1T-07

ライフログデータ抽出のための行動ツイート分類

安江　駿亮†　　六沼　元貴†　　杉本　徹‡

芝浦工業大学大学院理工学研究科†　　芝浦工業大学工学部‡

1. 研究背景と目的

近年、個人の行動の記録であるライフログに注目が集まっている。ライフログデータは個人の行動予測や生活改善への活用が期待できる。竹内らは、ライフログとあらかじめ決まっているスケジュールに基づいた未来予測提示によるタスク管理手法を提案した[1]。この研究では、ライフログを分析して未来のタスクの進捗を予測提示することで、タスクを円滑に進めるように利用者を促すことができた。また相澤は、特定の出来事だけを記録し応用するライフログとして、食に特化したライフログシステムを提案し、Web 上で公開している[2] [3]。このシステムは、携帯やデジタルカメラで撮影した写真を解析し、それが食事画像であれば食事バランスを推定し、その結果と写真をログとして保存し可視化するものである。毎日の食事バランスを確認できるので、生活改善の助けになると期待される。

一方で、マイクロブログの一種である Twitter[4]が普及している。Twitter では、日々多様なユーザにより様々な発言（ツイート）が投稿されている。この中にはユーザのライフログを含んだツイートも多く存在する。ツイートに含まれるライフログデータを抽出し分析することでユーザの行動予測や生活改善への活用が期待できる。しかし、膨大なツイートの中からユーザの行動を含むツイートのみを人手で収集し分析するのは困難である。

そこで、本研究では Twitter に投稿されるツイートからライフログデータを抽出する前段階として、ユーザの行動を含むツイート(以下行動ツイートと呼ぶ)と含まないツイート(以下非行動ツイートと呼ぶ)を機械学習によって分類する手法を提案する。機械学習は教師あり学習法である SVM(Support Vector Machine)を用いる。

2. 研究手順

以下のような手順で研究を行った。

Behavioral-tweet Classification for Lifelog Data Extraction
†Shunsuke Yasue, †Genki Rokunuma, ‡Toru Sugimoto
†Graduate School of Engineering and Science, Shibaura Institute of Technology
‡College of Engineering, Shibaura Institute of Technology

① ツイートの収集

分類器を訓練するためのツイートは Twitter Streaming API を使用して収集した。この時、人手での分類が効率的に出来るように以下の条件を設けてツイートを収集した。

・日本語である
・リプライやリツイートでない
・URL を含まない
・文字数が 5 文字以上 40 文字以下

② ツイートの分類

①で収集したツイートを人手で行動ツイートと非行動ツイートに分類し、それぞれ 500 件の計 1000 ツイートを用意した。これらのツイートは分類が妥当であるかを協力者に確認してもらった。

③ ツイートのベクトル化

ツイートを BoW(Bag of Words)でベクトル化した。まず分類したツイート群を MeCab[5]を用いて形態素解析し、表 1 に示す品詞の単語を特徴語として抽出した。その後、各ツイート中の特徴語の出現回数を求めて特徴ベクトルに変換した。

表 1 抽出する特徴語の品詞

品詞	品詞細分類	活用型
動詞	自立	-
形容詞	自立	-
名詞	サ変接続	-
助動詞	-	特殊・タ
助動詞	-	特殊・ダ
助動詞	-	特殊・タイ

④ 分類器の訓練

③でベクトル化したツイート群を用いて分類器を訓練した。

⑤ 評価実験用ツイートの収集

評価実験用のツイートを同様の手順で収集しベクトル化した。評価用ツイートは行動ツイートと非行動ツイートそれぞれ 250 件の計 500 ツイートを用意した。

⑥ 評価実験
分類器で評価用ツイートを分類した。

3. 評価実験

3.1. データセット

評価実験にはあらかじめ用意した 500 件のツイートを使用した。

3.2. 分類器

分類器は SVM を用いた。実装には Python の機械学習ライブラリである scikit-learn を使用した。カーネルについてはグリッドサーチと 10 分割交差検定を行い最も結果が良かった linear カーネルを用いた。また、C パラメータは 1 とした。表 2 にその結果を示す。

表 2 カーネル毎の分類精度

カーネル	線形	RBF	シグモイド	多項式
分類精度	89.5%	86.2%	87.2%	50.1%

3.3. 実験結果

評価値として分類の正解率を以下の式で求めた。

$$\text{正解率} = \frac{\text{正しく分類できたツイートの総数}}{\text{評価用ツイートの総数}}$$

実験の結果、正解率は 0.882 となった。表 3 に実験結果の混同行列を示す。

表 3 混同行列

		分類結果		
		行動	非行動	合計
正解データ	行動	233	17	250
	非行動	42	208	250
	合計	275	225	500

表 3 より正解が非行動ツイートであるのに誤って行動ツイートに分類されているものが多いことが分かる。誤って分類されたツイートの特徴語を調べたところ、すべて「助動詞-特殊・タ」を含んでいた。

例①「岩手勝ったんだ」
例②「PS2 今買ったらいくらかな…」

例①は主語を考慮せず「助動詞-特殊・タ」を含むため誤って行動ツイートに分類されたと考える。主語を正確に判断し素性に加えることができればこの誤りを無くすことができると考える。

例②は「たら」が助動詞「た」の活用形であるために「助動詞-特殊・タ」として抽出され起きた誤りである。そこで「たら」を抽出しないように条件を変更し再度実験を行った。その結果、誤って行動ツイートに分類された「たら」を含む 8 ツイート中 7 ツイートを正しく分類することができた。しかし、「たら」を含む行動ツイートを誤って非行動ツイートとして分類してしまう誤りも発生した。例を以下に示す。

例①カミソリで頭剃ったら寒くて仕方ない
例②ふぁぼ整理してたら良いの見つけた

これは過去の完了の意味で用いられている点を考慮していないために起きた誤りである。より細かく意味を解析することができれば改善できると考える。

また、行動ツイートに正しく分類されたものの特徴語を確認した結果、こちらも「助動詞-特殊・タ」を多く含んでいた。そこで「助動詞-特殊・タ」の影響を調べるために、ベクトル化の条件を変更したときの正解率を調べる実験を行った。なお、使用する訓練データと評価データは前の実験と同じである。実験結果を表 4 に示す。

表 4 ベクトル化の条件を変更したときの正解率

	正解率
提案手法	0.882
助動詞無し	0.624
「タ」無し	0.668
「ダ」無し	0.878
「タイ」無し	0.880

表 4 より分類精度は「タ」に大きく影響を受け、「ダ」、「タイ」はほとんど影響がないことが分かった。

4. 結論

Twitter に投稿されるツイートが行動ツイートか非行動ツイートであるかを機械学習で分類する手法を提案した。その結果 88.2%の精度でツイートの分類を行うことができた。

参考文献

[1] 竹内 俊貴, 田村 洋人, 鳴海 拓志, 谷川 智洋, 廣瀬 通孝 : ライフログとスケジュールに基づいた未来予測提示によるタスク管理手法,情報処理学会論文誌,Vol.55, No.11, pp.2441-2450, 2014.
[2] 相澤 清晴 : ライフログの実践的活用:食事ログからの展望, 情報処理, Vol.50, No.7, pp.592-597, 2009.
[3] 写真で簡単ごはん日記 : http://www.foodlog.jp/
[4] Twitter : https://twitter.com/
[5] MeCab : http://taku910.github.io/mecab/