1T-06

経験を述べたツイートを対象とした行動カテゴリ分類

六沼　元貴†　　安江　駿亮†　　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. はじめに

近年，Facebook や Twitter など，ユーザが自由に短文をインターネット上に投稿できるサービスが多く利用されている．投稿されたこれらの情報は，企業のマーケティングやトレンド分析を目的として収集される機会も多い．特に Twitter 上で呟かれるツイートには，ユーザが何をした，どこへ行ったなど，行動や経験に関する情報を含むものが多く見られる．これらの情報を解析することは，コンピュータがユーザの行動パターンや嗜好情報を理解し，正しく記録するために重要な処理である．またこれらの情報は，例えば飲食店の推薦システムにおいて，ユーザ一人ひとりの好みを考慮した上で適切な情報を提供したり，対話システムにおいて，ユーザが親近感を抱きやすい応答を行う技術などに応用することができる．

本研究ではこのような経験情報を含むツイートを対象とし，それらを機械学習を用いて行動カテゴリへと分類する方法を提案する．

2. 行動カテゴリ

本研究では，行動カテゴリという概念を定義し，収集したツイートをそれらに分類する．行動カテゴリとは，ユーザが経験した行動の種類を表す分類である．総務省統計局が『社会生活基本調査』という国民アンケートで使用している分類[1]を雛形としており，人手による分類実験を行うことで，不要なカテゴリの削除，および必要なカテゴリの追加を行った．結果として，本研究では 21 個の行動カテゴリを定義した．行動カテゴリの一覧を，表 1 に示す．

3. 研究手法

本研究では，機械学習を用いてツイートを自動的に適切な行動カテゴリへと分類することを目標としている．具体的な分類の手法としてロジスティック回帰分析を用いる．ロジスティック回帰分析は，説明変数の値から目的変数の発生確率を予測する統計学的な手法である．本研究では，1 つのツイートに含まれる単語を説明変数とし，各行動カテゴリを目的変数として設定する．

表 1. 行動カテゴリの一覧

睡眠	食事	通勤・通学
移動	仕事	学業
学習・訓練	家事	買い物
テレビ・ラジオ	読書	サブカルチャー
趣味・娯楽	休養	旅行・外出
スポーツ	社会参加活動	医療
育児	交際・付き合い	その他

4. ツイートの収集と教師データ作成

本研究では，分類の対象として『経験情報を含むツイート』を扱う．経験情報とは，ユーザが何をした，どこへ行ったなど，行動や経験に関する情報である．日本語で書かれたツイートを多数収集し，その中からリプライ（他ユーザへの返信文），ハッシュタグ（#から始まる文字列），絵文字や記号などを除く．さらにそれらのツイートから，人手で『経験情報を含むツイート』を抽出する．抽出作業は大学院生 2 名による人手の作業で行った．

こうして集められた『経験情報を含むツイート』を行動カテゴリへと分類する．表 1 の 21 個の行動カテゴリそれぞれについて詳細な説明文と例文を設定し，それに従って人手で行動カテゴリへと分類する．この作業も大学院生 2 名によって行った．こうして集められた『経験情報を含むツイート』と行動カテゴリのペアを教師データとし，機械学習に用いるものとする．

5. 評価実験

5-1. 行動カテゴリ分類の有用性に関する実験

行動カテゴリ分類の有用性を確認するために評価実験を行った．内容は，雑談対話システムを想定して，被験者が行った行動や経験した出来事を入力してもらい，それに対するシステムの応答文を見て，いくつかの質問に答えてもらうアンケート形式の実験である．被験者は大学生 5 名である．今回の実験を行うにあたり，行

Behavior category classification of tweets that express user's experiences
†Genki Rokunuma, †Shunsuke Yasue, ‡Toru Sugimoto
†Graduate School of Engineering and Science, Shibaura Institute of Technology
‡College of Engineering, Shibaura Institute of Technology

動カテゴリごとにそれぞれ 3 つずつの応答文を事前に設定した．被験者によって入力された発話文を行動カテゴリに分類した後に，該当する行動カテゴリに設定された応答文をランダムに 1 つ出力する．被験者とシステムのやり取りの例を図 1 に示す．ベースラインとして，被験者の行動カテゴリと無関係に，別途用意した汎用的な 10 個の応答文からランダムに 1 つを選んで出力するシステムをシステム I とし，行動カテゴリを利用して応答文を選択するシステムをシステム II とした．質問内容は，各システムの応答文の多様性や適切さを 5(とても当てはまる)から 1(当てはまらない)の 5 段階で評価してもらうものである．被験者 5 名の評価の平均値を表 2 に示す．どの質問に対してもシステム II はシステム I の評価を上回り，行動カテゴリ分類の有用性を確認することができた．

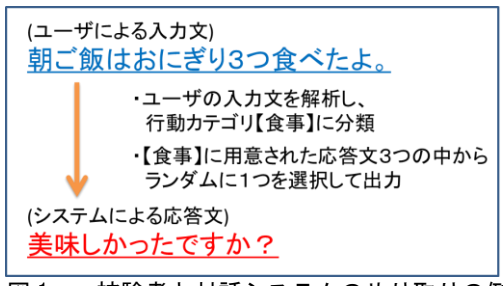

図 1． 被験者と対話システムのやり取りの例

表 2． 行動カテゴリ分類の有用性に関する実験結果

質問内容	システム I	システム II
応答の多様性	2.8	4.8
話題の適切さ	2.2	4.2
対話の面白さ	2.2	4.0
また対話したいか	2.0	3.8
行動の理解度	1.8	4.2
親近感を抱いたか	2.4	4.2

5-2. 行動カテゴリ分類の精度に関する実験

ロジスティック回帰分析を用いた行動カテゴリ分類の精度を確認する評価実験を行った．今回の実験では，収集した『経験情報を含むツイート』765 個それぞれを人手で行動カテゴリへ分類し，教師データとする．ロジスティック回帰分析の計算には LIBLINEAR[2]を用いる．テストデータとしては，収集データ数が最も多かった【サブカル】カテゴリのツイートを使用する．【サブカル】カテゴリは，ユーザがゲームで遊んだり漫画を読んだり等，サブカルチャーに関する経験を含んだツイートが分類されるカテゴリである．収集した教師データ 765 個のうち 9.5%にあたる 73 個のツイートが【サブカル】カテゴリへと分類されている．テストデータとして，教師データに用いたツイートとは別に 30 個の【サブカル】カテゴリへと分類されたツイートと，30 個の【サブカル】以外へと分類されたツイートを用意した．今回はこれらのテストデータが【サブカル】カテゴリへと分類されるか否かの 2 値分類を行う．

はじめに各ツイートに対して形態素解析を行い，MeCab[3]を用いてツイートから動詞および名詞を素性として抜き出して単語辞書を作る．ここで，文の意味を表しにくい数値，記号，顔文字や，多くのツイートに共通して含まれる動詞『する』および『てる』をストップワードとして除外した．作成した単語辞書をもとに，ツイートごとに Bag of Words ベクトルを生成する．この Bag of Words ベクトルを説明変数とし，ロジスティック回帰分析を行った．

教師データ 765 個，およびテストデータ 60 個を用いて行った実験の結果を表 3 に示す．

表 3. 行動カテゴリ分類の精度に関する実験結果

		分類結果		
		【サブカル】	【サブカル】以外	計
正解	【サブカル】	6	24	30
	【サブカル】以外	0	30	30
	計	6	54	60

分類精度は 60%であった．このように高いとは言えない値になった理由として，教師データ数の不足や，説明変数に対して次元削減などの処理を行っていないことが考えられる．

6. おわりに

本研究では経験情報を含むツイートを対象とし，それらを行動カテゴリへと分類する方法を提案した．ユーザによって投稿されたツイートを適切な行動カテゴリへ分類することは，コンピュータがユーザの行動パターンや嗜好情報を理解し，正しく記録することにつながる．評価実験の結果，行動カテゴリの有用性を示すことができたが，分類精度については十分なものでなかった．今後は，分類精度の向上を目指すとともに，ユーザの情報を正しく理解する対話システムへの実装を実現したい．

参考文献

[1] 統計局ホームページ 平成 23 年社会生活基本調査 用語の解説 http://www.stat.go.jp/data/shakai/2011/pdf/kaisetua.pdf
[2] LIBLINEAR https://www.csie.ntu.edu.tw/~cjlin/liblinear/
[3] MeCab http://taku910.github.io/mecab/