RO-016

行動を表す単語に着目した Twitter からの行動抽出

Behavior Extraction from Behavioral Words on Twitter

矢野 裕司† 横井 健‡ 橋山 智訓†
Yuji Yano Takeru Yokoi Tomonori Hashiyama

1. はじめに

近年, ライフログに関する研究が注目されている. ライフログとは人の行動を記録したものであり, 生活改善や行動予測に活用することが期待されている. 従来のライフログでは, センサなどの自動で記録されるデータを利用したもの[1]や, メモ, 写真などの手動で入力したデータを利用したもの[2][3]が提案されている. また, 行動を記録するために必要となる行動抽出については, センサを用いた研究のほかに, 個人活動を記録したブログなどのテキストデータを用いた研究[4][5]が行われている. しかし, ブログなどでは投稿頻度が少なく, 行動を細かく抽出することが出来ない.

一方, 個々のユーザが短文を投稿し閲覧できるソーシャルネットワーキングサービスの一種である Twitter[6]が近年爆発的に普及しており, 膨大な投稿の中には投稿者の行動を表すものも多く存在している. Twitter での投稿は tweet と呼ばれており, 以降本稿でも tweet と呼ぶ. Twitter は, パーソナルコンピュータや携帯電話などを用いて気軽に投稿を行うため, ユーザは楽しみながら, 行動を記録されているという実感が少なく行動を記録することができる. また, Twitter は思ったことをリアルタイムに高頻度で投稿する場合が多く, 一般的なブログよりも, それぞれの行動そのものと, 行動の時間を細かく取得することが可能である. これらの点で tweet データは手動によるメモや一般的なブログなどの従来の行動抽出の対象よりも優れているといえる. しかし tweet データは, Web ページや一般的なブログよりもさらに文法として不正確な表現が多いため, 形態素解析などの従来の解析手法だけで解析することは難しい場合がある.

そこで本研究では, Twitter における個人の tweet データに含まれる行動を表す単語に着目し, その単語に目的語が必要な場合には目的語を補い, 投稿者の行動を抽出することを目的とする. そのために, 行動を表す単語を集めた辞書を作成する. 目的語に関しては, 従来の解析手法である形態素解析, 係り受け解析を用いて, 従来手法と提案手法を組み合わせることで抽出を行う. 目的語抽出の部分のみに形態素解析などを用いるため, 目的語が不要な場合については, 形態素解析の辞書に影響を受けない. なお本稿では, 行動とは「tweet の直前, 直後または tweet 中の動きを伴う自発的な行い」と定義し, 以降取り扱うこととする. また行動を表す単語は, 一般的な動詞と名詞, Twitter 特有の「なう」といった単語とし, 過去の tweet を用いて選択する.

以降, 第 2 章では本研究に関連する従来研究について述べる. 第 3 章では, 本研究で提案する手法について述べる. 第 4 章では実験方法を述べ, 第 5 章では実験結果を示し得られた実験結果について考察する. 最後に第 6 章で本研究の結論を述べる.

† 電気通信大学大学院情報システム学研究科
‡ 東京都立産業技術高等専門学校ものづくり工学科

2. 従来研究

本研究に関連した従来研究を紹介し, それらの従来研究と本研究との違いを示す.

2.1. テキストデータからの行動抽出

2.1.1. Web ページ・ブログからの行動抽出

Web ページのテキストデータから, 行動情報の抽出と今後の行動予測を行う研究[4]が行われている. この研究ではまず, 少量のテキストデータに形態素解析と係り受け解析を適用し, この結果からパターンマッチングを用いて行動情報を抽出し, 行動の推移の結果数を用いて行動間の推移を学習する. その学習結果を用いて, 同様に行動情報を抽出した新たなテキストデータに対して行動予測を行う. 行動の属性としては, "行動主", "動作", "対象", "場所", "時間"を対象としている. なお, 行動は動作単独ではなく上述の行動の属性を少なくとも一つ含むことを条件としている. 同様に, 外出状況について述べられたブログから行動情報を抽出する研究[5]も行われている. この研究ではまず, 動詞を行動動詞と非行動動詞に分類した行動動詞判定辞書を作成する. 次にこの行動動詞判定辞書と, ブログ本文を形態素解析にかけた結果と照らし合わせて行動を抽出している. 行動の属性としては, "何を", "どこで", "いつ", "誰と", "どのように"を対象とし, 「(名詞)を(動詞)」であれば"何を"に相当する情報である, というようにパターンマッチングを用いて判断している. また, パターンだけでなく名詞に条件も付け加えることにより, 文法的に有り得ない名詞と動詞の組み合わせを抽出することを防いでいる. これらの研究では, ブログや Web ページを対象として行動を抽出しており, それぞれの行動を行った時間や行動そのものが大まかにしか取得することができない. しかし, Twitter を用いて行動抽出をすることで, 行動を行った時間や行動そのものを, 細かく取得することができる. またこれらの研究では形態素解析の結果で抽出された動詞に着目することで行動の抽出を行っているが, tweet データはブログや Web ページよりもさらにインフォーマルなデータであるため, 形態素解析に依存したこれらの手法では行動を抽出することは困難であると考えられる. そこで本研究では形態素解析に加えて, 行動辞書を作成し, 行動辞書中の単語が含まれる tweet から行動情報を抽出する. また, 目的語の抽出には従来研究と同様に形態素解析, 係り受け解析を利用する. これにより, 形態素解析への依存を減らし, 非常にインフォーマルな tweet データに対しても行動抽出を行うことができると考えられる.

2.1.2. tweet からの行動抽出

本研究と同様, Twitter からの行動抽出に関する研究としては, 動詞に着目して tweet から行動となるものを抽出し, その行動を学内等小規模のマップ上に表示する研究[7]が報告されている. この研究では, 対象となるそれぞれの tweet における動詞を行動として抽出している. この際, 「コーヒーなう」のように動詞が抽出されなかった場合は, 予め作成された行為抽出エンジンにより動詞を補完する. 行為抽出エンジンでは, まず大量の tweet データから名詞と動詞の組合せを抽出する. そして, そのすべての名詞と動詞の組合せに対して自己相互情報量を求め, その値が最も高いものを名詞に対する動詞として補完し, 行為を抽出している. この研究では, 動詞の判別において形態素解析を用いているため, Twitter 特有の単語や形態素解析が難しい崩れた表現については行動を抽出することが困難である. そこで本研究では, 2.1.1 節でも述べたように, 形態素解析に加えて, 行動辞書を作成し, 行動辞書中の単語が含まれる tweet から行動情報を抽出する. しかし, 「なう」のように動詞が曖昧な単語に対しては, この研究と同様の手法を用いて動詞の補完を行うことで, 行動を抽出する.

また筆者らは, Twitter における tweet から行動を表す単語を集めた行動辞書を作成し, この行動辞書を用いて Twitter から行動を抽出する研究[8]も報告してきた. この研究では, あらかじめ大量の tweet データに対し文字 n-gram を適用し, その中から行動を表す単語を人手で選択し, 行動辞書を作成した. 次に, それぞれの tweet データと行動辞書の単語を比較し, tweet データ中に行動辞書の単語が含まれていれば, 行動を表していると判断して, 行動を抽出している. この研究では, データ数が不十分であり, 適合率および再現率が 0.4～0.5 程度であった. そこで, 本研究ではこの手法を元に, 適合率および再現率の向上を目指すとともに, 行動辞書の作成過程と行動の抽出結果についてより深く考察を行う.

2.2. 画像データからの行動抽出

画像データからの行動抽出に関する研究としては, 食事に着目した Foodlog[3]と呼ばれる食事ログの研究が行われている. Foodlog では, 栄養管理のサポートを目的としており, 食事の画像を撮影, 抽出および分析することによって, 栄養組成等を推測することができ, ユーザは自分が摂取した食品に関する情報を知ることができる. 具体的には, まずユーザが撮影デバイスで撮影したすべての画像を, 食事画像とその他の日常画像に分類する. 次に食事画像と分類された画像にどのような栄養バランスで構成されているかを推定する. 最後に集められた食事画像と栄養バランスを可視化し, 食生活の改善や次に食べる望ましいメニューを提案する. これにより, ユーザの食生活が改善されるという利点がある. しかし Foodlog では食事のみに着目しており, それ以外の行動に関する情報は取り扱っていない. また, 取り扱うデータは画像データである点も本研究とは異なる.

3. 提案手法

本研究では行動を抽出するために, 行動を表す単語に着目する. まず行動を表す単語の辞書(行動辞書)を作成する. 次にこの行動辞書を利用して, 行動を抽出する.

3.1. 行動辞書の作成

本研究で作成する行動辞書では, 登録する単語を図 1 のように, 目的語を必要とせず単語単体で行動を表す単語と, 目的語を必要とする単語の 2 種類に分類する. また目的語を必要とする単語についてはさらに, 一般動詞と, 「なう」という Twitter 特有の単語の 2 種類に分類する. 目的語が不要な単語としては, 「おはよう」や「おやすみ」, 「ほかる(風呂に入る)」といった単語, 目的語を必要とする一般動詞としては「食べる」や「買う」といった単語が例として挙げられる. これらの行動辞書の単語を収集するために, 大量の tweet から文字 n-gram により抽出し, 行動辞書における単語の候補とする. 単語の抽出手法には形態素解析も挙げられるが, 一般的な語を辞書として持つ形態素解析器では Twitter 特有の単語を抽出することができない. また行動辞書に登録すべき単語の候補は非常に多く, 人手で探索することは困難であるため, 人手で探索する前に, 以下の 1 から 7 および図 2 に示す手順に従い, 候補を削減する.

1. 文字 n-gram におけるnを 2 から Nまで変更しながら,複数の文字 n-gram の結果による候補$\boldsymbol{w_n}(n = 2,3,\dots,N)$を取得する. ここで$\boldsymbol{w_n} = \{w_{n1}, w_{n2}, \dots, w_{nm_n}\}$であり, m_nは n-gram の結果数, $w_{ni}(i = 1,2,\dots,m_n)$はそれぞれ n-gram の結果である.
2. 候補を削減する基準値を二つ設ける. それぞれの基準値は上の基準値を$Whigh^{xy}(x = 3,\dots,N),(y = 2,3,\dots,N-1)$, 下の基準値を$Wlow^{xy}$とする. ここで, xとyは比較する対象のそれぞれのnであり. $x > y$である. また, $Whigh^{xy}$および$Wlow^{xy}$は$0 < Wlow^{xy} < Whigh^{xy} < 1$である.
3. w_{xi}に含まれるw_{yj}の出現頻度とw_{xi}の出現頻度の比率r_{ij}^{xy}を(1)式により求める. ここで, $freq(x)$は文字列xの出現頻度である.

$$r_{ij}^{xy} = \frac{freq(w_{xi})}{freq(w_{yj})} \quad (1)$$

4. r_{ij}^{xy}が$Whigh^{xy}$以上であれば、w_{yj}を削除する。
5. r_{ij}^{xy}が$Wlow^{xy}$以下であれば、w_{xi}を削除する。
6. 3 から 5 をすべてのnの組み合わせにおけるすべての候補の組み合わせに対して行う。

最後に, 削減した候補から人手で主観的に行動を表す候補を選択して, 行動辞書の単語とする. 行動辞書には, 単語とその単語が示す行動および行動の経過を結合したもの, 行動主, 目的語の必要性, tweet と対比した行動の時間, 単語の読みを表記する. 単語は対象となる単語を表す. その単語

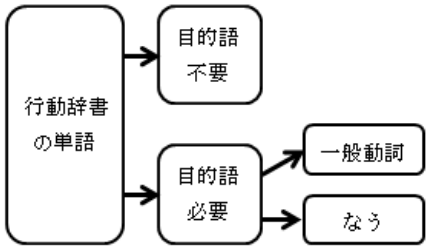

図 1 行動辞書の単語

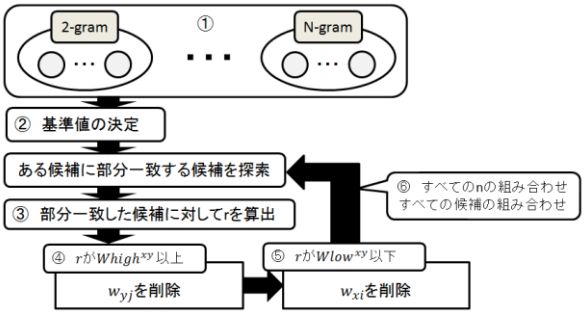

図 2 候補削減の手順

が示す行動には起床や帰宅等を表し，行動の経過として開始，終了または中を用いてこの二項目を組み合わせる．また，行動主は自分，他人のいずれかを用い，目的語の必要性は必要，不要のいずれか，tweet と対比した行動の時間には過去，現在，未来のいずれかを用いる．これらのラベル付けは人手で行う．

3.2. 行動の抽出

本研究では，個人の tweet の集合に行動辞書の単語が含まれた場合に，その単語が示す行動を行ったと判断する．行動辞書の単語は，図 1 に示したように目的語が不要な単語，目的語が必要な一般動詞，「なう」といった Twitter 特有の単語の 3 種類に分類できるため，行動辞書の単語が含まれるか否かの判断を行った後に，含まれた行動辞書の単語の種類に応じて，それぞれの手法で行動を抽出する．

目的語が不要な単語の場合には，行動辞書の単語が表す行動をそのまま抽出し，tweet の投稿時間と行動を記録する．例として，tweet に「おはよう」という単語が含まれていた場合はその tweet を投稿した時間に起床したと判断する．

目的語を必要とする一般動詞の場合には，行動抽出のために対応する目的語を抽出する必要があるため，行動辞書の目的語を必要とする単語を含む tweet における，単語以前の文章に形態素解析器および係り受け解析器を適用し，係り受け関係を求める．そして，助詞がなく直接行動辞書の単語に係っている名詞をすべて抽出し，目的語とする．また，格助詞である「を」，「に」に着目して，それらの助詞を通して行動辞書の単語に係っている名詞もすべて抽出し，目的語とする．上記以外の助詞に係っている言葉は抽出しない．これにより，目的語を必要とする単語に対しての目的語を得て，行動を抽出する．例として，「ご飯を食べた」という tweet であれば，「食べた」という行動辞書の単語に対して，「ご飯」という目的語を得て，行動を抽出する．

また Twitter には，Twitter 特有の「なう」という表現がある．この「なう」は一般的に"場所＋「なう」", "行動＋「なう」", "食べ物＋「なう」", "飲み物＋「なう」"の形で使われる場合が多く，行動を表している割合が高い．また，"動詞＋「なう」"といった二重になった状態で使われる場合も存在する．そのため「なう」に対しては，動詞補完の必要性を調べる必要がある．また，動詞補完が必要な場合に関しては動詞を補完する必要がある．そこで，「なう」は「居る」，「している」，「食べている」，「飲んでいる」およびただ語尾につけている場合の 5 パターンを想定して，動詞の補完を行う操作をする．まず，目的語を必要とする一般動詞と同様に，「なう」を含む tweet における，「なう」以前の文章に形態素解析器および係り受け解析器を適用し，係り受け関係を求める．そして，助詞がなく直接行動辞書の単語に係っている名詞をすべて抽出し，目的語とする．また，格助詞である「を」，「に」に着目して，それらの助詞を通して行動辞書の単語に係っている名詞もすべて抽出し，目的語とする．これらによって目的語が得られなかった場合には，この「なう」はただ語尾につけていると判断し，削除する．それ以外の場合には，文献[7]と同様に動詞の補完を行う．まず，予め大量の tweet データを形態素解析し，名詞と動詞の組合せの頻度を求める．そして，目的語となる名詞に対して(2)式に示す自己相互情報量 (Pointwise Mutual Information)PMI が高い尤もらしい動詞を補完する．

$$PMI(Noun, Verb_k) = \log \frac{p(Noun, Verb_k)}{p(Noun)p(Verb_k)} \qquad (2)$$

ここで，$Noun$は目的語となる名詞，$Verb_k$はそれぞれの動詞，$PMI(Noun, Verb_k)$は$Noun$と$Verb_k$の自己相互情報量である．また(2)式から，固定している$Noun$の確率$p(Noun)$や対数を除いても，大小関係は変わらない．従って，自己相互情報量の値が最も高い動詞を目的語に対する動詞として補完すると考えると(3)式のようになる．

$$Verb_{Noun} = arg \max_{Verb_k} p(Noun|Verb_k) \qquad (3)$$

ここで，$Verb_{Noun}$は目的語となる名詞$Noun$に対して補完する動詞である．「なう」は進行形で用いる場合が多いため，(3)式により補完された動詞は，その動詞の進行形を実際に補完する．

4. 実験方法

本実験では，行動の抽出を行う際に用いる行動辞書を作成し，その行動辞書を用いて行動を表すと考えられる tweet を抽出する．抽出結果は，適合率，再現率と F 値で評価を行う．また，Twitter 特有の「なう」という表現については動詞の補完をし，行動抽出を行う．なお，【自動】や【auto】といったものが含まれる tweet は，分析の妨げとなる自動発信された tweet であると判断し，行動抽出を行う tweet の対象から除外する．

tweet の取得には，Twitter 社が提供している Twitter Search API[9]を利用する．

4.1. 行動辞書の作成

行動辞書を作成するために，2011 年 6 月 7 日から 2012 年 12 月 21 日までの 564 日間で，150 人の無作為に選んだユーザ及び 2,081 人の有名人ユーザの tweet，計 10,509,978 件に対して文字 n-gram を適用する．n-gram における n は 2 から 10 まで変更して行う．tweet の平均文字列長は 53.10 文字である．また n-gram の結果の中から 11 回以上出現したパターンを n-gram の結果として利用した．なお，"あああ"のように同一文字のみで構成されている，または平仮名，カタカナ，漢字，「ー」以外の文字が含まれているものについては日本語で行動を表す単語はないと考え除外する．候補を削減する基準値である$Whigh^{xy}$は 0.95，$Wlow^{xy}$は 0.01 として実験を行う．削減した n-gram の候補の中から，行動を表して

いると考えられる単語を人手で抽出する. さらに 5 人によって, 抽出した単語に行動内容, 行動主, 目的語の必要性, tweet と対比した時間を人手でラベル付けし, それぞれで最も多く付けられたラベルを採用する. 人手によるラベル付けでは, ラベルを予め定められたいくつかの候補から選択する. また, 削減する基準値$Whigh^{xy}$および$Wlow^{xy}$による n-gram の候補数や候補の内容への影響を調べるために, $Whigh^{xy}$および$Wlow^{xy}$を変更して n-gram の候補の削減を行い, それぞれの基準値による n-gram の候補数と得られた行動を表す単語数を求める. 本実験では, $Whigh^{xy}$の値を 1.0, 0.99, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, $Wlow^{xy}$の値を 0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 として行う.

4.2. 行動の抽出

行動辞書が正しく作成されているかを調べるために, 実際のユーザにおける tweet から行動の抽出を行う. 今回の実験では, 無作為に選んだ 10 ユーザの 2012 年 1 月 21 日から 1 月 31 日までの 11 日間におけるリプライやリツイートを除いた tweet を対象とする. 11 日間の 1 ユーザ当たりの tweet 件数の平均は 445.8 件, 標準偏差は 218.0 件である. 従って標準偏差より, 様々な tweet 頻度のユーザが選択されたことがわかる. 正解データは, 5 人によって対象の tweet をすべて行動を表すか否かに分類し, その多数決を取り行動を表すとされた tweet を抽出して作成した. 正解データは 1 ユーザ当たり平均 64.4 件, 標準偏差は 41.3 件である.

行動の抽出では, 同時に目的語の抽出も行う. 対象 tweet は行動辞書の単語以前の文章を抽出してから, 日本語形態素解析システム JUMAN[10]により形態素解析を行い, 日本語構文・格解析システム KNP[11]により係り受け解析を行って, 目的語の抽出を行う.

評価は, 正解データを利用して, (4)式に示す適合率$precision$および(5)式に示す再現率$recall$と, 適合率および再現率の調和平均をとった(6)式に示す F 値$F\text{-}measure$で行う.

$$precision = \frac{R}{N} \tag{4}$$

$$recall = \frac{R}{C} \tag{5}$$

$$F\text{-}measure = \frac{R}{\frac{1}{2}(N+C)} \tag{6}$$

ここで, R は抽出結果のうち正解データと適合した tweet 数, N は抽出結果の tweet 数, C は正解データの tweet 数である. この適合率, 再現率, F 値について 10 ユーザそれぞれの値と 10 ユーザの平均値を算出し, 評価を行う.

また動詞に着目して, 形態素解析のみを用いた行動抽出の手法との比較を行うために, 本手法による実験と同じ tweet データや正解データを実験対象とし, 行動の抽出を行う. 比較実験として以下のように実験方法を設定し行った. まず対象の tweet データに対して形態素解析を適用し, 動詞が含まれていれば行動を表すとし, 抽出を行う. 最後に(4.2)式に示す適合率および(4.3)式に示す再現率と, (4.4)式に示す F 値を求め, 本手法で求めた評価の値と比較する.

4.3. なうに対する動詞の補完

本実験で対象としている「なう」というキーワードには, 3.2 節で述べたように複数の意味での使われ方がある. そこで, 本実験では「なう」という単語が含まれていた場合の動詞補完も行う. この実験では, 動詞補完のための学習データとして 2011 年 6 月 7 日から 2012 年 2 月 29 日までの 268 日間における, 150 人の無作為に選んだユーザ及び 2,081 人の有名人ユーザの tweet 約 500 万件のうち, 「居る」, 「する」, 「食べる」, 「飲む」の基本形, 連用形, 未然形, 仮定形を含む tweet 約 100 万件を用いる. この tweet をすべて日本語形態素解析システム JUMAN により形態素解析し, 名詞と動詞の組み合わせパターンを抽出する. なお, 名詞と動詞の組み合わせパターンとしてカウントするのは, 「名詞＋動詞」または「名詞＋助詞(複数を含む)＋動詞」の形式のパターンであり, それ以外のパターンはカウントしない. そして実際に動詞を補完するテストデータとして, 2011 年 6 月 7 日から 2012 年 2 月 5 日までの 184 日間における, 「なう」というキーワードを含む約 5 万件の中から, 無作為に 100 件の tweet を選び, 動詞を補完する.

5. 実験結果と考察

4 章で述べた実験方法に従い, 実験を行った結果を示す. 行動の抽出は, 本実験で作成した行動辞書を用いて行った.

5.1. 行動辞書の作成結果と考察

表 1 に, n-gram を適用して人手で抽出し, ラベル付けを行って行動辞書に記録した単語の例を示す. 表 1 における n-gram の順位は, 本研究の削減手法により削減した後の n-gram の件数の順位である. 本実験では, 表 1 のような単語を

表 1 行動辞書の単語の一部

n-gram の順位	単語	行動	行動主	目的語	時間	読み
96	します	行動開始	自分	必要	未来	します
287	おはよう	起床	自分	不要	過去	おはよう
557	おやすみ	就寝	自分	不要	未来	おやすみ
586	なう	行動中	自分	必要	現在	なう
842	やる	行動開始	自分	必要	未来	やる
1433	食べた	食事終了	自分	必要	過去	たべた
2632	美味しい	食事中	自分	必要	現在	おいしい
13369	帰ろう	帰宅開始	自分	不要	未来	かえろう
13381	もぐもぐ	食事中	自分	必要	現在	もぐもぐ

80 語抽出できた. 表 1 より, 行動を表す単語はある程度抽出できていると考えられる. また Twitter では行動を表す単語は上位に存在しており, n-gram を用いた本手法によってそれらの単語を抽出できたことがわかる. Twitter 上ではこれらの単語が頻繁に用いられているため, Twitter 上から抽出し作成した行動辞書によってある程度の行動を抽出できると考えられる. また, 上位の単語としては, 「します」や「やる」のように行動全般を表す一般的な単語や, 「おはよう」や「おやすみ」といった睡眠に関する単語が多いことがわかる. それらに加えて, 「なう」といった Twitter 特有の単語も多く用いられているため, Twitter における行動抽出を行う際には, このような Twitter 特有の単語に対する処理を行う必要があるといえる.

表 1 より, 本実験で作成した行動辞書では, 「おはよう」という単語が tweet に含まれる場合は「自分がこの tweet 以前に起床した」という解釈ができる. また, 「もぐもぐ」という単語が tweet に含まれる場合は「自分がこの tweet 中に食べている」という解釈ができ, 目的語抽出の処理により目的語を補うことで, 「何」を食べているのかという情報を追加する. このように, 本実験で作成した行動辞書は 80 語の単語に対しての解釈を与えている.

図 3 にそれぞれの$Wlow^{xy}$における$Whigh^{xy}$を変化した時の削減後の n-gram の候補数, 図 4 にそれぞれの$Wlow^{xy}$における$Whigh^{xy}$を変化した時の行動を表す単語として抽出できた単語数を示す. 図 4 では, $Wlow^{xy} = 0.01$, $Whigh^{xy} = 0.95$として抽出できた単語数である 80 語を基準として, それぞれの$Wlow^{xy}$および$Whigh^{xy}$として削減した後の候補に, この 80 語のうち何語含まれているかを求めた. n-gram により 11 回以上出現した候補のパターン数は 843,890 件であった. ここから, 同一文字のみで構成されている, または平仮名, カタカナ, 漢字, 「ー」以外の文字が含まれているものについて除外した後のパターン数は, 166,809 件であった. 図 3 より, $Wlow^{xy} = 0.01$, $Whigh^{xy} = 0.99$とした場合でも, パターン数を 42291 件と, 25%程度まで削減できていることがわかる. 従って, 本手法による削減は有効であると考えられ, 人手で単語を抽出する際の負担が軽減されている. また, 図 3 より, 削減された n-gram の候補数は$Whigh^{xy}$の値の増加に対してほとんど線形的に増加しているが, $Wlow^{xy}$に対しては, $Wlow^{xy}$を 0 から 0.01 に増加した場合に急激に減少しており, また, 0.01 から 0.05 に増加した場合にも減少している. それ以降の増加に対しては減少の度合いが小さい. 一方で, 図 4 より, 行動を表す単語として抽出できた単語数は, $Whigh^{xy}$の値の増加に対してはあまり増減がないが, $Wlow^{xy}$に対しては, $Wlow^{xy}$を 0.01 から 0.05 に増加した場合に急激に減少しており, それ以前または以降の増加に対しては減少の度合いが小さい. この結果より, 行動を表す単語を多く抽出するためには, $Wlow^{xy}$の値は 0.01 が望ましいと考えられる. 一方で, $Whigh^{xy}$の値は変化に対してほとんど増減がなく, 候補数は$Whigh^{xy}$の増加に対してほとんど線形的に増加しているため, 0.7 や 0.8 まで減少しても単語数への影響は少なく, 候補削減として有効であると考えられる. 本研究では, より多くの行動を表す単語を抽出することが望ましいため, $Whigh^{xy} = 0.9$, $Wlow^{xy} = 0.01$として削減することが最も良いと考えられる. これは, 二つの単語の候補を比較する際に, 文字数が長い単語の候補に対しては削減条件を厳しく, 文字数が短い単語の候補に対しては削減条件を緩く設定することを表す.

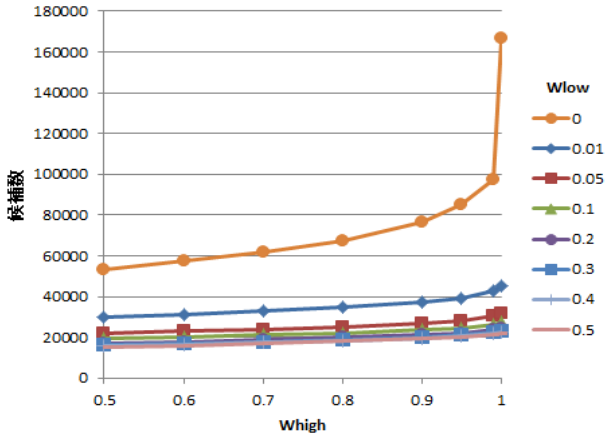

図 3 削減基準値を変更した時の候補数

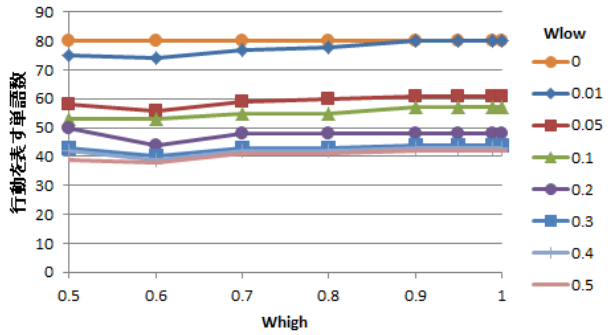

図 4 削減基準値を変更した時の抽出単語数

5.2. 行動の抽出結果と考察

5.2.1. 評価の結果と考察

提案手法および動詞に着目した手法による対象とした tweet の分析結果の分類数を表 2 に示す. 表 2 における tweet 件数はリプライおよびリツイートを除いた総 tweet 件数, 正解件数は正解データとして人手で抽出した tweet 件数, 行動 tweet 率は tweet 件数に対する正解データ件数の割合, 提案抽出件数および動詞抽出件数はそれぞれの手法により抽出した tweet 件数, 提案真陽性および動詞真陽性はそれぞれの手法で正解データと抽出結果が一致した場合(真陽性)の tweet 件数である. 表 2 より, tweet 数や行動を表している tweet の割合にはユーザによりばらつきがあることがわかる. また, 全ての tweet のうち行動を表している tweet の割合は平均で 0.15 程度であり, これらの tweet に含まれている行動を抽出することが目的となる. 提案手法では正解データの tweet 件数に対して抽出した tweet 件数が同等の件数であることがわかる. 一方で動詞に着目した手法では正解データの tweet 件数に対して抽出した tweet 件数が非常に多いため, 余分に抽出した tweet が非常に多いことがわかる. これは, tweet の文章には崩れた文章が多く, 形態素解析では動詞と誤って判断される確率が高いことを示している. また, 真陽性の件数は提案手法と動詞に着目した手法では

表 2 それぞれの手法における tweet 分析結果の分類数

	tweet 件数	正解件数	行動 tweet 率	提案抽出件数	提案真陽性	動詞抽出件数	動詞真陽性
ユーザ 1	962	172	0.18	190	119	673	130
ユーザ 2	384	60	0.16	60	39	236	51
ユーザ 3	290	25	0.09	33	13	176	20
ユーザ 4	675	58	0.09	67	31	345	26
ユーザ 5	529	65	0.12	59	26	350	47
ユーザ 6	338	34	0.10	48	21	263	32
ユーザ 7	347	61	0.18	64	46	269	38
ユーザ 8	352	52	0.15	86	33	294	43
ユーザ 9	235	36	0.15	46	28	158	26
ユーザ 10	446	81	0.18	72	54	276	60
平均	445.8	64.4	0.14	72.5	41	304	47.3

あまり変わらないが, 動詞に着目した手法では tweet 件数の 68%程度の tweet を抽出しているため自然と真陽性の件数が増えると考えられる. 一方で提案手法では 16%程度の tweet しか抽出していないため, 余分に抽出することは少なく適切に行動を表す tweet を抽出できていると考えられる. 本手法による行動抽出の評価結果を表 3, 動詞を着目した手法による行動抽出の評価結果を表 4 に示す. 表 4 より, 動詞に着目した手法での適合率の平均値は 0.15 であり, 最大のユーザでも 0.22 である. 行動を表している tweet の割合は 0.14 であり, この値と近く変わらないため, 無作為に tweet を抽出した結果とほとんど同じ結果である. 再現率は 0.74 程度であるが, これは正解データの tweet 件数に対して抽出した tweet 件数が非常に多いことによるものであると考えられる. 従って, 動詞に着目した手法は tweet データに対して有効ではないと考えられる. 一方で表 3 より, 本手法での適合率は約 0.55 であり, 最大のユーザで 0.75 であった. また, 本手法での再現率は約 0.63 であり, 最大のユーザで 0.78 である. 行動抽出では, 抽出漏れ無く行動を抽出できるように, 再現率が高いことが望まれると考えられる. 一方で誤って行動を抽出することも望ましくない. 適合率と再現率の調和平均である F 値を見ると, 動詞に着目した手法では平均値が約 0.25 である一方で, 本手法では平均値が約 0.58 である. 従って, tweet データから行動抽出を行う際には, 動詞に着目した手法よりも本手法のほうが有効であるといえる.

表 3 提案手法による行動抽出の評価結果

	適合率	再現率	F 値
ユーザ 1	0.63	0.69	0.66
ユーザ 2	0.65	0.65	0.65
ユーザ 3	0.39	0.52	0.45
ユーザ 4	0.46	0.53	0.50
ユーザ 5	0.44	0.40	0.42
ユーザ 6	0.44	0.62	0.51
ユーザ 7	0.72	0.75	0.74
ユーザ 8	0.38	0.64	0.48
ユーザ 9	0.61	0.78	0.68
ユーザ 10	0.75	0.67	0.71
平均	**0.55**	**0.63**	**0.58**

表 4 動詞に着目した手法による行動抽出の評価結果

	適合率	再現率	F 値
ユーザ 1	0.19	0.76	0.31
ユーザ 2	0.22	0.85	0.35
ユーザ 3	0.11	0.80	0.20
ユーザ 4	0.08	0.45	0.13
ユーザ 5	0.13	0.72	0.23
ユーザ 6	0.12	0.94	0.22
ユーザ 7	0.14	0.62	0.23
ユーザ 8	0.15	0.83	0.25
ユーザ 9	0.16	0.72	0.27
ユーザ 10	0.22	0.74	0.39
平均	**0.15**	**0.74**	**0.25**

図 5 に本手法および動詞に着目した手法における評価結果である表 3 と表 4 の適合率と再現率をグラフにしたものを示す. ここで, 図 5 におけるプロットのラベルはユーザの番号である. 図 5 より, 適合率および再現率のグラフでは, ユーザが二つのグループに別れていることがわかる. 提案手法における評価結果で適合率および再現率がともに高いユーザ 1, ユーザ 2, ユーザ 7, ユーザ 9, ユーザ 10 に関して, 動詞に着目した手法における評価結果を見ると, 動詞に着目した手法でも高い適合率であることがわかる. 一方で, 適合率および再現率がともに低いユーザ 3, ユーザ 4, ユーザ

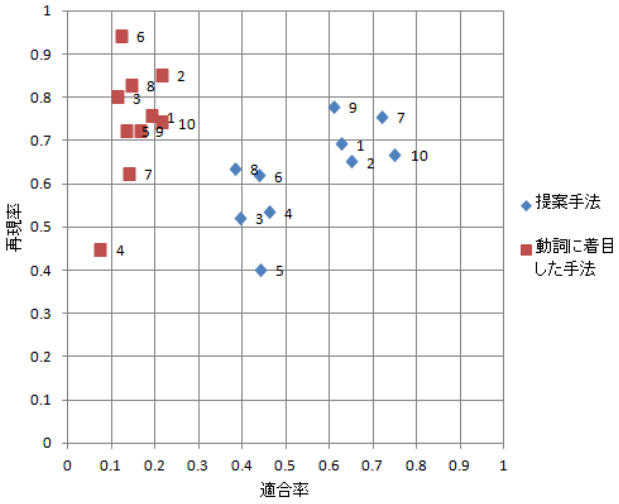

図 5 抽出結果の適合率と再現率

5, ユーザ 6, ユーザ 8 に関しては, 動詞に着目した手法でも低い適合率である. 動詞に着目した手法では, 形態素解析に依存しており, 崩れた表現が少ないユーザに関しては高精度で動詞を判別できると考えられる. しかし, 崩れた表現が多いユーザでは動詞を判別する精度は低くなると考えられる. 従って, 適合率および再現率が高いユーザは崩れた表現が少ないユーザ, 適合率および再現率が低いユーザは崩れた表現が多いユーザと考えられる. そのため, 本手法においても崩れた表現が多いユーザに比べ, 崩れた表現が少ないユーザのほうが高精度で行動を抽出することができる.

5.2.2. 抽出内容の結果と考察

正しく抽出された(真陽性の)結果の例を表 5, 誤って抽出された(偽陽性の)結果の例を表 6, 所望の tweet が抽出されなかった(偽陰性の)結果の例を表 7 に示す.

表 5 より就寝や起床, 入浴, 食事など人間の基本的な行動については, よく抽出されているといえる. これらの人間の基本的な行動の中では, 「もぐもぐ」や「二度寝」といった動詞以外の単語がキーワードとなっている場合でも抽出することができた. また, 「むくり」や「ほかいま」といった高頻度で用いられている Twitter 特有の単語がキーワードとなっている場合でも抽出することができた. これは, このような動作を表す際には, 殆どの場合において対応する単語が tweet に含まれるためである. 目的語についても真陽性の例で食事に対して「メロンパン」という食べ物の名称を抽出することができており, 正しく目的語を抽出できていると考えられる. また, 人間の基本的な行動以外の場合でも, 行動を明確に tweet の文章に示している場合には高精度で抽出できている. 従って, 人間の基本的な行動と tweet の文章に明確に示している行動については, 本手法によって抽出できた.

表 6 より, 偽陽性の結果は以下の 6 種類に分けることができると考えられる. それは, 表している行動が自分の行動でない, 表している要素が動きを伴わない, 表している行動を起こした時間が tweet の前後または tweet 中でない, 表している行動が実際に起こした行動ではなく願望を表す, 行動辞書の単語が別の言葉の一部となっている, 他者に働きかけているの 6 種類である. 1 章で述べたように, 行動は「tweet の直前, 直後または tweet 中の動きを伴う自発的な行い」と定義しているため, 表している行動が自分の行動でない, 表している要素が動きを伴わない, 表している行動を起こした時間が tweet の前後または tweet 中でない場合は行動としていない. そのため, これら 6 種類に該当する tweet は抽出されるべきではない. 表している行動が自分の行動でない場合では, 主語として人名や人を表す単語などが用いられているという特徴が多くみられる. また, 表している要素が動きを伴わない場合では, 「飽きた」のような思考を表す動詞が含まれている特徴があった. 更に, 表している行動を起こした時間が tweet の前後または tweet 中でない場合では, 時間を表す単語が含まれているまたは, 条件を満たしたら行動を行うといったような表現を用いているという特徴があった. 従って, これらの場合の tweet には, 人名や人を表す単語, 思考動詞, 時間を表す単語を集めた辞書を用意し, それらの単語が含まれた場合には抽出しないといったようなことを行うことで, 誤抽出を防ぐことができると考えられる. また, 願望を表す場合では, 助動詞の「たい」などが含まれる特徴が多くみられる. 他者に働きかけている場合では, 助動詞の「よう」などが含まれる特徴があった.従って, これらの場合の tweet には, 助動詞の「たい」や「よう」などが行動辞書の単語の直後などに現れた場合には抽出しないといったようなことを行うことで誤抽出を防ぐことができると考えられる.

更に表 7 より, 偽陰性の結果は以下の 6 種類に分けることができると考えられる. それは, 使用頻度の低い Twitter 特有や擬音などの単語がキーワードとなる, キーワードとなる単語に誤字や脱字およびカタカナ表記や「っ」を間に入れるといった表記, キーワードとなる単語が省略されている, キーワードとなる単語に動詞の活用形を用いている, キーワードとなる名詞を動詞化している, 行動を直接的に示していないの 6 種類である. これらの場合については, 現在作成した行動辞書では行動が抽出できていない. 使用頻度の低い Twitter 特有や擬音などの単語がキーワードとなる, 単語に誤字や脱字およびカタカナ表記や「っ」を間に入れるといった表記, キーワードとなる単語が省略されている場合, キーワードとなる名詞を動詞化している場合に対しては, 様々なユーザが多用した場合には n-gram により単語が得られるが, それ以外の場合では抽出することは難しいと考えられる. また, キーワードとなる単語に動詞の活用形を用いている場合には, 行動辞書の単語中の動詞に対して活用形も行動辞書の単語として加えることで抽出できると考えられる. 更に, 行動を直接的に示していない場合に

表 5 真陽性の結果の例

tweet	行動	目的語
二度寝しました	起床	―
お風呂なうなう！	行動中	―
おやすみなさいですー！	就寝	―
メロンパンもぐもぐ	食事中	メロンパン
俺はメモ帳で ホームページ作るぞ！	作成開始	ホームページ
お仕事終わりーっ！	行動終了	仕事

表6 偽陽性の結果の例

Tweet
GX 終わりなのか
課題飽きたなんて言ってられないのは わかってるけど飽きたなう
お家帰ったらお薬飲む
はーほっとけーきたべたい！
やる夫シリーズは昔から読み出すと止まらない
みんなビブリボンやろう

表 7 偽陰性の結果の例

tweet	行動
珈琲ゼリーじゅるじゅる	食事中
むぬり	起床
バイトいてきま	行動開始
パソコン買っちゃった！	購入
さて、揺れる前にシャワっとくか	入浴開始
あと 4 分でアキバ	移動中

は, 比喩表現などを集めた辞書を用いることで抽出できると考えられる.

5.3. なうに対する動詞の補完結果と考察

動詞補完の結果例を表 8 に示す. 表 8 より, 飲み物には「飲んでいる」, 食べ物には「食べている」など適切な動詞が補完され, ただ付けている場合についても補完しないといった適切な処理を行うことができた. 表 8 の 5 番目の例では, 目的語が得られていないが, これは「夕飯」が「作る」に対して係っているため, 「なう」に対して係っている名詞はなく, 動詞も補完されない. 即ち, ただ語尾につけている場合と判断している. このような場合については従来手法では考慮していないが, 本手法ではただ語尾に付けている場合を想定しているため, 適切に処理を行うことが出来る. 本手法による動詞の補完では食べ物や飲み物は, 製品名についても補完することができる. これは, 学習に tweet データを用いており, 本実験で用いた形態素解析器 JUMAN が辞書に wikipedia などに記載されている単語を加えているため, 製品名と「食べる」や「飲む」, 「する」等との共起パターンが現れることによる. 従って, 一般的な名詞や人気の製品名では動詞の補完を行うことができた. 学習するテキストデータの数を増やすことにより, 更なる精度の向上が見込まれる. また, 形態素解析の辞書に新たな単語を加えることで, より広範囲の名詞について動詞を補完できると考えられる. さらに, 現在「なう」は, 「居る」, 「している」, 「食べている」, 「飲んでいる」およびただ語尾につけているパターンの 5 パターンを想定して動詞補完を行ったが, その他のパターンについても検討し同様の手法を用いて動詞補完を行うことで, 「なう」の様々な使い方による動詞補完が可能となる.

6. 結論

行動を抽出するために, 行動辞書の作成手法を提案し, 提案手法に従って行動辞書を作成した. そして, 作成した行動辞書を用いて実際のユーザの行動を抽出し, 評価を行った. その結果, tweet から「起床」や「入浴」などの人間の基本的な行動とそのおおよその時間を抽出することができた.

また, tweet 中に明確に示されている行動についても抽出することができた. しかし, 行動辞書に含まれる単語が言葉の一部になっている場合や, 使用頻度の低い単語が含まれている場合には抽出精度が低かった. 行動抽出と同時に行った目的語の抽出に関しては, 目的語が必要な単語に対しての目的語の抽出を高精度で行うことができた. さらに Twitter 特有の表現である「なう」に対する動詞の補完ができた. 本手法では, 行動辞書の単語は人手で選択したため, 従来手法では抽出が難しいと考えられる, 「なう」や「ほかいま」といった Twitter 特有の表現をした単語を抽出することができた. また, 形態素解析に加えて行動辞書を利用して抽出を行ったため, 崩れた日本語にも対応できる場合がある. これらにより, 正しい日本語を用いているユーザだけではなく, 崩れた日本語や Twitter で頻繁に用いられる特有の単語を用いているユーザの行動も抽出できると考えられる.

今後は, 行動抽出に関して誤って抽出された場合(偽陽性)の tweet と所望の tweet が抽出されなかった場合(偽陰性)の tweet についてのデータを用いて行動辞書を改善し, 行動抽出の精度を向上させる. 本手法の精度を向上させることにより, 手軽に行動抽出を行うことが可能になり, 生活改善や行動予測に利用することができると考えられる.

表 8 補完した動詞の例

抽出前 tweet	目的語	補完した動詞
黒豆の甘酒なう	甘酒	飲んでいる
無事に風呂から離脱。アイスなう。ウマー	アイス	食べている
面談まで待機なう	待機	している
ジントニックなう	ジントニック	飲んでいる
夕飯作るなう	―	―

参 考 文 献

[1] Ig-Jae Kim, Sang Chul Ahn, Heedong Ko, Hyoung-Gon Kim, “Automatic Lifelog Media Annotation based on Heterogeneous Sensor Fusion”, Proceedings of IEEE International Conference on Multi Sensor fusion and Integration for Intelligent systems, pp703-708, 2008.

[2] 永徳真一郎, 茂木学, 望月理香, 八木貴史, 武藤伸洋, “ライフログを用いた「行動タグ」生成・利用に関する研究”, 電子情報通信学会技術研究報告, 110, pp55-60, 2011.

[3] Kiyoharu Aizawa, Gamhewage C. de Silva, Makoto Ogawa, Yohei Sato, “Food log by Snapping and Processing Images”, Virtual Systems and Multimedia, 16th, pp. 71-74, 2010.

[4] グェンミンティ, 川村隆浩, 中川博之, 田原康之, 大須賀昭彦, “条件付確率場と自己教師あり学習を用いた行動属性の自動抽出と評価”, 人工知能学会論文誌, Vol.26, No.1, pp. 166-178, 2011.

[5] 佐々木健太, 長野伸一, 長健太, 川村隆浩, “Web 上のライフストリームからのユーザ行動情報の抽出”, 第 25 回人工知能学会全国大会論文集, 25th, ROMBUNNO.3F3-4IN, 2011.

[6] Twitter, http://www.twitter.com/, 参照日：2013/03/18.

[7] 岡瑞起, 李明喜, 橋本康弘, 宇野良子, 荒牧英治, “Augmented Campus: 拡張するキャンパス”, 第 24 回人工知能学会全国大会論文集, pp. 239-242, 2010.

[8] 矢野裕司, 横井健, 橋山智訓, “行動辞書を利用した Twitter からの行動抽出”, 情報科学技術フォーラム, 11th , pp. 51-56, 2012.

[9] Twitter Search API, http://search.twitter.com/, 参照日：2013/03/18.

[10] 京都大学 黒橋河原研究室, 日本語形態素解析システム JUMAN, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN, 参照日：2013/03/18.

[11] 河原大輔, 黒橋禎夫, “自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル”, 自然言語処理, vol.14, No.4, pp. 67-81, 2007.