RO-003

行動辞書を利用した Twitter からの行動抽出

Extraction of Behavior using the Behavior Dictionary from 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抽出したものであり、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、特殊なセンサを用いたものや、手動によるメモ、写真などを用いたものが提案されている。また、ブログ等のテキストデータから行動情報を抽出する研究も行われており、ライフログに注目が集まっていることがわかる。

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[1]は近年爆発的に普及しており、膨大な投稿の中には行動を表す投稿も多く存在している。よってそれらの投稿から行動を抽出することで、従来のライフログにはない有用な情報が得ることができる可能性がある。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ライフログを取得することができる。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。しかし、tweet データは新聞記事などとは異なり、インフォーマルなデータであるため、形態素解析といった従来の解析手法だけで解析することは困難である。

そこで本研究では、個人のライフログを生成するために、Twitter における個人の tweet の集合に含まれる行動を表す単語に着目し、行動を抽出することを目的とする。なお、行動を表す単語は、一般的な動詞とサ変名詞、Twitter 特有の「なう」といった単語とする。

2. 関連研究

ライフログに関する研究としては、食事に着目した Foodlog[2]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、食事の写真を撮影、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。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。しかし、Twitter を用いてライフログの生成を行うことにより、様々な行動に関する情報を取り扱うことができる。

また、テキストデータからの行動情報の抽出に関する研究としては、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[3]が行われている。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を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”を対象としている。同様に、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間の予測を行う研究[4]も行われている。この研究ではまず、少量のテキストデータから行動情報を抽出し、行動の推移の結果数を用いて行動間の推移を学習する。そしてその結果を用いて、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。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を用いて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、形態素解析だけでは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。

本研究と同様に、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、行動を学内等小規模のマップ上に表示する研究が報告[5]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、大量の tweet データから名詞と動詞の組合せを抽出して、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。本研究においても、「なう」のように動詞が曖昧な単語に対して、同様の手法を用いて動詞の補完を行い、行動を抽出する。

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」という特例の 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」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」といった単語が挙げられる。これらの行動辞書の

† 東京都立産業技術高等専門学校
‡ 電気通信大学

単語を収集するために、Twitter でよく用いられる単語を文字 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ため、文字 n-gram が適当であると考えられる。この辞書は、Twitter から行動を抽出するための辞書であり、である。また行動辞書における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$\boldsymbol{w_n}(n = 2,3, \dots, \mathrm{N})$を取得する。ここで$\boldsymbol{w_n} = \{w_{n1}, w_{n2}, \dots, w_{nm_n}\}$であり、$w_{ni}$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$Whigh_{xy}$、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。
③w_{xi}に部分一致するw_{yj}の出現頻度とw_{xi}の出現頻度の比率r_{ij}を(1)式により求める。

$$r = \frac{freq(w_{xi})}{freq(w_{yj})} \tag{1}$$

ここで、$freq(x)$は文字列xの出現頻度である。
④rが$Whigh_{xy}$以上であれば、w_{yj}を削除する。
⑤rが$Wlow_{xy}$以下であれば、w_{xi}を削除する。
⑥③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。

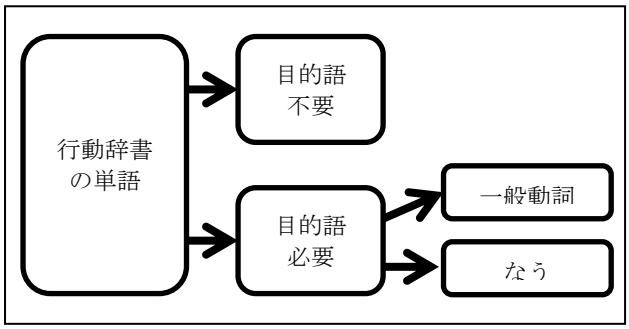

図 1 行動辞書の単語の分類

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、表 1 のように単語とその単語が示す行動、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し、その行動が示す行動は起床や帰宅等を表す。また、行動主は自(自分)、他(他人)、不(自分または他人)のいずれかを用い、目的語の必要性は単(単語単体で行動を表す)、複(目的語を必要とする)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。このラベル付けは人手で行う。

3.2. 行動の抽出

図 1 に示したように、行動辞書の単語は目的語が不要な単語、目的語が必要な一般動詞、「なう」といった特例の 3 種類に分類できるため、本研究ではそれぞれの単語に応じた方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 tweet に含まれていた場合、その行動をそのまま抽出する。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抽出する。

目的語を必要とする一般動詞の場合には、対応する目的語を抽出する必要があるため、形態素解析器および係り受け解析器を利用して tweet の係り受け関係を求める。目的語を抽出する際には、格助詞である「が」「を」「の」、および係助詞である「は」に着目し、それらの助詞に係っている言葉は抽出し、それ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」に対しては動詞を補完する必要がある場合も存在する。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、「なう」の前の単語が行動辞書の単語と一致した場合には、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

表 1 行動辞書の一部

単語	行動	行動主	目的語の必要性	時間	読み
おはよう	起床	自	単	過去	おはよう
おやすみ	就寝	自	単	過去	おやすみ
行ってきます	外出開始	自	複	未来	いってきます
おかえり	帰宅	他	単	過去	おかえり
しています	行動	不	複	現在	しています
なう	行動・場所	自	複	現在	なう
ほかった	入浴終了	自	単	過去	ほかった

場合には、文献[5]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、名詞と助詞の組合せをカウントする。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 PMI が高い尤もらしい動詞を補完する。

$$PMI(N,V_i) = log\frac{p(N,V_i)}{p(N)p(V_i)} \quad (2)$$

ここで、N は目的語となる名詞、V_iはそれぞれの動詞、$PMI(N,V_i)$はその自己相互情報量である。また、比較するうえでの式を簡単化するために、(2)式を変形し、固定する N の確率$p(N)$を無視して log を外した値で比較し、値の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)式が得られる。

$$V_N = arg\max_V p(N|V) \quad (3)$$

ここで、V_Nは目的語となる名詞 N に対して補完する動詞である。また、「なう」は進行形で用いる場合が多いため、(3)式により「する」が補完する動詞に選択された場合は、進行形の「している」を実際に補完する。

また、【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。

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。また抽出結果は、適合率、再現率と F 値で評価した。行動辞書には行動を単体で判別する単語だけでなく、目的語を必要とする単語も存在するため、その目的語の抽出結果を適合率で評価した。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、行動抽出を行った。

なお tweet の取得には、Twitter 社が提供している Twitter Search API[6]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 2011 年 10 月 14 日までの 130 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 約 200 万件に対して文字 n-gram を適用した。nは 2 から 10 まで変更して行った。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半角英数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した。rの基準値である$Whigh_{xy}$はすべて 0.95、$Wlow_{xy}$はすべて 0.01 として実験を行った。この行動辞書単独での評価は行わず、この辞書を用いた行動抽出を行うことで、辞書の評価とした。

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った。今回の実験では、無作為に選んだ 3 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 を対象とした。11 日間の 1 ユーザ当たりの tweet 件数の平均は 545 件、標準偏差は 364 件であった。正解データとして人手で抽出した tweet は 1 ユーザ当たり平均 97 件、標準偏差は 92 件であった。ユーザは無作為に選んでおり、標準偏差より tweet 頻度の高いユーザと tweet 頻度の低いユーザが選択されたことがわかる。

評価は、人手で抽出した正解データを利用して、(4)式に示す適合率および(5)式に示す再現率と、適合率および再現率の調和平均をとった(6)式に示す F 値を求め行った。

適合率 $$\text{precision} = \frac{\text{R}}{\text{N}} \quad (4)$$

再現率 $$\text{recall} = \frac{\text{R}}{\text{C}} \quad (5)$$

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

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。

4. 3. 目的語の抽出

目的語を必要とするキーワードに関して、目的語を抽出する実験を行った。今回の実験では、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、目的語を抽出した。対象 tweet は「なう」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[7]により形態素解析を行い、日本語構文・格解析システム KNP[8]により係り受け解析を行った。

評価は、抽出されるべきである目的語を人手で抽出した正解データを利用して、(4)式に示す適合率を求めて行った。

また、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されている。そこで、本実験ではこの場合の動詞補完も行う。この実験では、2011 年 6 月 7 日から 2011 年 2 月 29 日までの 268 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 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 万件を対象とした。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した。なお、名詞と動詞の組み合わせパターンとしてカウントしたのは、「名詞＋動詞」または「名詞＋助詞(複数を含む)＋動詞」の形の tweet であり、それ以外のパターンはカウントしなかった。

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、本実験で行った行動辞書を用いて行った。

5. 1. 行動辞書の作成結果

n-gram をかけて人手で抽出した結果の例として、行動を表すと考えられるキーワードを表 2 に示す。表 2 における順位と件数は、それぞれ n-gram による候補を削減した後の順位と件数である。また、表 3 にそれぞれのnの組合せに対するrの平均値を示す。表 2 より、行動を表すキーワードはある程度抽出できていると考えられる。これらのような単語を集め、行動辞書を作成した。行動辞書の単語数は 95 語であった。今後は、行動辞書中にはないこれらの

単語の連用形等を加えるとともに、n-gram の結果を再び見直し、辞書の改善を行う。また、本実験では影響は殆ど出なかったが、表 3 のようにnの組合せによりrの平均値は異なる。そこで、それぞれの比較対象 x と y ごとに基準値$Whigh_{xy}$と$Wlow_{xy}$を変更することでより良い結果が得られると考えられる。これは$n=2$ の n-gram の結果は、$n=3$の n-gram の結果と$n=4$の n-gram の結果を比較する場合、$n=4$の n-gram の結果のほうが出現率rが小さくなるためである。また、$n=2$の n-gram の結果と$n=10$の n-gram の結果は、出現率が高くともあまり関係ないため、直接比較せず、比較はnの差が 2 以内の場合に制限する等の方法も考えられる。

5. 2. 行動抽出結果

対象とした tweet の分類数を表 4、抽出結果のクロス表を表 5、行動抽出の評価結果を表 6 に示す。ここで表 4 における tweet はリプライおよびリツイートを除いた総 tweet 数、suc は正解データとして人手で抽出した tweet 数、ext は本手法により抽出した tweet 数、tp は正解データと抽出結果が一致した場合(真陽性)の tweet 数、br は tweet に対する suc の割合である。また、表 5 における行動(人手)は人手による正解データとして選択した tweet 数、行動以外(人手)は人手による正解データとして選択しなかった tweet 数、行動(抽出)および行動以外(抽出)はそれぞれ本手法により抽出した tweet 数と抽出しなかった tweet 数である。3 ユーザのユーザ名はそれぞれユーザ A、ユーザ B、ユーザ C と置き換えた。表 4 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。表 5 では、全ユーザの合計 tweet 数における人手による抽出(正解データ)と本手法による抽出の結果を示した。また、表 6 の適合率は 4 割程度と正しく抽出された割合は高くないが、割合にはユーザによりばらつきがあることがわかる。表 5 では、全ユーザの合計 tweet 数における人手による抽出(正再現率は 5 割程度抽出できていることがわかる。行動抽出では、抽出漏れ無く行動を抽出できるように、出来る限り再現率が高いことが望まれると考えられるため、今後は F 値と再現率の上昇を目指す。更に、真陽性の結果の例を表 7、偽陽性の結果の例を表 8、偽陰性の結果の例を表 9 に示す。表 7 より就寝や起床、入浴等の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。しかし、単語が含まれていても、形態素解析を利用した場合では意図していない場所で区切られることがあるため、従来手法より本手法のほうが取得できると考えられる。また、表 8 より「しよう」や「やめ

表 2 n-gram をかけて人手で抽出した結果

順位	単語	件数
22	います	155,466
200	しました	46,417
234	いました	41,885
250	おやす	40,469
263	おはよう	39,357
318	おやすみ	32,524
403	おはようございます	26,689
411	なう	26,397
485	おかえり	24,050
1,162	風呂	12,445
4,069	ほかえり	3,534

表 4 それぞれのユーザにおける結果

	tweet	suc	ext	tp	br
ユーザ A	962	203	199	91	0.211
ユーザ B	290	32	47	16	0.110
ユーザ C	384	57	88	32	0.148
平均	545	97	111	46	0.178

表 5 抽出結果のクロス表

	行動(人手)	行動以外(人手)	合計
行動(抽出)	139	195	334
行動以外(抽出)	153	1149	1302
合計	292	1344	1636

表 6 評価結果

	適合率	再現率	F 値
ユーザ A	0.457	0.448	0.453
ユーザ B	0.340	0.500	0.405
ユーザ C	0.364	0.561	0.441
平均	0.416	0.476	0.444

表 3 それぞれのnの組合せにおける r の平均値

	2	3	4	5	6	7	8	9	10
2		0.233	0.179	0.150	0.135	0.124	0.116	0.110	0.105
3			0.527	0.394	0.348	0.319	0.305	0.292	0.285
4				0.704	0.617	0.565	0.537	0.520	0.504
5					0.828	0.757	0.718	0.694	0.676
6						0.886	0.839	0.809	0.784
7							0.924	0.885	0.860
8								0.943	0.911
9									0.953
10									

表 7 真陽性の結果の例

対象 tweet	行動
おやすみなさいー！	就寝
起きたーっ！おはようございますっ！	起床
お風呂入ろうっと	入浴
それではお仕事行ってきますーっ！	外出
ご飯食べよっと！	食事

表 8 偽陽性の結果の例

対象 tweet
年賀状どうしよう
作業の途中でニコ動観だすのやめたい
うおっれきさん潜伏してたの
何やってるんだろう私
アイコンは自分で描きました！見て下さいこの質素な絵！

表 9 偽陰性の結果の例

対象 tweet	行動
17 時間ぶりくらいの浮上です (´・ω・`)	Twitter 開始
みそ汁うまーい	食事
ブログ更新完了	ブログ執筆
今はふぁぼ規制です	Twitter 特有
フォロー完了	Twitter 特有

た」等の行動辞書に含まれる単語が、他の言葉の一部になっているものや、「描きました」といった長い時間経過していると考えられる過去のことを表した tweet が抽出されている。しかし、表 8 に示すような例は行動を表していない、または最近の行動ではないため、抽出されるべきではない。そこでそれを改善する処理を行う必要がある。また、表 9 より浮上等の比喩表現を用いている、ふぁぼ等の Twitter 特有の行動である、および単語の間に「ー」等の文字が含まれている場合には、現在作成した行動辞書では行動抽出が取得できていなかった。この改善策としては、比喩表現に関する辞書を用いて単語に対応する比喩表現も対象に含む、n-gram の結果を再び見直し今回判断しなかったその他の単語も行動辞書に加える、「ー」等の文字が含まれていても一致するといったことが挙げられる。これらのことを行い、更なる再現率の向上を行う必要がある。「ー」等の文字に対する改善策としては、「ー」を無いものとして扱うことおよび前の文字の母音と同じ文字を補完することが考えらえる。この時、前の文字の母音が「お」であった場合は「う」の補完も行い、一致を調べ抽出する。

5. 3. 目的語抽出結果

目的語を抽出した結果、適合率は 0.83 であった。抽出した目的語の例を表 10 に示す。適合率および表 10 より、目的語の抽出は高精度で行えているといえる。一般的には Twitter のようなインフォーマルな文書には形態素解析は向かないといわれているが、今回の結果より係り受け関係を解析するうえでは、やや高精度で係り受け関係を抽出できているといえる。また、着目する助詞を変更することや、助詞だけではなく連用形＋”タ”および連用形＋”ダ”も考慮して行うことにより、更なる適合率の増加が出来ると考えられる。

また、動詞補完の結果を表 11 に示す。表 11 より、飲み物には「飲んでいる」、食べ物には「食べている」等適切な動詞が補完されている。表 11 の 3 番目の例では、”作る”までが目的語となっているが、これは「なう」に対する目的語であるため、正しく抽出出来ている。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が、あまり人気でないものの場合や、新製品の場合には補完することが難しい。これは、大量のテキストデータの中でも出現数が少なく、更に「食べる」や「する」等との共起パターンが現れないこともあるからである。しかし、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。よって、動詞補完には適切な手法であると考えられる。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。

表 10 抽出した目的語の例

抽出前 tweet	抽出した目的語
授業なうだけど、通信科目でみんなレポートなうだから静かすぎるｗｗ	・授業 ・みんなレポート
腰骨にカバン当たって激痛なう。	・激痛
電車なう眠い(・ω・)	・電車
面談まで待機なう	・待機
ジントニックなう	・ジントニック

表 11 補完した動詞の例

抽出前 tweet	抽出した目的語	補完した動詞
黒豆の甘酒なう	・黒豆の甘酒	・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
夕飯作るなう	・夕飯作る	補完なし
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

6. まとめ

行動を抽出するために、行動辞書作成手法を提案し作成した。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た。その結果、tweet から「起床」や「入浴」等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と「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 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の tweet と所望の tweet が抽出されなかった場合の 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。そして、行動抽出と同時に必要に応じて目的語の抽出と動詞の補完を行い、行動抽出の際に目的語を付与することも行う。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ため、精度の向上を目指す。

また近年では Facebook[9]や mixi[10]など様々なサービ

スを同時に使用している人も少なくない。そのため、同一ユーザのそれぞれでの投稿を組み合わせることで、より詳細な行動を抽出することが出来るため、この課題についても検討する必要がある。

参考文献

[1] Twitter, http://www.twitter.com/, 参照日：2012/03/28.

[2] Kiyoharu Aizawa, Gamhewage C. de Silva Makoto Ogawa, Yohei Sato, “Food log by Snapping and Processing Images”, Int Conf.Virtual Systems and Multimedia, pp71-74, 2010.

[3] 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.

[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, pp166-178, 2011.

[5] 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”, The 24th Annual Conference of the Japanese Society for Artificial Intelligence, pp239-242, 2010.

[6] Twitter Search API, http://search.twitter.com/, 参照日：2012/03/28.

[7] 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：2012/03/28.

[8] 京都大学 黒橋・河原研究室, 日本語構文・格解析システム KNP, http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP, 参照日：2012/03/28.

[9] Facebook, http://www.facebook.com/, 参照日：2012/03/28.

[10] mixi, http://mixi.jp, 参照日：2012/03/28.