

# 深層学習を用いたツイートからの観光情報抽出手法

橋田 修一<sup>†</sup> 酒井 達弘<sup>††</sup> 田村 慶一<sup>††</sup>

<sup>†</sup> 広島市立大学 情報科学部

<sup>††</sup> 広島市立大学大学院 情報科学研究科

## 1. はじめに

近年、国内の観光産業に大きな注目が集められている。しかし、観光地運営において、観光客らが感じた不満や満足の調査はアンケート方式で行われることが多く、幅広い意見が収集されにくい。そこで、本研究では観光客の意見を幅広く収集するために、観光地に対して投稿されたツイート(観光ツイートと呼ぶ)から分析に必要なツイートの分類を行う手法を提案する。分類手法として、畳み込みニューラルネットワーク(以下、CNN)と分散表現を用いた手法を提案する。

## 2. 観光ツイートの分類

観光地に対して投稿されるツイートに分類ラベルを与えることで区別する。本研究では、観光地に対して意見・感想が含まれているツイートを観光地に関する意見が含まれるクラス(OIクラス)、事実や観光地に関係ないもしくは意見・感想でないツイートは観光地に関する意見を含まないクラス(ONIクラス)とする。OIクラスの中で、満足に関するツイートをPositiveクラス、不満に関するツイートはNegativeクラスとして考える。このように観光ツイートを分類することで観光地に関する多角的な視点の解析に用いることができると考えられる。

## 3. 分散表現とCNNモデル

Kim[1]は分散表現とCNNモデルを組み合わせた分類モデル(以下、Yoon-Kim'sモデル)を提案した。このモデルでは、CNNモデルに分散表現を用いた表現方法を利用し、テキストの分類をモデルに学習させる。入力層として、埋め込み層が用いられ、単語は入力された識別番号から  $m$  次元のベクトルに変換される。このベクトルを分散表現と考えることができる。また、畳み込み層ではフィルタサイズを(高さ, 幅)=( $k, m$ )と設定し、 $k$  単語の分散表現を畳み込み、1つのフィルタで1つの表現が抽出できることが考えられ、プーリング層では最も反応したフィルタの値を抽出することができる。最後に全結合層と出力層が構成される。

## 4. テキスト情報の多チャンネル化

提案モデルとして入力層を2層構成し、単語列情報と品詞列情報をそれぞれの入力層に与え、単語の分散表現と品詞の分散表現を重ねる手法を提案する。このとき、埋め込み層から出力された2次元行列を「チャンネル」と考え、2つの埋め込み層の出力を重ねて多チャンネル」と考えることができる。この手法を用いること

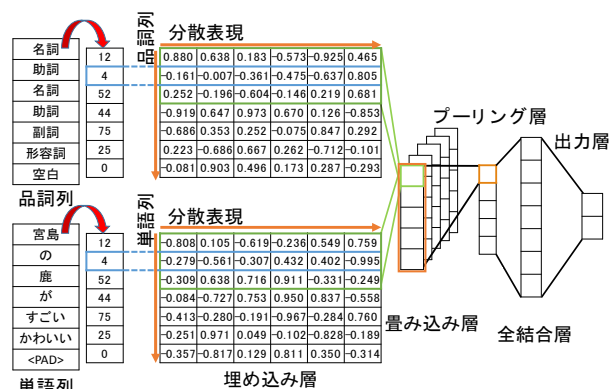


図1: 提案モデルの概要

で、訓練データ内に出現しない単語が現れた時、品詞情報により学習の補完を行えることが期待できる。モデルの概要を図1に示す。

## 5. 評価実験

評価実験では、キーワード“宮島”で検索したジオタグ付きツイートに対して手作業でラベル付けを行ったデータセットを用いる。OIクラスのツイート2740件中その中でNegativeクラスのツイート826件、Positiveクラスのツイート1914件、ONIクラスのツイート3253件である。また、実験では深層モデルとして、畳み込み層が3つのYoon-Kim'sモデル、提案モデルの2つを用いた。交差検定10回の平均を算出し、この処理を20回繰り返すことで、分類精度の統計をとった。モデルの精度を表1に示す。

表1: モデルの分類精度(正解率)[%]

モデル	OI⇔ONI	Neg ⇔ Pos
ナイーブベイズ	77.6	78.4
Yoon-Kim'sモデル	82.6	87.8
提案手法	<b>83.0</b>	<b>88.4</b>

## 6. まとめ

本研究では、観光ツイートの分類をCNNモデルを用いて行う手法を提案した。評価実験では提案手法の分類精度が最も高いことを確認した。また、評価実験から、品詞の情報を用いることで学習の補完を行うことができることを確認した。今後はツイートを分類後の不満要素、満足要素の解析手法の実装を行いたい。

## 謝辞

本研究の一部は、総務省SCOPE(受付番号:162308002)と広島市立大学・特定研究費の支援により行われた。

## 参考文献

- [1] Yoon Kim : Convolutional neural networks for sentence classification, Proceedings of EMNLP 2014, pp.1746-1751, 2014.