The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

畳み込みニューラルネットワークを用いた観光ツイートの分類手法

Sightseeing Tweet Classificaiton using Convolutional Neural Networks

橋田修一 *1　田村慶一 *2　酒井達弘 *2
Shiuichi Hashida　Keiichi Tamura　Tatsuhiro Sakai

*1広島市立大学 情報科学部
Faculty of Information Sciences, Hiroshima City University

*2広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

Recently, the icreasing interest in the social media networking service (SNS) is remarkable. Many researchers work for the discovery of new information from the posted tweets. In this study, we focus on the tweets related to tourist spots for finding the spot's improvement and appeal point. However, tweets on Twitter site include not only descriptions that are related to sightseeing but also descriptions that are not related to sightseeing. As the solution, we use convolutional neural networks for classifying these tweets. Moreover, we propose the multi-channel method to improve the performance of classification. In this method, the model was passed two test information that is words and form classes. In experiments, several models are used for comparing each other, Naive Bayes, convolutional neural networks (CNN), long short-term memory, and the proposed model. The experimental results show that the proposed model can classify tweets better than the other models.

1. はじめに

近年，国内の観光産業における経済効果は目覚ましく，観光振興に大きな注目が集められている．そこで，観光地の発展を目指すため，観光地の満足点や不満点を調査し，その後の改善策や PR 計画に役立てる事業が盛んに行われている．しかし，実際に行われている観光客への調査はアンケート形式が多く *1，その結果が出るまでに時間がかかることや分析に多くの人手が必要となるようなコスト面の問題が挙げられる．そこで，ソーシャルメディア，特に Twitter 上に投稿されるツイートをリアルタイムに分析することで，観光地の満足点や不満点を明らかにすることが期待されている．

SNS ユーザの増加に伴って，投稿されたツイートから新たな情報の発見や分析を行う研究 [Jansen 09, Sakai 15] が盛んに行われており，その一つとして，観光情報に着目した研究がある [松本 14, 佐藤 15]．観光客が投稿したツイートから観光情報を抽出することで，リアルタイム性のある調査結果が得られることが期待できるが，実際には，Twitter 上に投稿されるツイートは様々なバリエーションのあるツイートが存在する．そのため，観光地の分析に使用できるツイートと使用できないツイートを分類する必要がある．

そこで，本研究では，観光地に関して投稿されたツイート（以下，観光ツイートと呼ぶ）に対する，畳み込みニューラルネットワークを用いた分類手法を提案する．具体的には，観光地に対する意見を含むツイートと含まないツイートの分類，観光地に対する満足を示すツイートと，不満を示すツイートの分類を行う．また，分類精度の向上のために，単語列だけでなく品詞列も組み合わせてモデルに入力する新しいモデルを提案する．

連絡先: 橋田修一，広島市立大学情報科学部，広島市安佐南区大塚東三丁目 4 番 1 号，082-830-1676，email : shu.hashida@gmail.com

*1　http://www.mlit.go.jp/kankocho/siryou/toukei/shouhidoukou.html

2. 観光ツイートの分類

観光ツイートとして投稿されたツイートは観光運営者に限らず，観光客が投稿したツイートも含む．そのため，さまざまなバリエーションを含んでいるツイートが多く存在する．ここでは，観光ツイートを分類するために，それぞれのツイートに対して付与される分類ラベルの定義を行う．

まず，投稿されたツイートが観光地に関する意見を含んでいるツイートであれば *Opinion_is_Included*(以下,*OI*) クラスとし，観光地に関する意見を含んでいないツイートであれば *Opinion_is_Not_Included*(以下,*ONI*) クラスと分類する．本研究における意見には観光地への評価や苦情が含まれるものとする．具体的には，*OI* クラスには観光地に対する満足を示す *Positive* クラスと不満を示す *Negative* クラスが含まれる．

次にツイートを分類する処理の流れを示す (図 1)．

1. Twitter 上に投稿されたツイートを，キーワード検索を用いて取得する．この時，キーワードには観光地名を示す単語が設定される．例として，広島県内の観光地であ

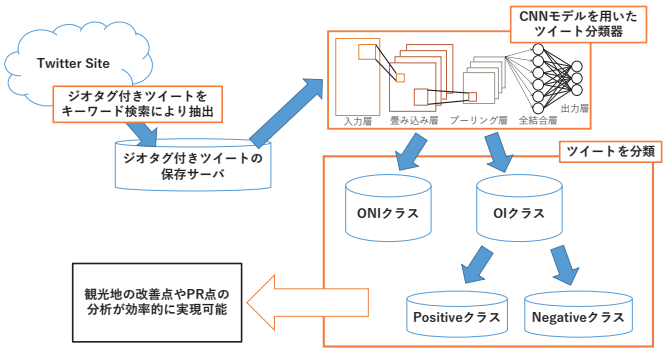

図 1: 観光ツイートの分類の流れ

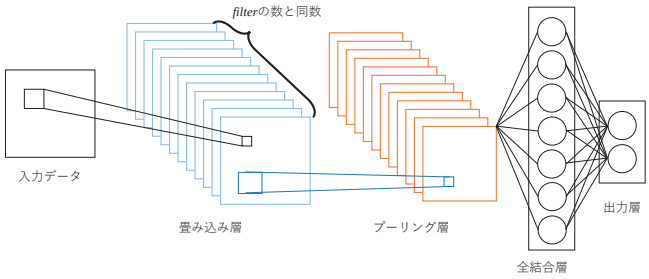

図 2: 畳み込みニューラルネットワークの概要

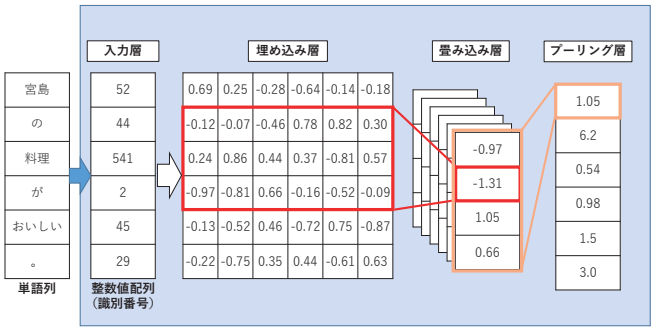

図 3: テキスト分類のための畳み込みニューラルネットワーク

る宮島に注目するときは，キーワードを “宮島” と設定する．

2. 取得したツイートを分類器で分類ラベルの予測を行う．取得したツイートは最初に $OI-ONI$ 分類器によって，OI クラスと ONI クラスに分類される．

3. 次に OI クラスに分類されたツイートに対して，$Positive-Negative$ 分類器が $Positive$ クラスと $Negative$ クラスの分類を行う．

3. 畳み込みニューラルネットワーク

畳み込みニューラルネットワーク [Lecun 98] とは，主に画像認識で用いられている手法であり，畳み込み層とプーリング層から構成される．図 2 に畳み込みニューラルネットワークの概要を示す．まず最初に，畳み込み層では，入力された 2 次元，もしくは 3 次元行列に対してフィルタを適用し，スライドしながら対象領域の畳み込み処理を行う．出力では，入力データの特徴マップが生成され，畳み込み層内にあるフィルタの数だけ特徴マップが生成される．次に，プーリング層では畳み込まれた特徴マップのそれぞれに対して，注目領域の代表値を算出し，情報圧縮が行われる．

4. 畳み込みニューラルネットワークを用いたテキスト分類手法

本章では，畳み込みニューラルネットワークと分散表現を用いたテキスト分類の研究について Kim の行った研究 [Kim 14] を紹介する．

4.1 分散表現

畳み込みニューラルネットワークを用いたテキスト分類の学習を行うため，テキストを数値空間上に表現する必要がある．そこで一般的に分散表現 [Mikolov 13b] を利用し，テキスト中の単語を数値化し，モデルに入力として与える．分散表現とは，単語を示す記号に対してベクトルを付与する手法である．分散表現を用いることで，単語間の意味関係が数値空間上で表現され，テキストを分散表現の系列として表現できる．分散表現の獲得方法について，word2vec[Mikolov 13a] を用いて大量のテキストデータから事前学習を用いて獲得する手法や，モデル内の埋め込み層を用いた獲得方法 [Collobert 11] などが挙げられる．

4.2 モデル

Kim が提案したモデル (以下，Yoon-Kim’s モデルと呼ぶ) では，埋め込み層の事前学習を行う手法と畳み込み層で複数のフィルタサイズを設定する手法で実験が行われている．分散表現獲得方法については，事前学習ありの埋め込み層と事前学習なしの埋め込み層の両方が示されている．事前学習ありの埋め込み層を用いた手法では，Google News から得られた 1 千億単語を用いて事前学習で収得した分散表現が使用される．また，事前学習なしの埋め込み層を用いた手法では，単語ベクトルの初期値をランダムに決定し，テキスト分類の学習を通して単語ベクトルの重みを更新する．

次に，モデルの構成について，主な層の説明とその役割について示す．図 3 に Yoon-Kim’s モデルの概要を示す．埋め込み層では，入力層から得られた単語識別番号の配列 (配列長:N) を受け取り，対応する単語ベクトル (次元数:M) を埋め込まれた，$N \times M$ の 2 次元行列を出力する．次に，畳み込み層では，単語ベクトルの次元数と揃えられたフィルタが適用され，テキストの特徴を抽出している．また，プーリング層では，畳み込まれた特徴マップ全体に対してマックスプーリングが適用され，最大値が抽出される．その後，全結合層，出力層を構成することで分類ラベルの予測を行う．

実験結果として，分散表現を事前学習で獲得することで分類モデルの精度向上が行えることが確認された．しかし，事前学習により得られた分散表現は，文法的に類似する単語のベクトルが類似することが確認された．もう一方の，事前学習が行われていない埋め込み層を用いる手法では，意味的な単語関係が表現されるため，意味的に似ている単語のベクトルが類似することが確認された．

5. 提案手法

本章では，Yoon-Kim’s モデルをベースとした提案手法を示す．本研究のモデルは Yoon-Kim’s モデルでも，事前学習なしの埋め込み層を用いるモデルを用いる．従来手法では畳み込みニューラルネットワークに単語列のみを入力として与えていたが，提案手法では単語列に対応する品詞列を与える．品詞列を与えることで，学習の促進が期待でき，テキスト中の各単語が持つ情報を増やすことで分類精度の向上が期待できる．主な手順を次に示す (図 4)．

1. 訓練データ内のツイートの分かち書きを行い，そこから単語識別番号を記録する辞書を作成する．

2. 次に，分かち書きされたツイートに対して，それぞれの

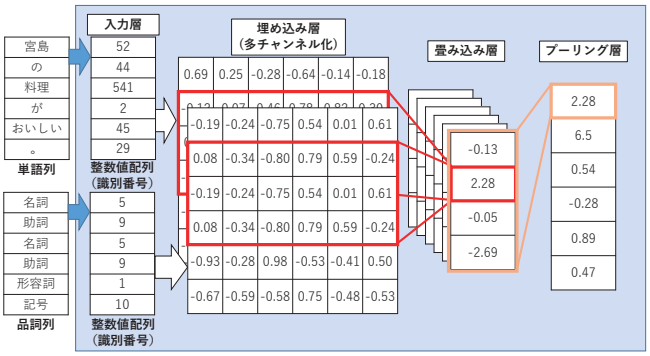

図 4: 提案モデルの概要

単語に単語識別番号を割り振る．テストデータ内において，出現した単語の単語識別番号を記録した辞書に存在しなければ，単語には空白を示す識別番号 “0” を格納し，品詞列にはその単語の品詞が格納される．これにより，単語列と品詞列の位置関係の差がなくなり，テキストの特徴を抽出することができる．

3. 整数値番号に変換された単語列と品詞列はモデルの入力層に与えられる．モデル内には 2 つの入力層と埋め込み層が存在し，それぞれの埋め込み層で単語の分散表現，品詞の分散表現を学習し，埋め込み処理を行う．その後，畳み込み層に与えるために，それぞれのチャンネルを積み重ねることで「多チャンネル化」を行う．

6. 実験

提案手法を評価するために評価実験を行った．本章では実験結果を示す．

6.1 データセット

本実験で用いるデータセットについて説明する．本研究では，実際に投稿された観光ツイートを，Twitter API を用いてキーワード検索 “宮島” により取得し，手作業でラベル付けを行った．ラベルは，OI クラスに含まれる $Positive$ クラスと $Negative$ クラス，そして ONI クラスとした．これらの 3 つのラベルはそれぞれのツイートに付与され，1 つのツイートに 1 つのラベルを付与する．表 1 にそれぞれのクラスに属するツイート数の内訳を示す．

表 1: データセット内の内訳

$Positive$	$Negative$	ONI	計
1914	826	3253	5993

本研究では比較モデルとして，次に示すモデルを用いて，10 回交差検定を行い，テストデータに対する分類精度の平均値をそのモデルの精度として評価を行う．また，Yoon-Kim’s モデルでは 3 手法を用いて，実験を行った．

- Naive Bayes
  - Naive Bayes を用いて分類を行う．
- Long short-term memory(LSTM)
  - 深層学習の手法の一つとして知られ，自然言語処理分野で多くの研究に使われているモデルである．時系列データに対して有効な手段として，長期依存の学習が行える．
- Yoon-Kim’s モデル
  - Kim の提案したモデルである．本研究では，次の 3 つの手法を用いて分類の学習を行った．
    1. 訓練データに対して word2vec を使用し，事前学習を通して得られた重みを用いて単語の埋め込みを行うモデル
    2. 訓練データとは別のデータセットに対して事前学習を行い，その単語の重みを用いて単語の埋め込みを行うモデル
    3. モデルに対する分類の学習を通して，単語の重みを学習する埋め込み層を用いて単語の埋め込みを行うモデル
- 提案手法
  - 本研究の提案モデルであり，Yoon-Kim’s モデルをベースとして入力に品詞列を追加したモデルである．

また，モデル内に Dropout を設定しているため，10 回交差検定を 20 回行い，分類精度の統計値を算出した．それぞれのモデルの精度を $OI-ONI$ 分類は表 2 に，$Negative-Positive$ 分類は表 3 に示す．

実験結果から，$OI-ONI$ 分類において，比較モデルよりも提案モデルが高い分類精度を示すことが確認できた．しかし，$Negative-Positive$ の分類では提案モデル以外のモデルで高い精度が確認された．これはデータセット内の $Negative$ クラスと $Positive$ クラスのツイート数に大きな偏りがあることが考えられる．

7. まとめ

本研究では，観光地の改善点や PR 点を分析するため Twitter 上に投稿される観光ツイートに着目し，畳み込みニューラルネットワークを用いて観光ツイートを分類する手法を提案した．また，多チャンネル化を行うことで分類精度の向上を図った．実験では，実際に投稿されたツイートに対してモデルの学習，分類を行い比較実験を行った．実験結果として，提案モデルの $OI-ONI$ 分類における分類精度が最も良いことが分かった．しかし，$Negative-Positive$ クラスの分類精度では他のモデルに劣る部分があることが確認できた．今後の課題として，分類されたツイートに対して観光地の改善や PR に関する情報の抽出が考えられる．具体的には，観光地の「どの施設が」「悪いか，もしくは良いか」，また,「それはなぜなのか」などの要素の抽出が考えられる．

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

[Collobert 11] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P.: Natural Language Processing (Almost) from Scratch., *Journal*

表 2: モデルの評価値 ($OI - ONI$ 分類)

モデル	OI クラス			ONI クラス			正解率
	再現率	適合率	F 値	再現率	適合率	F 値	
ナイーブベイズ	0.832	0.649	0.728	0.752	0.890	0.814	0.776
LSTM	**0.815**	0.777	0.795	0.821	0.851	0.835	0.817
Yoon-Kim's CNN model (1. 事前学習済み分散表現)	0.718	0.645	0.677	0.727	0.784	0.753	0.721
Yoon-Kim's CNN model (2. 事前学習済み分散表現)	0.591	0.705	0.642	0.705	0.587	0.639	0.641
Yoon-Kim's CNN model (3. 埋め込み層)	0.830	0.781	0.804	0.825	0.864	0.844	0.826
Yoon-Kim's CNN model (品詞列のみ)	0.784	0.743	0.761	0.794	0.826	0.808	0.788
提案手法 (Yoon-Kim's CNN model)	**0.832**	**0.789**	**0.809**	**0.831**	**0.865**	**0.847**	**0.830**

表 3: モデルの評価値 ($Negative - Positive$ 分類)

モデル	$Negative$ クラス			$Positive$ クラス			正解率
	再現率	適合率	F 値	再現率	適合率	F 値	
ナイーブベイズ	0.605	**0.831**	0.700	**0.914**	0.763	0.831	0.784
LSTM	0.825	0.752	0.784	0.898	0.930	0.913	0.876
Yoon-Kim's CNN model (1. 事前学習済み分散表現)	0.708	0.448	0.541	0.795	0.917	0.851	0.775
Yoon-Kim's CNN model (2. 事前学習済み分散表現)	0.618	0.205	0.299	0.733	**0.943**	0.825	0.721
Yoon-Kim's CNN model (3. 埋め込み層)	0.835	0.746	0.789	0.897	0.936	0.915	0.878
Yoon-Kim's CNN model (品詞列のみ)	0.754	0.471	0.569	0.804	0.929	0.861	0.791
提案手法 (Yoon-Kim's CNN model)	**0.840**	0.764	**0.797**	0.903	0.936	**0.918**	**0.884**

of Machine Learning Research, Vol. 12, pp. 2493–2537 (2011)

[Jansen 09] Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A.: Twitter Power: Tweets As Electronic Word of Mouth, *J. Am. Soc. Inf. Sci. Technol.*, Vol. 60, No. 11, pp. 2169–2188 (2009)

[Kim 14] Kim, Y.: Convolutional Neural Networks for Sentence Classification, in Moschitti, A., Pang, B., and Daelemans, W. eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1746–1751, ACL (2014)

[Lecun 98] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, in *Proceedings of the IEEE*, pp. 2278–2324 (1998)

[Mikolov 13a] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781, (2013)

[Mikolov 13b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality, in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3111–3119 (2013)

[Sakai 15] Sakai, T., Tamura, K., and Kitakami, H.: Emergency Situation Awareness during Natural Disasters using Density-based Adaptive Spatiotemporal Clustering, in *Database Systems for Advanced Applications, DASFAA 2015 International Workshops, SeCoP, BDMS, and Posters, Hanoi, Vietnam, April 20-23, 2015*, Vol. 9052, pp. 155–169 (2015)

[佐藤 15] 佐藤 圭, 関 良明 : Twitter を用いた観光地の混雑状況分析手法の検討, 情報処理学会研究報告. SPT, セキュリティ心理学とトラスト, Vol. 2015, No. 15, pp. 1–6 (2015)

[松本 14] 松本 義之, 藪内 賢之 : B3-3 Twitter からの地域・観光情報収集とその有用性の検討 (B3 時系列・言語情報, 一般講演), バイオメディカル・ファジィ・システム学会大会講演論文集, Vol. 27, pp. 87–88 (2014)