日本オペレーションズ・リサーチ学会
2019年 秋季研究発表会

レビューデータを用いた商品評価に関する特徴的な表現の抽出

入会申請中　中央大学大学院　三宅　伸　MIYAKE Shin
05000907　東海大学　大竹　恒平　OTAKE Kohei
01405390　中央大学　生田目　崇　NAMATAME Takashi

1. はじめに

現在広く使われている EC サイトにおいては購入者によるロコミ（商品レビュー）が購買に大きく影響しており，サービスや商品に対する評価や，レビューそのものに対する評価のマーケティング活動への活用が期待されている．本研究ではゴルフ場の予約サイトを対象に，ゴルフ場へのレビュー及び評価得点を利用し，商品評価に関する特徴的な表現を抽出することを試みる．

具体的には，畳み込みニューラルネットワーク（CNN）によりレビューの特徴的な表現を学習し，投稿者のこれまでの評価得点よりも高い評価を有するレビューを判別する．また，モデルに入力するデータの形式を変化させ，精度の違いを検証する．

2. 先行研究および本研究の位置付け

近年，深層学習を用いたモデルはコンピュータビジョンにおいて驚くべき成果をあげているが，消費者行動・消費者理解や広告戦略等，マーケティング領域においても，活用が期待されている．このような状況の下，口コミや商品レビューといった，消費者が投稿したテキストデータを対象とした取組みも行われておりカテゴリ分類などの分類問題 [1]，感情分析やスパム検出，およびその他の従来の自然言語処理タスクにおいて，深層学習を用いたモデルが有効であることが示されている．

本研究では，レビューの特徴量行列からのパターン認識を行い，高い評価（本研究ではこれを「高評価レビュー」と呼ぶ）を付けた商品レビューに含まれる，特徴的な表現を抽出することを目的とする．

3. 対象レビューの抽出

本研究では，ゴルフポータルサイトを運営している企業より提供された，ゴルフ場に関するレビューデータを用いて分析を行う．対象とするレビューは，あるゴルフポータルサイト内に掲載されているゴルフ場に対して投稿されたものである．なお，評価は実施にゴルフ場を予約し，プレーした会員のみが行うことができる．会員は五点満点でゴルフ場に対する評価を行う．本研究では，投稿者個人の過去の評価平均点よりも高い評価が付くレビューを高評価レビューととらえ，ラベル付けを行う．なお，分析の前処理として，レビュー文が空白のものは削除した．また，対象とする投稿者は，期間内に 5 回以上 50 回以下の投稿を行った者に限定する．なお，期間は期間：2013 年 11 月 1 日 ～ 2016 年 11 月 30 日である．

4. 分析方法

4.1 word2vec

word2vec の構造は，隠れ層と出力層の 2 つの層から構成される単純なニューラルネットワークであり，用いることで単語の分散表現を獲得することができる．分散表現を獲得することにより，単語の概念を低次元の密なベクトルによって表現できる．

CBOW 法は，着目する単語に隣接している前後n個の単語を用いて，中心の単語を推定するモデルである．入力単語とする特定の単語周辺の単語に対し，特定の単語共起する事後確率を求めている（式 (1)）．

$$f(w_o|w_s) = \frac{\exp\{ \boldsymbol{v'}_{w_0}^{T} \cdot \boldsymbol{v}_{w_s}\}}{\sum_{w_v \in V} \exp\{\boldsymbol{v'}_{w_v}^{T} \cdot \boldsymbol{v}_{w_s}\}} \qquad (1)$$

w_oが特定の単語を表し，w_sが特定単語の周辺に位置する単語を表している．$\boldsymbol{v}_{w_s}$と$\boldsymbol{v'}_{w_o}$は単語を表すベクトルである．また，$\boldsymbol{v}$は入力ベクトルを表し，$\boldsymbol{v'}$は出力ベクトルを表す．Vは全ての語彙とする．

4.2 畳み込みニューラルネットワーク（CNN）

畳み込みニューラルネットワーク（CNN）とは，一般の順伝播型のニューラルネットワークとは異なり，全結合層に加えて畳み込み層とプーリング層から構成されるニューラルネットワークのことである．本研究では，一般的なCNNに見られる畳み込み層，プーリング層，全結合層が続き，最後に活性化関数としてsoftmax関数を用いた出力層から構成されるCNNを利用する．また，モデルの中には単語を高次元の実数ベクトルに変換するための埋め込み層も利用する．埋め込み層を追加することで，モデルの学習と同時に単語の重みも学習することができる．

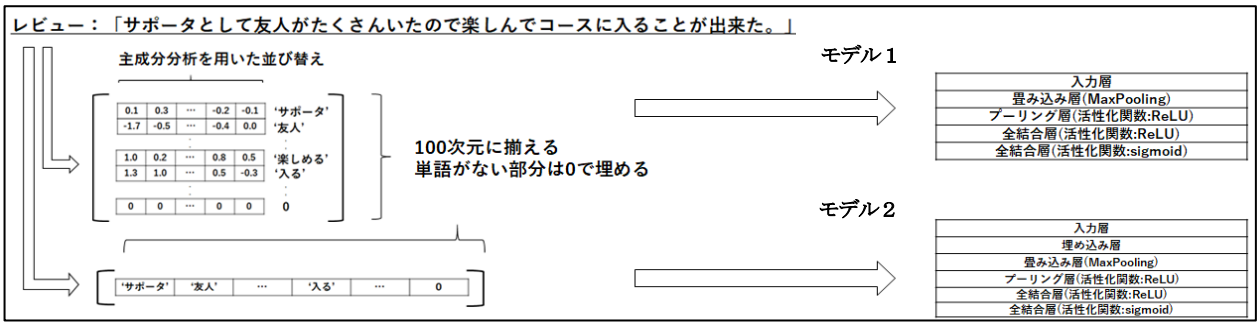

図 1: 入力層に入力するデータの作成方法および CNN 内の層構造

4.3 高評価レビュー判別モデルの提案

本研究では2種類の入力データを作成し，それぞれをCNNに入力した時のモデルの精度の違いを比較する．どちらも各レビュー内に含まれる名詞，形容詞，動詞のみを用いる（図 1）．

1つ目の入力データは，word2vecと主成分分析により，レビューデータを加工した入力データである．はじめに，word2vecを利用し，レビューの中に登場する単語全てを100次元のベクトルで表現する．次に主成分分析を行い寄与率が高い主成分の主成分負荷量の順番にベクトルを並べ替えたものを各レビュー内に登場する単語に照合し，列に単語ベクトルの情報を持ち，行にレビュー本来の配置情報を有する１チャネルの画像を作成し入力する．（モデル1）．

2つ目の入力データは，レビュー中に出現した単語とリスト番号（単語の出現順に定めた番号）を紐付け，そのまま入力データとしたもので，モデル内には埋め込み層を追加し，分散表現を獲得した後，畳み込み層へと伝達する方法である．（モデル2）．

それぞれのデータ処理により，CNN内で学習することが可能となる．どちらも単語をベクトルで表現した後，畳み込みを行うという部分は同じであるが，単語の重み付け方法が異なる．なお，モデルの精度の違いを比較するため，ここでは畳み込み層以降の層にプーリング層，全結合層，出力層を順に結合したモデルのaccuracy評価のみ記述する．

4. 分析結果および考察

モデル1の入力画像は，2つの次元で単語の並びの情報を有することがわかる．しかし，画像データの場合の場合，単語が本来入らない部分を0で置き換えているため，画像的特徴が画像の一部分に集中した結果，うまく特徴が捉えられなかったと考えられる（図 2）．

一方，モデル2（図3）の結果からは，モデル1よりもが良いことがわかった．これは重みの学習と同時に単語自体の重みも学習しているため，共起の特徴をうまく捉えながら学習することができたためだと考えられる．

しかし，どちらのモデルも損失関数は収束せず，テストデータでのaccuracyも低いままであることから，過学習が起きており，さらなるモデルの工夫が必要である．

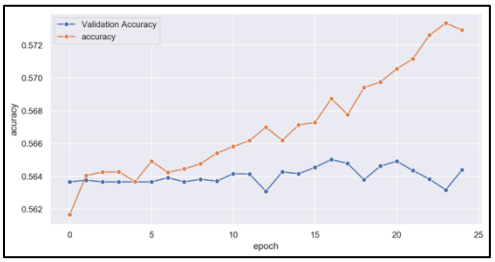

図 2：モデル1によるaccuracy推移

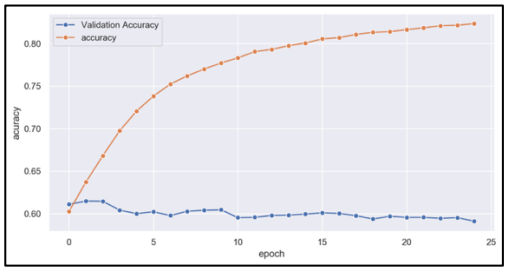

図 3：モデル2によるaccuracy推移

参考文献

本研究にあたり，データを提供いただいた企業に感謝申し上げます．本研究は JSPS 科研費 19K01945, 17K13809 の助成を受けたものです．

参考文献

[1] Y. Kim, “Convolutional Neural Networks for Sentence Classification”, *Conference on Empirical Methods in Natural Language Processing*, pp.1746-1751 (2014)