_{fij} \tag{4}$$

ここで、$z_{c,i+p,j+q}^{(l-1)}$ は、前の層 $l-1$ の出力値である。H はフィルタの幅である。h_{fcpq} はチャネル c に対するフィルタ f の重みである。b_{fij} はフィルタ f のバイアスである。ただし、バイアスの値は

位置によらず一つのフィルタ内で共有することが多い。(i, j) は出力画像における位置である。(p, q) はフィルタ f における位置である。最終的な畳み込み層の出力値 z_{fij} は u_{fij} の値に活性化関数を適用し計算する。

$$z_{fij} = \sigma(u_{fij}) \tag{5}$$

ここで、σ は活性化関数である。実際の畳み込み層は複数のフィルタを持つことが多い。

次に、プーリング層とは入力画像の局所的な平均や最大値を取る層である。これによって、入力画像における特徴の位置に関するノイズを緩衝できる。式 6 に、平均プーリング層の出力値の位置 (i, j) における出力値 u_{cij} を示す。

$$u_{cij} = \frac{1}{H^2} \sum_{(p,q) \in P_{ij}} z_{cpq} \tag{6}$$

ここで、H はプーリングする範囲の幅である。P_{ij} は出力画像の位置によるプーリングの範囲であり、位置の集合で表される。c はチャネルである。(i, j) は出力画像における位置である。(p, q) はあるプーリングの範囲 P_{ij} における入力画像での位置である。

2.3 レーティング予測

2.3.1 隠れ状態を用いたホテルレビューのレーティング予測

藤谷ら [1] は複数のカテゴリにおけるレーティング予測に対して、Multi-Instance Multi-Label learning for Relation Extraction (MIML-RE) [7] モデルを用いた手法を提案している。その手法では、レビュー内の各文毎に予測した隠れレーティングからレビュー全体のレーティングを予測する。図 3 のように、文毎のレーティングからレビュー全体のレーティングを予測する際のカテゴリ間の繋がりを手動で変化させカテゴリ間の関係性を考慮している。各文の素性には Bag Of Words (BOW) または Bag Of n-grams を用いている。藤谷ら [1] は各文毎に隠れレーティングを予測することによって 0.4832 の正答率が得られることを示した。また、カテゴリ間の繋がりによって正答率が変化することを示した。

この手法では、文同士の位置関係を考慮しておらず、カテゴリ間については考慮しているものの複雑な関係を捉えることができていない。

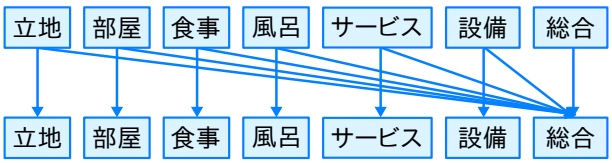

図 3: 藤谷ら [1] のモデルにおけるカテゴリ同士の繋ぎ方の例

2.3.2　ニューラルネットワークを用いたレーティング予測

ニューラルネットワークを用いたレーティング予測の手法が、Nal ら [3]、Rie ら [4]、Duyu ら [5] 等によって提案されている。これらの方法に共通するのは、単語の意味表現から畳み込みニューラルネットワークと全結合ニューラルネットワークを用いて分類を行うことである。まず、単語の意味表現から畳み込みニューラルネットワークを用いて単語同士の関係を捉えた特徴量を抽出する。その後、そこから得られた文書全体の特徴量を全結合ニューラルネットワークの入力とし多値または二値分類を行う。また、Duyu ら [5] と Nal ら [3] の手法はニューラルネットワークのモデルの中にパラメータとして単語の意味表現を取り込んでいる。これにより、特定の分類問題に対してそれらを最適化することができる。

これらの手法は 1 つのカテゴリにおける多値または二値分類を対象としている。よって、多カテゴリのレーティング予測において、これらの手法をカテゴリ毎に適用しただけではカテゴリ間の関係を考慮することができない。

2.4　形態素解析

形態素解析とは、文等の文字列を形態素に分割する処理である [10]。ここで、形態素とは意味を持った言語の最小単位である。日本語における形態素解析は、単語分割とその語形変化の解析に相当する。以降、日本語の文字列を形態素解析する場合の単語分割について述べる。単語分割の手法として、最長一致法と bi-gram マルコフモデルによるものについて説明する。

最長一致法では、解析すべき文字列を順に読み進めながら解析結果となる単語を決定していく。まず、文字列の現在の位置からの部分文字列が辞書内の単語と一致するか検査する。次に、一致した単語の内最も文字数の多いものを解析された単語として記録する。最後に、その単語の文字数だけ文字列を読み進め、再び現在の位置から始まる部分文字列に一致する単語を辞書で検索する。これを繰り返していくことによって、文字列の終端まで形態素解析を行う。

これに対して、bi-gram マルコフモデルによる手法では単語の bi-gram 確率を考え、この積が文字

列全体で最大となるように文字列を単語に分割する。以下に最も適切な単語列 $\hat{W}$ を求める式を示す。

$$\begin{aligned} \hat{W} &= \mathrm{argmax}_W P(W), \\ P(W) &= \sum_{i=1}^{n} P(w_i|w_{i-1}), \\ W &= w_1, w_2, ..., w_n \end{aligned} \tag{7}$$

ここで、W は元の文字列から分割された単語列であり、$P(W)$ はその単語列の同時確率である。w_i は単語、$P(w_i|w_{i-1})$ は単語 w_i と単語 w_{i-1} の bi-gram 確率である。

2.5 Adam

Adam[8] は確率的最適化のためのパラメータ更新のアルゴリズムである。実験によって、Adam がニューラルネットワークに適用できること、及び、パラメータを SGD や AdaGrad[9] より速く収束させることが確かめられている。式 8 に Adam による目的関数 L のパラメータ $\mathbf{w}$ の更新式を示す。式 8 は、実際には Diederik ら [8] によって示されている逐次的なアルゴリズムによって効率良く実装できる。

$$\begin{aligned} \mathbf{w}_t &= \mathbf{w}_{t-1} - \alpha \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t} + \epsilon}, \\ \mathbf{m}_t &= \frac{1-\beta_1}{1-{\beta_1}^t} \sum_{i=1}^{t} {\beta_1}^{t-i} \mathbf{g}_i, \\ \mathbf{v}_t &= \frac{1-\beta_2}{1-{\beta_2}^t} \sum_{i=1}^{t} {\beta_2}^{t-i} {\mathbf{g}_i}^2, \\ \mathbf{g}_i &= \frac{\partial L_i}{\partial \mathbf{w}_i} \end{aligned} \tag{8}$$

ここで、$\mathbf{w}_t$ は更新回数 t におけるパラメータ $\mathbf{w}$ を表す。$\mathbf{m}_t$ と $\mathbf{v}_t$ は更新回数 t におけるパラメータ $\mathbf{w}$ の一次・二次モーメントである。α と β_1、β_2、ϵ は Adam のパラメータである。$\mathbf{g}_t$ は更新回数 t におけるパラメータ $\mathbf{w}$ の勾配である。

Adam の特徴として、$\alpha \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}} \lessapprox \alpha$ である。$\alpha \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}$ は全ての t について g_t を定数倍しても変化しない。よって、更新幅の大まかな上限を実際に計算される勾配 g_t に依らず α のみによって決定することができる。また、Diederik ら [8] は $\frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}$ を signal-to-noise ratio (SNR) と呼んでいる。SNR の分子 $\mathbf{m}_t$ は、更新回数 t までの更新幅が小さい、あるいは、勾配の向きがよく変わっているとき小さくなる。よって、パラメータが最適解や局所解、鞍点付近にあるときは更新幅が小さくなり、それ以外

のとき大きくなる。

3　提案手法

提案手法は、文書・文間及びカテゴリ間の関係を考慮したレーティング予測の手法である。提案手法では、パラグラフベクトルによってレビューの文書及び各文の意味表現を生成し、それらをニューラルネットワークによる分類器入力としてレーティング予測を行う。以下にその基礎となる考えと具体的なアルゴリズムを示す。

3.1　文書・文間及びカテゴリ間の関係の考慮

先行研究 [1] の実験結果から、レビュー内の文毎の素性を元にレビューの分類を行うことが正答率の向上に有効であると考えられる。また、カテゴリ間の繋がりの変化が正答率に影響していることから、これをパラメータとして機械学習のモデルに組み込めば正答率を向上させることができると考えられる。さらに、レビュー内の文毎に意味表現を生成し分類器の入力とすれば、その位置関係を考慮した学習を行うことができる。

以下に、文の位置関係が重要となる例を示す。2 つ目の例は、1 つ目の例の 2 つ目の文と 3 つ目の文を入れ替えたものである。

> 食事が美味しかった。
> **とても良かった。**
> 部屋から眺めが素晴らしかった。

> 食事が美味しかった。
> 部屋から眺めが素晴らしかった。
> **とても良かった。**

1 つ目の例では、「とても良かった。」という文が直前の食事に関する文の意味を補完しているのに対し、2 つ目の例では、部屋に関する文の意味を補完している。このように、文の位置関係によってどの文がどの文と強く関連しているかが変化する。それによって、予測すべきレビュー全体のレーティングも変化すると考えられる。このように、文の位置関係の考慮はレーティング予測において重要であると考えられる。

次に、カテゴリ間の関係が重要となる例を図 4 に示す。図 4 のように「総合」カテゴリのレーティングは一般に他のカテゴリのレーティングに応じて高くなる。このような関係は「立地」と「部屋」、「食事」と「サービス」等の他のカテゴリ間にも存在する。よって、このことからも、カテゴリ間の関係をレーティング予測において考慮することで、正答率を向上させられると考えられる。

図 4: カテゴリ間の関係の例

しかし、個々のレビューに対して全ての文ベクトルをニューラルネットワークによる分類器の入力とすることは問題がある。なぜならば、文の数はレビュー毎に異なっており、複数レビュー内の複数の文ベクトルを単純な行列として表せないためである。これは、分類器のミニバッチ方式の訓練を難しくし、プログラムの実行時間を増加させる。この問題に対処するため、本手法では各レビュー内の文ベクトルに対して重み付け平均を行う。これにより、全てのレビューで文ベクトルの数が統一され、複数レビュー内の複数の文ベクトルをまとめて 3 次元行列として表すことができる。つまり、ミニバッチ方式の訓練における計算が容易となる。

分類器はニューラルネットワークを用いて構成することによって、文書・文間の関係とカテゴリ間の関係を同時に考慮した分類を行う。従来手法 [3, 4, 5] では、分類器として畳込みニューラルネットワークと全結合ニューラルネットワークが用いられている。しかし、提案手法においては畳み込みニューラルネットワークより全結合ニューラルネットワークを用いた方が正答率が高かったため、分類器には後者のみを用いた。

3.2 アルゴリズム

提案手法のレーティング予測の流れを説明する。図 5 にアルゴリズム全体の概略を示す。提案手法では、PV-DM によってレビュー内の文書全体及び各文の意味表現を生成し、それらをニューラルネットワークによる分類器の入力としてレーティング予測を行う。

初めに、PV-DM を用いて、各レビューの文書全体のベクトルとそれに含まれる各文のベクトルを生成する。以降、これらをそれぞれ文書ベクトル、文ベクトルと呼ぶ。文書ベクトルと文ベクトルについては別々に学習し生成する。式 1 の目的関数における h には引数のベクトルを結合する関数を用いる。また、学習の高速化のため、Quoc ら [2] によって用いられている階層的 softmax 関数の代替としてネガティブサンプリングを行う。ネガティブサンプリングとは、文脈外の単語をデータセットにおける出現確率でサンプリングし、それらと文脈の意味が遠ざかるように学習する手法である。

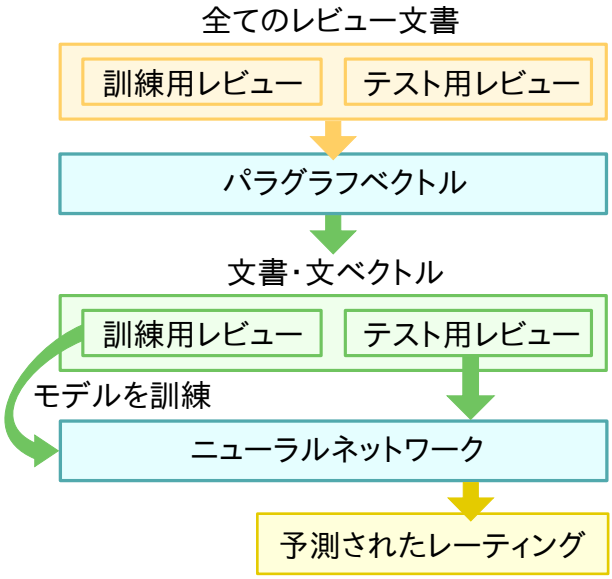

図 5: 提案手法におけるアルゴリズムの概略

ただし、出現頻度は極端に頻出する単語の影響を抑えるため各単語に対して 3/4 乗している。現在の単語と同じ単語や同一回の学習で一度サンプリングされた単語はサンプリングしない。

次に、各レビュー内の全ての文ベクトルに対して重み付け平均を行い、圧縮された文ベクトルを生成する。この過程により、各レビューで疎らだった文の数を統一する。式 9 に重み付け平均によって圧縮した文ベクトル $t_{i_{part}}$ を示す。各文ベクトルは圧縮後の各文ベクトルと位置が近いほど重みが大きくなるように重み付け平均する。

$$\mathbf{t}_{i_{part}} = \sum_{i_{sent}} \frac{w(x_{i_{part}}(i_{sent}))}{|\sum_{i'_{sent}} w(x_{i_{part}}(i'_{sent}))|}\mathbf{s}_{i_{sent}}, \tag{9}$$

$$x_{i_{part}}(i_{sent}) = \frac{i_{sent}}{\#sent - 1} - \frac{i_{part}}{\#part - 1},$$

$$w(x) = \begin{cases} \frac{1}{2}(\cos(\pi|x|) + 1) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

ここで、$\mathbf{s}_{i_{sent}}$ はレビュー内の文ベクトル、$\mathbf{t}_{i_{part}}$ は重み付け平均された文ベクトルである。i_{sent} はレビュー内の文ベクトルのインデックス、i_{part} は重み付け平均された文ベクトルのインデックスである。$\#sent$ はレビュー内の文ベクトルの数、$\#part$ は重み付け平均された文ベクトルの数である。[*1]

最後に、分類器によってレーティング予測を行う。分類器は全結合ニューラルネットワークによって構成される。図 6 に各層の結合の様子を示す。 入力層はレビュー毎の文書ベクトルと圧縮された文

[*1] 重み付けの関数には cos 関数の他に、x に対して線形に重みを減少させるような関数や、単純に文を区画毎に平均するような関数も考えられる。区画毎に平均する関数は他の 2 つより正答率が低く、線形な関数と cos 関数はほぼ同じ正答率を示したため、cos 関数を採用した。

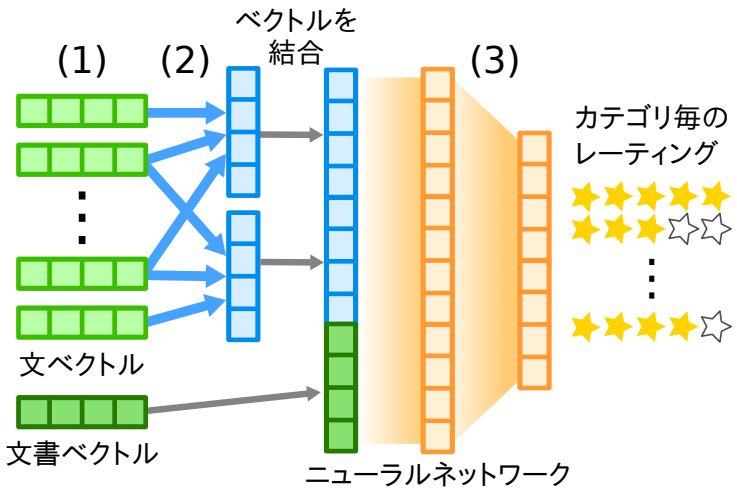

図 6: 全結合ニューラルネットワークによる分類器

ベクトルの結合ベクトルである。ニューラルネットワークの活性化関数には、シグモイド関数を用いる。また、出力層はカテゴリの数とレーティングの場合の数の積だけのニューロンを持ち、各ニューロンの出力はあるカテゴリがあるレーティングであることの正規化されていない対数確率を表す。各ニューロンの出力はカテゴリ毎に交差エントロピー誤差関数によって損失に変換される。ニューラルネットワークは式 10 に示す目的関数 E を最小化するように学習を行う。

$$E = -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K} d_{nck} \log y_{ck}(x_n; w), \tag{10}$$

$$y_{ck}(x_n; w) = \frac{e^{u_{ck}(x_n;w)}}{\sum_{j=0}^{K} e^{u_{cj}(x_n;w)}}$$

ここで、u_{ck} は出力層のニューロンの出力値、y_{ck} はカテゴリ c においてクラス k が選ばれる確率、w はニューラルネットワークのパラメータである。d_{nck} は n 番目の文書がカテゴリ c でクラス k ならば 1、それ以外で 0 となる値である。N はミニバッチサイズ、C はカテゴリの総数、K はクラスの総数である。ニューラルネットワークによる多値分類では一般に、出力層のニューロンの出力値全体を softmax 関数で正規化する。しかし、複数カテゴリの多値分類問題において、同様に出力層のニューロンの出力値を正規化してしまうと softmax 関数の値が確率としての意味を成さない。よって、式 10 では、カテゴリ毎に softmax 関数でニューロンの出力値が示す対数確率を正規化する。そして、それらを各カテゴリ毎の目的関数とし足し合わせることであるレビューに対する目的関数を構成する。これを全てのレビューに対して足し合わせ、ニューラルネットワークの目的関数とする。さらに、ニューラルネットワークの学習時にはドロップアウト及び重み減衰を行う。ただし、重み減衰において全結合層の重みの内、バイアスは減衰していない。パラメータ最適化の手法には Adam[8]

を用いる。

4　実験

実験の目的は、提案手法のレーティング予測における優位性と、提案手法に反映されている意図が実際にレーティング予測において有効であるかの検証である。具体的には、提案手法が従来手法 [1] に対して優れているかを検証する。また、提案手法を単純にした複数の比較手法を用意し、提案手法がそれらより有意に優れていることを示す。さらに、提案手法と比較手法及び比較手法同士の比較により、文同士の位置関係の考慮や文書ベクトルと文ベクトルを同時に用いることがレーティング予測に有効であるかを検証する。

4.1　実験設定

実験には、正答率を測定する実験と提案手法における予測レーティングと正解レーティングの RMSE を測定する実験の 2 つを行った。比較手法として、提案手法における分類器の入力をそれぞれ (1) DV (Document Vector)、(2) ASV (Averaged Sentence Vector)、(3) Weighted ASV に変えた手法を用意した。DV とはレビュー全体の文書ベクトルであり、Quoc ら [2] の手法に相当する。ASV とはレビュー内で平均した文ベクトルであり、Weighted ASV とはレビュー内で重み付け平均によって圧縮された文ベクトルである。Weighted ASV において重み付け平均の方法は提案手法と同様であるが、重み付け平均の式のハイパーパラメータは別に調整した。有意差検定にはマクネマー検定を用い、p 値が 0.05 より小さいとき有意とした。また、RMSE の計算において正解または予測レーティングが 0 点であるものは評価から省いた。これは用いるデータセットのレビューにおいて、レーティングの 0 点はレーティングが不可能であることを意味するためである。

データセットとしては、先行研究 [1] と同様に、ホテル予約サイト楽天トラベルにおけるレビュー 337,266 件からレビューの番号順に訓練データ 300,000 件、開発データ 10,000 件、評価データ 10,000 件を用いた。楽天トラベルによって提供されている元のレビューデータはレーティングを含むファイルとレビューの文書を含むファイルとに分かれている。それぞれ Tab Separated Values (TSV) フォーマットで 1 行 1 レビューとして情報が記述されている。レーティングを含むファイルと文書を含むファイルのフォーマットをそれぞれ図 7 と図 8 に示す。以下に、それぞれのファイルから実験のために抽出した情報とそれらの前処理について記述する。まず、レーティングを含むファイルからは、レビューの ID と「立地」から「総合」カテゴリまでのレーティングを取り出した。レビューの

第 1 フィールド（レビューの ID）[Tab] 第 2 フィールド [Tab] ...
第 7 フィールド（「立地」カテゴリのレーティング）[Tab] ...
第 13 フィールド（「総合」カテゴリのレーティング）
図 7: レーティングを含む TSV ファイルのフォーマット

第 1 フィールド [Tab] 第 2 フィールド（レビューの文書）[Tab]
第 3 フィールド（レビューの ID）[Tab] ...
図 8: 文書を含む TSV ファイルのフォーマット

文書を含むファイルからは、レビューの ID とレビューの文書を取り出した。ただし、元のレビューの文書に含まれる「【ご利用の宿泊プラン】」以降の文字列はユーザが記述したものではないため取り除いた。その後、レーティングとレビューの文書をレビュー ID が一致するように組にした。このとき、レーティングを含むファイル、または、レビューの文書を含むファイルのどちらかにしか存在しないレビュー ID を持つレビューは取り除いた。

レビューの文書に対する前処理について以下に示す。まず、全てのレビューの文書に対して文字コードを UTF-8 に変換し、以下の正規化処理を行った。記号「！”＃＄％＆’（）＊＋, − . ／：；＜＞？＠［￥］ ＾＿｀｛｜｝\u301C」は全て NFKC 形式で正規化した。記号「\u02D7\u058A\u2011\u2012\u2013 \u2043\u207B\u208B\u2212」は全て記号「-」で置き換えた。記号「\uFE63 －‐−―—﹘━ー」は全て記号「ー」で置き換えた。チルダ記号「\u007E\u223C\u223E\u301C \u3030\uFF5E」は全て削除した。ここで、\uXXXX は 16 進数で表現された Unicode のコードポイントを示す。次に、各レビューの文書を文に分割した。「。」、「.」、「!」、「?」を文の終端文字とし、文の終端文字でない文字の 1 回より多い繰り返しと文の終端文字または文書の終端の連続を一つの文として正規表現によって解析した。ただし、文が 1 つも解析できなかった文書については、文書全体の文字列をその文書に含まれる唯一の文とした。最後に、形態素解析には形態素解析器 MeCab を用いた、辞書には IPA 辞書を用いた。単語の情報は表層のみを利用し、MeCab によって出力される表層が無い特殊な単語は取り除いた。

表 1 に各手法におけるニューラルネットワークのパラメータ設定を示す。全ての手法において、中間層の数は 1、入力層及び中間層におけるドロップアウト率はそれぞれ 0.2 と 0.5 で共通である。Adam[8] のハイパーパラメータは [8] と同様の値を用いた。Weighted ASV と提案手法において圧縮された文ベクトルの数はそれぞれ 3 つと 2 つとした。全ての実験において文書及び文ベクトルについては、学習回数は 1,024 回、学習する単語の範囲は前 3 単語、単語の最少出現回数は 5 回、ネガティ

ブサンプリングの回数は 5 回、ベクトルの次元数は 600 次元に設定し学習したものを用いた。

4.2 結果

まず、提案手法と 3 つの比較手法、従来手法 [1] を正答率で比較したものを表 2 に示す。そのグラフを図 9 に示す。また、表 3 に提案手法と従来手法におけるカテゴリ別の正答率を示す。そのグラフを図 3 に示す。表 2 において、提案手法が従来手法の正答率を 0.0198 有意に上回っている。また、提案手法が DV と Weighted ASV の正答率をそれぞれ 0.0050 と 0.0163 有意に上回っている。Weighted ASV が ASV を 0.0029 有意に上回っている。

次に、表 4 にレーティングの RMSE を測定した結果を示す。そのグラフを図 4 に示す。提案手法は従来手法が欠点としていた食事と風呂のカテゴリにおいてそれぞれ 0.65 及び 0.34 だけ低い誤差を示した。また、その他全てのカテゴリにおいても提案手法は従来手法より低い誤差を示した。

最後に、正答率を測定する実験における、提案手法の精度と再現率、F 値をそれぞれ表 5 と表 6、表 7 に示す。表において N/A は計算不可能であることを示す。また、カテゴリ毎の正解レーティングの内訳と提案手法のカテゴリ毎の予測レーティングの内訳を表 8 と表 9 に示す。

表 1: 各手法のパラメータ設定

手法	学習回数	中間層でのユニット数
DV	20	512
ASV	55	256
Weighted ASV	24	256
提案手法	30	512

表 2: 各手法における正答率

手法	正答率
従来手法 [1]	0.4832
DV	0.4980
ASV	0.4838
Weighted ASV	0.4867
提案手法	**0.5030**

表 3: 提案手法と従来手法 [1] におけるカテゴリ別の正答率

手法	立地	部屋	食事	風呂	サービス	設備	総合
従来手法 [1]	0.4961	0.4706	0.5140	0.3973	0.4783	0.4265	0.5660
提案手法	0.5140	0.4984	0.5353	0.4347	0.5116	0.4479	0.5794

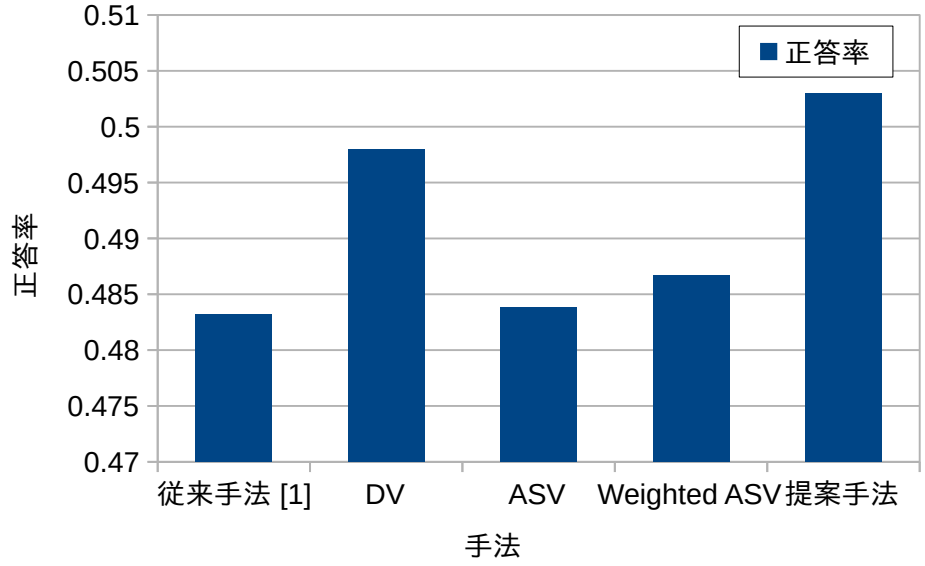

図 9: 各手法における正答率

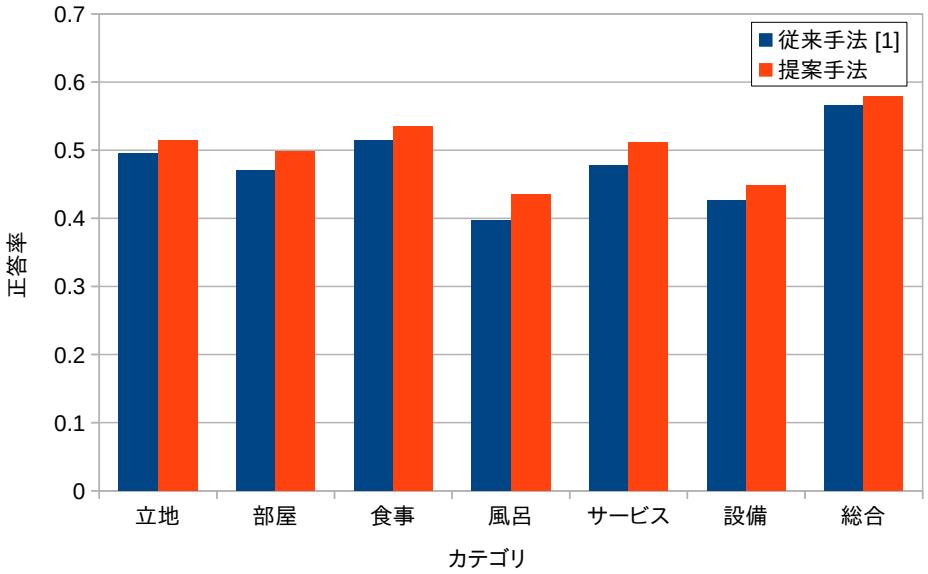

図 10: 提案手法と従来手法 [1] におけるカテゴリ別の正答率

表 4: 提案手法と従来手法 [1] におけるカテゴリ別のレーティングの RMSE

手法	立地	部屋	食事	風呂	サービス	設備	総合
従来手法 [1]	0.97	0.97	1.53	1.27	0.94	0.95	0.81
提案手法	0.88	0.88	0.93	1.03	0.86	0.90	0.73

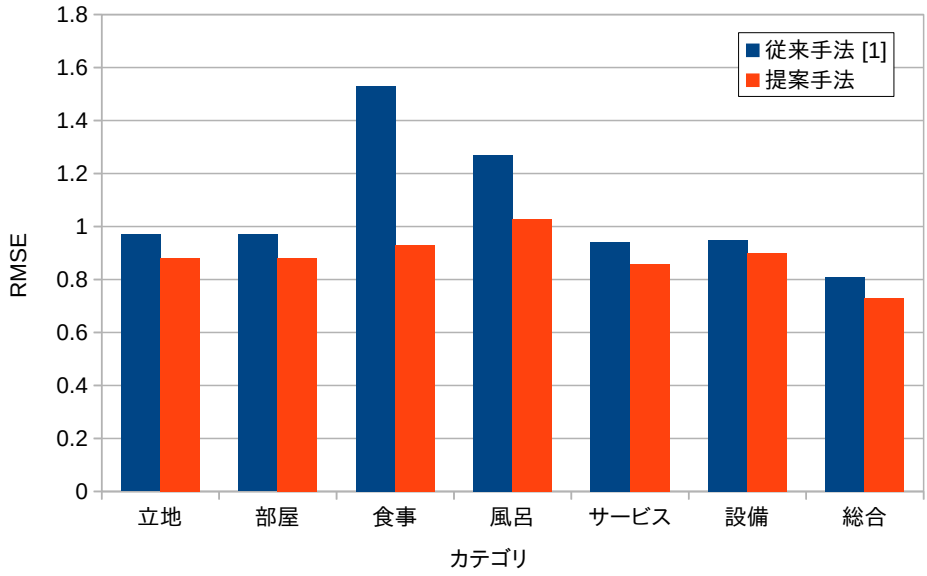

図 11: 提案手法と従来手法 [1] におけるカテゴリ別のレーティングの RMSE

表 5: 提案手法の精度

レーティング	立地	部屋	食事	風呂	サービス	設備	総合	全カテゴリ
0	N/A	N/A	0.5474	N/A	N/A	N/A	N/A	0.5474
1	N/A	0.6029	0.4182	0.5000	0.5520	0.5122	0.5570	0.5330
2	0.1111	0.3372	0.3387	0.3092	0.3830	0.2785	0.2685	0.3076
3	0.4507	0.4424	0.2874	0.4133	0.4454	0.4353	0.4186	0.4251
4	0.4197	0.4371	0.3928	0.3917	0.4481	0.4194	0.5493	0.4512
5	0.5811	0.6135	0.6334	0.5167	0.6177	0.5459	0.6820	0.5994

表 6: 提案手法の再現率

レーティング	立地	部屋	食事	風呂	サービス	設備	総合	全カテゴリ
0	N/A	N/A	0.9278	0.0000	N/A	0.0000	N/A	0.8011
1	0.0000	0.2470	0.1825	0.1267	0.3855	0.1207	0.3308	0.2186
2	0.0038	0.0621	0.0598	0.0740	0.0577	0.0407	0.1086	0.0589
3	0.0472	0.2165	0.0382	0.6736	0.3724	0.3933	0.1573	0.3359
4	0.4674	0.6485	0.3003	0.3128	0.5405	0.6204	0.7765	0.5520
5	0.7147	0.5582	0.5467	0.5053	0.6056	0.3853	0.5125	0.5613

表 7: 提案手法の F 値

レーティング	立地	部屋	食事	風呂	サービス	設備	総合	全カテゴリ
0	N/A	N/A	0.6886	N/A	N/A	N/A	N/A	0.6504
1	N/A	0.3504	0.2541	0.2022	0.4539	0.1953	0.4151	0.3100
2	0.0074	0.1049	0.1017	0.1194	0.1003	0.0710	0.1547	0.0989
3	0.0854	0.2908	0.0674	0.5123	0.4056	0.4132	0.2287	0.3753
4	0.4423	0.5222	0.3404	0.3479	0.4900	0.5004	0.6434	0.4966
5	0.6410	0.5846	0.5869	0.5109	0.6116	0.4518	0.5852	0.5798

表 8: カテゴリ毎の正解レーティング件数

レーティング	立地	部屋	食事	風呂	サービス	設備	総合	全カテゴリ
0	0	0	3365	468	0	64	0	3897
1	35	166	126	221	179	174	133	1034
2	260	467	351	635	312	541	267	2833
3	1356	1898	1258	2852	2124	2718	1144	13350
4	3607	3698	2191	3075	3597	3664	4573	24405
5	4742	3771	2709	2749	3788	2839	3883	24481

表 9: 提案手法のカテゴリ毎の予測レーティング件数

レーティング	立地	部屋	食事	風呂	サービス	設備	総合	全カテゴリ
0	0	0	5703	0	0	0	0	5703
1	0	68	55	56	125	41	79	424
2	9	86	62	152	47	79	108	543
3	142	929	167	4648	1776	2456	430	10548
4	4017	5486	1675	2456	4338	5420	6465	29857
5	5832	3431	2338	2688	3714	2004	2918	22925

5　考察

まず、提案手法と従来手法を正答率及び正解レーティングと予測レーティングの RMSE について比較する。表 2 より、提案手法が従来手法 [1] の正答率を 0.0198 有意に上回っている。よって、提案手法が従来手法 [1] より正答率において優れていることが分かった。さらに、図 3 においても、提案手法の正答率は従来手法 [1] を全てのカテゴリにおいて上回っている。レーティングの RMSE については、図 11 より、提案手法の RMSE が従来手法 [1] のそれを全てのカテゴリにおいて下回っている。よって、レーティングの RMSE においても提案手法が従来手法 [1] より優れていることが分かった。ここで、実験に用いたレビューデータの特徴から提案手法と従来手法 [1] の違いについて考察する。藤谷ら [1] より、実験で用いたデータセットには、「食事」のカテゴリで 0 点が付与されたレビューが 108,079 件存在する。さらに、0 点が付与されたレビューは「風呂」のカテゴリでは 13,332 件、設備のカテゴリでは 2,011 件、他のカテゴリでは 0 件存在する。一般に、0 点はユーザが何らかの理由でレーティング不可能と判断したことを示す。例えば、「食事」のカテゴリならばホテルで食事を取っていない、「風呂」のカテゴリならば別の入浴施設を利用した等である。よって、提案手法は従来手法 [1] よりレビュー中の上記のような意味をよく捉えていると考えられる。また、表 8 と表 9 を比較すると、正解レーティングが 0 点であるレビューが 0 件であるカテゴリにおいて、提案手法は 0 点を 1 件も予測していない。このことからも、提案手法は、0 点が 1 から 5 点のレーティングと異なりレーティングの度合いを示すものではないということを上手く考慮できているといえる。

次に、文の位置関係の考慮がレーティング予測に正答率の向上に有効であるかを検証する。表 2 より、Weighted ASV の正答率が ASV の正答率を 0.0029 有意に上回っている。ここで、ASV はレビュー内の文ベクトルを足し合わせただけの素性を分類器の入力としている。すなわち、文の特徴は考慮しているが、その位置関係は考慮していない。また、Weighted ASV は文ベクトルを位置によって重み付け平均したベクトルを素性としている。すなわち、文の特徴とその大まかな位置関係を考慮している。以上より、文の位置関係を考慮することはレーティング予測に有効であることが分かった。

次に、文書ベクトルと文ベクトルを同時に素性として用いることが正答率の向上に有効であるかを検証する。表 2 より、提案手法が DV や Weighted ASV に比べ有意に高い正答率を示している。提案手法と DV の手法における差は重み付け平均された文ベクトルを素性として用いるか用いないかである。提案手法と Weighted ASV の差は文書ベクトルを素性として用いるか用いないかである。以

上より、文書ベクトルと文ベクトルを同時に特徴量として用いることがレーティング予測に有効であることが分かった。また、このことは文書ベクトルと文ベクトルがいくらか異なる特徴を学習していることを示す。なぜならば、文書ベクトルと文ベクトルが同じ特徴を学習していた場合、提案手法とDV の正答率における有意差は無くなるはずだからである。ただし、このとき提案手法と Weighted ASV の正答率における有意差は一般に無くならない。なぜならば、文ベクトルを分類器において評価する方法として重み付け平均が最適とは限らないからである。

次に、提案手法のその他の性質について考察する。表 8 より、「食事」カテゴリでレーティングが 0 点である場合を除いて、正解レーティングが 0 から 2 点であるレビューの件数は 3 から 5 点であるものに比べ少ないことが分かった。この正解レーティングが 0 から 2 点であるレビューについて、表 6 より、「食事」カテゴリでレーティングが 0 点である場合と精度または再現率が計算できなかった場合を除いて、提案手法の再現率は全て精度に比べて低い結果となった。このため、表 7 のように、F 値も 0 から 2 点のレーティングにおいて低くなってしまった。特に、表 9 より、「立地」カテゴリでレーティングが 1 点の場合と「風呂」カテゴリでレーティングが 0 点の場合、「設備」カテゴリでレーティングが 0 点の場合は全て予測レビュー件数が 0 件となった。しかし、表 8 より、これらのカテゴリとレーティングの組み合わせを持つレビューは評価データに含まれている。そのため、表 6 のように、そのようなカテゴリとレーティングの組み合わせでは再現率が 0 となってしまった。以上のように、提案手法は 0 から 2 点のレーティングにおいて、再現率、及び、F 値が低くなってしまうことが分かった。

最後に、提案手法そのものの問題点について考察する。提案手法の問題点の一つは、レビューの素性の生成と分類のモデルが分離していることである。具体的には、パラグラフベクトルによってレビューの文書とそれが含む文の素性を生成する段階、及び、それらをニューラルネットワークによって分類を行う段階の 2 つに分離している。このことは、問題を 2 つに分けることで個々の問題を単純にしているが、同時にいくつかの問題を伴う。1 つは、レビューを一つずつレーティング予測することができないことである。提案手法では、新しいレビューを訓練データに加える場合、全てのレビューの文書・文ベクトルを再構築する必要がある。レビューの件数が多い場合、これは大量の計算を必要とし効率的でない。これは提案手法が実際に応用されるときに問題となる。なぜなら、実際の商品レビューの件数は時間と共に増加していくためである。2 つ目は、文書や文の素性やそれらを生成するモデルのパラメータが分類時に調整できないことである。最大の正答率を達成するためには、

これらは分類器のパラメータと同様に対象とする分類問題に対して最適化されることが望ましい。3つ目は、単語等のより細かい言語要素間の関係が分類時に十分に考慮できないことである。単語間の関係は文書・文ベクトルによってある程度表現されている。しかし、それらは分類の正答率が最大になるように表現されているとは限らない。以上の問題点から、提案手法の素性の生成と分類のモデルは統合するべきであると分かる。なぜなら、モデルを統合すれば1つ目と2つ目の問題点が解消されることは明らかなためである。3つ目の問題点も、レビューの文書中における単語や文字等の細かい言語要素の特徴からレーティングを予測することで対処することができる。統合されたモデルには、入力間出力間の複雑な関係を考慮できるニューラルネットワークのようなモデルを引き続き用いるのがよいと考えられる。

6　結論

6.1　まとめ

本研究では、多カテゴリにおけるレーティング予測について、レビュー全体の文書ベクトルに加え重み付け平均された文ベクトルを用いた手法を提案した。

楽天トラベルのレビューデータを用いた実験では、提案手法が従来手法 [1] より高い正答率を示した。さらに、カテゴリ別の正答率、カテゴリ別のレーティングの RMSE を用いた評価基準においても、提案手法は従来手法より優れていることが分かった。文ベクトルのみを用いた 2 つの比較手法の実験結果より、レビューの文書内の文の並びがレーティング予測に重要であることが分かった。提案手法及び 2 つの比較手法の実験結果より、文書ベクトルと文ベクトルを同時に用いることが正答率の向上に有効であり、また、文書ベクトルと文ベクトルがレビューのいくらか異なる特徴を捉えていることが分かった。また、実験において使用したデータセットに対する提案手法の性質として、0 点から 2 点のレーティングにおいて再現率、及び、F 値が低くなってしまうことが分かった。

6.2　今後の予定

今後は、単語や文字等のより小さな言語要素間の複雑な関係を考慮することを課題とする。考察より、このためには各レビューの素性を生成するモデルと分類を行うモデルを 1 つに統合する必要がある。なぜならば、モデルが分かれていることによって小さな言語要素同士の関係を予測時に考慮できないためである。統合されたモデルには、単語や文字間の複雑な関係を考慮するためニューラルネットワークを用いる。また、これによって、訓練レビューを一つずつ学習できるようになり、提案手法のように新しい訓練レビューが追加される毎に訓練レビュー全てについて素性を再計算する必要が無くなる。さらに、単語・文字等の小さな言語要素の特徴から文書・文等のより大きな言語要素の特徴をニューラルネットワークによって自動的に構成すれば、文書・文間の関係も同時に考慮できると考えられる。修士研究では、これらを達成することによってレーティング予測の正答率をさらに向上させることを目指す。

謝辞

本研究を進めるにあたって、佐々木裕教授、三輪誠准教授に御指導いただきました。同研究室の先輩方、友人には多くの協力、助言をいただきました。また、本研究の実験において、楽天株式会社よりホテル予約サイト楽天トラベルにおけるレビューデータを使用させていただきました。この場を借りて感謝致します。

参考文献

[1] 藤谷宣典ら, 隠れ状態を用いたホテルレビューのレーティング予測. 言語処理学会第 21 回年次大会, 2015.

[2] Quoc Le et al., Distributed Representations of Sentences and Documents. ICML 2014, 2014.

[3] Nal Kalchbrenner et al., A Convolutional Neural Network for Modelling Sentences. ACL 2014, 2014.

[4] Rie Johnson et al., Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. NAACL 2015, 2015.

[5] Duyu Tang et al., Learning Semantic Representation of Users and Products for Document Level Sentiment Classification. ACL 2015, 2015.

[6] Yoshua Bengio et al., A Neural Probabilistic Language Model. The Journal of Machine Learning Research 3, 2003.

[7] Mihai Surdeanu et al., Multi-instance Multi-label Learning for Relation Extraction. CoNLL 2012, 2012.

[8] Diederik Kingma et al., Adam: A Method for Stochastic Optimization. ICLR 2015, 2015.

[9] John Duchi et al., Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Journal of Machine Learning Research 12, 2011.

[10] 田中穂積, 自然言語処理 -基礎と応用-. 電子情報通信学会.

[11] 岡谷貴之, 深層学習. 講談社.