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

1 序論 3

2 関連研究 4

2.1 隠れ状態を用いたホテルレビューのレーティング予測 4

2.2 パラグラフベクトル 4

2.3 ニューラルネットワークを用いた評判分析 5

3 提案手法 7

3.1 文書・文間及びカテゴリ間の関係の考慮 7

3.2 アルゴリズム 8

4 実験 11

4.1 実験設定 11

4.2 結果 12

5 考察 13

6 結論 14

表目次

1 各手法のパラメータ設定 11

2 各手法における正答率 12

3 提案手法と従来手法 [1] におけるレーティングの RMSE 12

図目次

1 藤谷ら [1] のモデルにおけるカテゴリ同士の繋ぎ方の例 4

2 パラグラフベクトルの学習の概略 5

3 カテゴリ間の関係の例 8

4 提案手法におけるアルゴリズムの概略 8

5　全結合ニューラルネットワークによる分類器 10

1　序論

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、その評判分類において、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深い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性を捉えることができる。さらに、多カテゴリの分類問題においてはカテゴリ間の関係性を捉えた分類が実現できる。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。これは評判分類問題に対し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。

本研究は、複数カテゴリにおける評判分類について、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分類の実現を目的とする。

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する。次に、その文書ベクトルと重み付け平均された文ベクトルを用いてニューラルネットワークによる分類器において分類しレーティング予測を行う。楽天トラベルのデータセットを用いた実験において、提案手法は従来手法 [1] に対して約 2pp 上回る正答率を示した。レーティングの平均二乗誤差 (RMSE) を元にした評価基準では、従来手法 [1] において欠点となっていたカテゴリについてそれを上回る結果を示した。

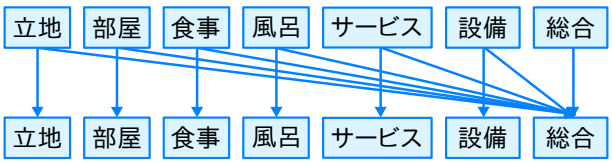

図 1: 藤谷ら [1] のモデルにおけるカテゴリ同士の繋ぎ方の例

2　関連研究

2.1　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 1 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。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が示された。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も示されている。

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。

2.2　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 2 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

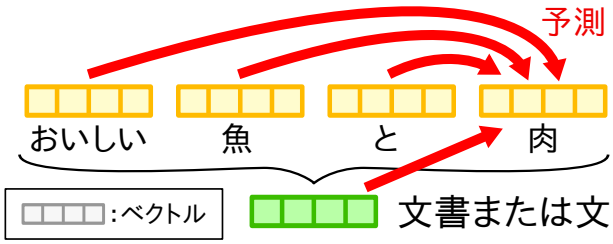

図 2: パラグラフベクトルの学習の概略

L を最大化するように各パラメータの学習を行う。

$$L = \frac{1}{T}\sum_{t=k}^{T} \log p(w_t|w_{t-k}, ..., w_{t-1}), \tag{1}$$
$$p(w_t|w_{t-k}, ..., w_{t-1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}},$$
$$y = b + Uh(w_{t-k}, ..., w_{t-1}, d; W, D)$$

ここで、d は文書、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。

2.3 ニューラルネットワークを用いた評判分析

ニューラルネットワークを用いた評判分析の手法が、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。

3　提案手法

提案手法では、パラグラフベクトルによってレビュー内の各文及び文章の意味表現を生成し、それらをニューラルネットワークの入力として分類を行う。以下にその基礎となるアイデア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ることで、その順序を考慮した学習を行う。

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。よって、文の位置関係を考慮することは重要である。

次に、カテゴリ間の関係が重要となる例を図 3 に示す。図 3 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。

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

図 3: カテゴリ間の関係の例

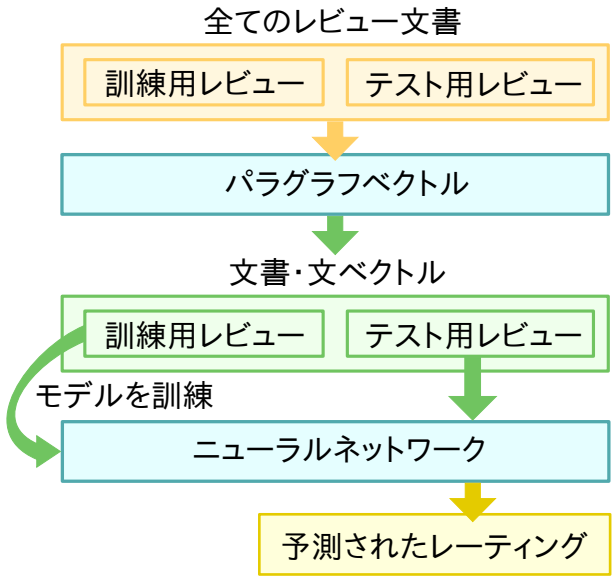

図 4: 提案手法におけるアルゴリズムの概略

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捉えた分類を行う。従来手法 [3] や [4]、[5] では、単語ベクトルに対して畳込み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

提案手法の処理の流れを説明する。図 4 にアルゴリズム全体の概略を示す。

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。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 2 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。

$$\mathbf{t}_{i_{part}} = \sum_{i_{sent}} \frac{w(x_{i_{part}}(i_{sent}))}{|\sum_{i'_{sent}} w(x_{i_{part}}(i'_{sent}))|} \mathbf{s}_{i_{sent}}, \tag{2}$$

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

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

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 5 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内であるレーティング値が選ばれることの正規化されていない対数

[*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。

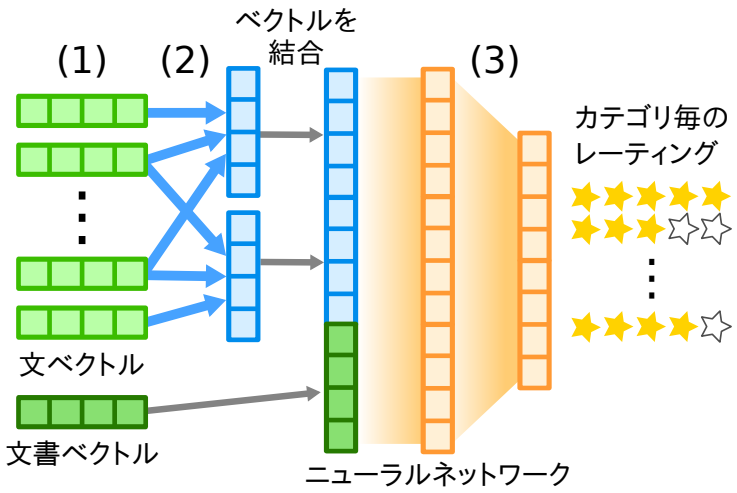

図 5: 全結合ニューラルネットワークによる分類器

確率を表す。ニューラルネットワークは式 3 に示す目的関数 E を最小化するように学習を行う。

$$E = -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K} d_{nck} \log y_{ck}(x_n; w), \tag{3}$$

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

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。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。ところで、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値である対数確率全てを 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 3 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。パラメータ最適化の手法には Adam[8] を用いる。

4　実験

4.1　実験設定

実験は、各手法の正答率を測定するものと、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平均二乗誤差 (RMSE) を測定するものの 2 つを行った。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。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 に変えた手法を用いた。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。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。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。

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。

レビューに対する前処理について以下に示す。まず、レビューの文書に対して、文字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

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

ダ記号「\u007E\u223C\u223E\u301C \u3030\uFF5E」は全て削除した。次に、文の解析は正規表現を用いて行った。「。」、「.」、「!」、「?」を文の終端文字とし、解析した最後の文の次の文字から文の終端文字または文書の終端までを一つの文として解析した。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、表層が無いものは取り除いた。

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。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。

次に、表 3 にレーティングの RMSE を測定した結果を示す。提案手法は従来手法 [1] が弱点としていた食事と風呂のカテゴリにおいてそれぞれ 0.60 及び 0.24 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。

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] におけるレーティングの RMSE

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

5　考察

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ことから、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。また、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とから、文の位置関係の考慮がレーティング予測に有効であることが分かった。さらに、提案手法が Weighted ASV に比べ有意に高い正答率を示していることから、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。これ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。

藤谷ら [1] より、実験で用いたデータセットには、食事のカテゴリにおいて 0 点が付与されたレビューが 108,079 件存在する。また、風呂のカテゴリでは 13,332 件、設備のカテゴリでは 2,011 件存在し、他のカテゴリでは 0 件である。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。

次に、提案手法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、PV-DM 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 つ目は、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はない。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表現されているが、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。以上より、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。

6　結論

本研究では、多カテゴリにおける評判分類問題について、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。

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。また、比較手法の結果より、レビュー内の文の並びが評判分類に重要であること、及び、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。

今後の課題は単語間や文字間などの言語要素間のより多様で複雑な関係を考慮することである。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。モデルの統合によって、学習手法の柔軟性を高めると共にさらなる正答率の向上を目指す。

謝辞

本研究において，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.