一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NLC2020-3(2020-09)

人物の属性を区別した定期的事象の獲得

山元　航平†　　嶋田　和孝††

† 九州工業大学大学院　情報工学府　〒820–8502 福岡県飯塚市川津 680–4
†† 九州工業大学大学院　情報工学研究院　〒820–8502 福岡県飯塚市川津 680–4
E-mail: †{k_yamamoto,shimada}@pluto.ai.kyutech.ac.jp

あらまし　本研究では，特定の時期，時間帯に発生する傾向がある人間の行動である「定期的事象」のうち，ある人物属性において特徴的なものの獲得を目指す．近年，知識獲得の研究が盛んであり，人間の行動や時間的常識知識などに関する研究が行われている．本研究で獲得を行う定期的事象も時間的な知識の中の 1 つである．人物属性ごとの定期的事象の獲得には，人物属性ラベル付きのデータセットが必要となるが，作成のコストは高い．そこで広く使用されている SNS である Twitter ユーザのデータに対して一定量のアノテーションを行い，それに基づいて人物属性分類器を作成する．この分類器を用いて人物属性ラベルの付与された大量の Tweet データセットを自動的に構築する．構築したデータセットをもとに，人物属性ごとの事象表現の抽出や解析を行うことで，各人物属性に特有の定期的事象を収集する．実験結果より，"学生は 13 時に大学に来る"や"社会人は平日に残業から帰る"などの事象知識の獲得に成功した．
キーワード　知識獲得，テキストマイニング，常識的知識

Acquisition of periodic events with person attributes distinction

Kohei YAMAMOTO† and Kazutaka SHIMADA††

† Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology
680–4 Kawadu, Iizuka-shi, Fukuoka, 820–8502 Japan
†† Department of Artificial Intelligence, Kyushu Institute of Technology
680–4 Kawadu, Iizuka-shi, Fukuoka, 820–8502 Japan
E-mail: †{k_yamamoto,shimada}@pluto.ai.kyutech.ac.jp

Abstract　In this paper, we aim to acquire "periodic events" that repersent siginificant actions happened by groups of people in particular seasons or timing. Recently, important of knowledge acquisition is increasing. Many studies about human action knowledge and temporal commonsense knowledge have been carrying out. We need a dataset to acquire periodic events. However, manually building the dataset with human attribute labels is expensive. Therefore, we construct a human attributes classifier of Twitter users and create large labeled tweets dataset automatically. Periodic events with specific human attributes are collected with our proposed method. Finally, we succeeded at acquisition of events knowledge: for example, "Students go to colledge at 1pm." and "Workers finish their overtime work and go home on weekdays."
Key words　Knowledge Acquisition, Text Mining, Commonsense Knowledge

1. はじめに

近年，知識獲得の研究が盛んに行われている．獲得対象となっている知識は，数値の大小感覚の知識 [1] や一般的に物理的距離の近いオブジェクトペアの知識 [2] など様々である．大規模な知識ベースも複数提案されており [3] [4]，常識的知識が質問応答 [5] や非タスク指向型対話 [6]，複数の言語処理ベンチマークタスク [7] において有効であることが報告されている．

我々はこれまで，人間の行動を中心として，自然現象，社会現象など（本研究ではまとめて「事象」と呼ぶ）のうち，特定の時期や時間帯に発生する傾向のあるもの（本研究では「定期的事象」と呼ぶ）についての知識に着目をしてきた [8] [9]．この知識は，{事象, 時期・時間帯}のように，事象と時期・時間帯の 2 つ組として表記する．知識の具体例としては，{雪が降る, 冬}，{湯船につかる, 夜}などが挙げられる．以前の研究 [9] で我々は，定期的事象のうち，学生や社会人など特定の人物属

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2020 by IEICE

性に関連のあるものの獲得を試みた．具体的には，Twitter[1]ユーザを Tweet 投稿時間分布でクラスタリングすることで人物属性ごとのクラスタを作成した．その後，作成したクラスタごとに Tweet からの知識抽出を行い，目的とする知識の獲得を行った．結果として，クラスタリングによって作成された学生ユーザや主婦ユーザのクラスタから，人物属性ごとの定期的事象知識を複数獲得した．

しかし，いくつかの問題も発生した．最も大きい問題が，作成したクラスタへの評価の難しさである．学生ユーザや主婦ユーザなど，分かりやすいユーザが大半を占めるクラスタも存在したが，そうでないクラスタも多く，結果的にどのような属性の人物に紐づくのか不明な定期的事象を大量に獲得することとなった．あらかじめ人物属性の定義を行い，その定義に沿う人物であるか判定してデータセットの分割を行うことができれば，上述の問題を回避でき，より厳密に社会を反映した知識の獲得につながることが期待できる．

そこで本研究では，ユーザの人物属性による Tweet データセット分割手法の改良を行うことで，人物属性ごとのより高品質な定期的事象知識の獲得を試みる．具体的には，自然言語事前学習モデルである BERT [10] を使用して Twitter ユーザ人物属性分類モデルを作成し，作成した分類モデルを使用して Tweet データセットを人物属性ごとに分割する．分割した Tweet データセットから．目的である人物属性ごとの定期的事象知識の獲得を行う．

2. 関 連 研 究

本研究では，人物の属性ごと特徴的な定期的事象知識を獲得することを目的としている．Ge らは Wikipedia データを利用して「Event」の知識の収集を行った [11]．しかし，ここで Ge らが獲得の対象としている「Event」とは，事件，事故，災害などの突発的な出来事やオリンピックなどの大規模な行事などであった．社会的な影響の大きな出来事や行事についての知識は重要なものであると思われるが，特定の時期や時間帯に関連するものではないことが多いであろう．本研究では，社会的影響の大きな出来事や行事と同様に人々の身の回りの出来事についての知識にも価値があると考え，人間の行動を中心とした事象の獲得を行う．人間の行動と時間情報に着目した知識獲得を行っている論文としては，Tandon らの研究 [12] と Yao らの研究 [13] が挙げられる．Tandon らはドラマ，映画，小説などの物語作品から人物の行動知識の獲得を行い，その際，行動時の時間などの情報も抽出をした．しかし，抽出の対象とした時間情報は昼，夜などで種類が少なく，それぞれの定義も曖昧であった．本研究では，平日・週末から 1 時間までの幅広い粒度の時間区分を設定する．Yao らは Web テキストから動詞およびその動作主を抽出することで，近しい行動のペアなど，行動同士の関係知識の獲得を行った．時間の情報も抽出しているが行動間の相対的な時間情報であり，事象どうしの関係知識に近いものであるといえる．さらに，Tandon らの研究と Yao らの研究のいずれにおいても行動を行った人物の属性については考慮されていない．本研究では，人物の属性の沿った知識の獲得を目指す．また，幅広い時間的常識知識に着目した論文としては Zhou らの研究 [14] がある．Zhou らは事象の長さや頻度，発生するタイミングなどをモデルに学習させ，時間的常識知識の必要な複数のタスクにおいて BERT を上回る精度を達成した．我々の研究で扱う定期的事象についての知識は Zhou らの定義した複数の時間常識的知識の一部分であるが，Zhou らの研究は時間的常識能力のあるモデルの作成であり，我々が目指す構造化された知識の獲得とは目的が異なる．

3. データセット

提案手法による定期的事象の獲得の前に，ランダムな Twitter ユーザデータセット（3.1 節）と，人物属性ラベル付き Twitter ユーザデータセット（3.2 節）の計 2 種類のデータセットを作成した．それぞれについて詳しく説明する．

3.1 ランダム Twitter ユーザデータセット

ランダムな Twitter ユーザのデータセットである．収集には，Twitter 社の提供する Twitter API[2]を用いた．Twitter の日本語話者ユーザのうち，自己紹介欄にテキストが存在するユーザの Tweet 投稿履歴を遡る方法で Tweet を獲得した．獲得した時系列順 Tweet データに，ユーザ名と自己紹介欄のテキストを加えて 1 ユーザ分のデータとした．なお，獲得できた Tweet 数が 100 件以下か，獲得できた Tweet の期間（最も古い Tweet から最新の Tweet までの期間）が 1 年以下のユーザは除外した．本研究では，収集したデータのうち 20 万ユーザ分のデータ（計 5500 万 Tweet）を使用した．

3.2 ラベル付き Twitter ユーザデータセット

人手で作成した，人物属性のラベル付きの Twitter ユーザデータセットである．人物属性ラベルに加え，ランダム Twitter ユーザデータセット（3.1 節）と同様に時系列順の Tweet データ，ユーザ名，自己紹介欄のデータを持つ．人物属性ラベルは，「学生」「社会人」「子育て世代」のいずれかである．それぞれの人物属性の定義を以下に示す．

学生

- 小学校，中学校，高等学校，専門学校，大学，大学院および，それに準ずる教育期間に学習者の立場で在籍している．
- 休学中でない．

社会人

- 学生でない．
- 何らかの手段で所得を得ているか，専業主ふである．

子育て世代

- 義務教育以下の年齢の子供を育てている．

（注1）：https://twitter.com

（注2）：https://developer.twitter.com/en/docs/twitter-api

表 1　各人物属性の自己紹介欄で使用されやすいキーワード

人物属性	キーワード
学生	学生, 大学, 短大, 専門学校, 院生, 学部, 学科
社会人	社会人, 会社, 企業, 勤務, 業務
子育て世代	育児, 子育て, 息子, 娘, の父, の母, のパパ, のママ

表 2　人物属性ラベル付きユーザ数と Tweet 数

人物属性	ユーザ数	Tweet 数
学生	1,032	100 万
社会人	1,216	120 万
子育て世代	1,182	117 万

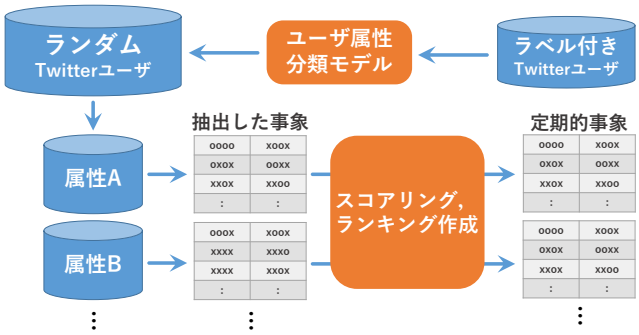

図 1　手法概要図

上記の人物属性に対し，その人物属性の Twitter ユーザが自己紹介欄で使用しているキーワードを人手で収集した．収集したキーワードを表 1 に示す．ランダム Twitter ユーザデータセット（3.1 節）のうち，表 1 のキーワードを自己紹介欄に含むユーザに対して，その人物属性のユーザであるかアノテーションを行い，本データセットを作成した．本データセットに含まれる，各人物属性とユーザ数，データ数を表 2 に示す．

4. 提 案 手 法

本研究では，人物属性に特有の定期的事象知識の獲得を目指す．まず，Twitter ユーザの人物属性分類モデルを作成し，人物属性が未知のユーザに対して属性分類を行うことで，人物属性ごとのユーザデータセットを作成する（4.1 節）．次にユーザデータセットごとに Tweet 中の事象の抽出を行う（4.2 節）．最後に，抽出した事象へのスコアリングおよびランキングの作成をし，人物属性ごとに特徴的な定期的事象知識を獲得する（4.3 節）．提案手法の概要を図 1 に示す．以下，各手順について説明する．

4.1　Twitter ユーザ人物属性分類モデルの作成

人物属性に特有の定期的事象を獲得するためには，事象の抽出元となる人物属性ラベルつきデータセットが必要である．知識獲得などのタスクでは使用するデータセットの規模が重要であるが，人手での大規模データセットの作成はコストが高い．そこで人手で作成した小規模なデータセットから Twitter ユーザの人物属性分類モデルを作成し，低コストでのラベル付きデータセットの拡張を行う．こうして作成したデータセットを後述の処理（4.2 節，4.3 節）で用いる．

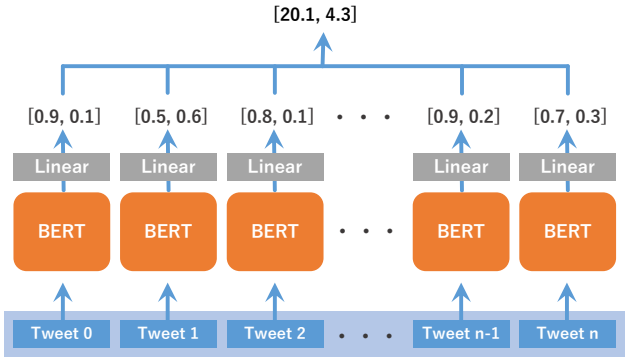

図 2　人物属性分類モデル概要図

使用する Twitter ユーザ人物属性分類モデルの説明を行う．まず BERT [10] の事前学習済みモデルに linear 層を追加し，Tweet 単位での One-vs-Rest の人物属性ラベル分類タスクにファインチューニングする．次に，ファインチューニングを行ったモデルを使い，あるユーザの投稿した Tweet に対して 1Tweet での属性分類をそれぞれ行う．1 ユーザ分の Tweet の属性分類の出力結果を足し合わせた値を，ユーザに対する人物属性推定値とする．モデルの概要図を図 2 に示す．

4.2　Tweet 中の事象の抽出

人間の行動をはじめとして，様々な事象は動詞や名詞と動詞のペアで表現されることが多い．そこでまず Tweet 中から動詞を抽出する．さらに，抽出した動詞と係り受けの関係にある名詞がある場合には，動詞単体とは別に名詞 + 動詞のペアを抽出する．本研究では，抽出を行う動詞及び動詞 + 名詞のペアを事象語と呼ぶ．なお，事象語の抽出の際にはモダリティ判定を同時に行い，現実でない事柄について述べるモダリティの文脈で使用されている動詞は除外する．具体的には，希望のモダリティを持つ文「映画を見たい.」中の「見たい」などである．このような語を除外するのは，本研究では実際に起こった事象を抽出対象としているためである．

4.3　事象へのスコアリング，ランキング作成

抽出した各事象語について，平日・休日，曜日，朝昼晩，24 時間の計 36 個の時間区分における頻度を算出する．各時間区分の頻度の，その事象語の総頻度に対する割合を，その事象語のその時間区分におけるスコアとする．最終的に，スコアをもとに人物属性ごとに各時間区分での事象語のランキングを作成し，上位の事象語を特定の人物属性における定期的事象を表す事象語として獲得する．なお，以下のいずれかの条件に当てはまる事象語はランキングから除外する．

（1） 設定した最低頻度より総頻度が少ない．
（2） ごく少数のユーザによって集中的に使用されている．
（3） 365 日の各日のスコア（総頻度に対するその日の頻度の割合）の分散が高い．

表 3　Twiiter ユーザ人物属性推定のファインチューニング結果

人物属性	Precision	Recall	F1
学生	0.54	1.00	0.71
Other	1.00	0.37	0.54
社会人	0.47	1.00	0.64
Other	1.00	0.24	0.39
子育て世代	0.62	1.00	0.76
Other	1.00	0.69	0.71

（4） 2 種以上の人物属性のランキング上位に存在する．

条件 1 を満たす事象語は，重要度の低いものであると推測できるため結果から除外する．条件 2 を満たす事象語は，一般性に乏しいものであると推測できるため除外を行う．条件 3 を満たす事象語は，いわゆる「バースト」的な事象に関連するものであると推測できる．例えば，2019 年の秋ごろにラグビーのワールドカップが日本で開催され大きな注目を集めた．日本代表の試合が 19 時に開始したとすると，その前後の時間にはラグビーについての Tweet が急増したはずであり，ラグビー関連の事象語が多く抽出されることとなる．しかし，ラグビー関連の事象は恒久的に 19 時に発生するものではなく，19 時の定期的事象としては適切でない．よって「バースト」的な事象語については上記の条件を用いて除外を行う．条件 4 を満たす事象語は，定期的事象ではあるが特定の人物属性における定期的事象に関連する事象語であるとは言えない．本研究では人物属性に特有の定期的事象の獲得を行うため，該当する事象語はランキング中から除外する．

5. 実　　験

本研究では，3.2 節で定義した「学生」「社会人」「子育て世代」の 3 種類の人物属性を解析対象の属性とする．この 3 種類を対象としたのは，社会的に属する人物の多い属性であり，属する人物の行動習慣に大きく影響を与える属性であると考えたためである．以下，行った実験について詳しく説明を行う．

5.1 Twitter ユーザ人物属性推定タスク

東北大学が公開している BERT 日本語事前学習モデルのうち，BERT-Base の whole-word-masking モデル[3]を使用した．ファインチューニングは，バッチサイズを 32，最適化関数を Adam，学習率を $2e-5$，エポック数を 3 に設定した．また正例には解析対象とする人物属性データ（表 2 参照）の Tweet をそれぞれ使用し，負例にはランダムユーザデータセット（3.1 節）中の Tweet を同数使用した．学習結果を表 3 に示す．さらに，本モデルを用いて人物属性ラベルが未知のユーザの属性分類を行い，ラベル付き Twitter ユーザデータセットを新たに作成した．データセットの詳細を表 4 に示す．

（注3）：https://github.com/cl-tohoku/bert-japanese

表 4　自動生成した人物属性ラベル付き Twitter ユーザデータセット

人物属性	ユーザ数	Tweet 数
学生	32,404	1,726 万
社会人	29,697	1,411 万
子育て世代	19,897	791 万

5.2 事象語ランキング作成

表 4 を対象に事象語のランキングを作成した．なお 4.3 節に示した条件 1 の最低頻度は 50，条件 2 は抽出元 Tweet の 2 割以上を占めるユーザがいるか，条件 3 の分散は $5e-5$，条件 4 の順位は上位 100 位までに設定した．紙面の都合上，作成した全てのランキングを示すことはできないが，学生の 24 時間の事象語ランキングを表 8 に，3 種類の人物属性の平日，休日ランキングをまとめたものを表 5 に，曜日のランキングのうち火曜日，水曜日，木曜日の結果をまとめたものを表 6 に，朝昼晩をまとめたものを表 7 に示す．なお，いずれのランキングも上位 5 件の結果である．

6. 考　　察

作成した Twitter ユーザ人物属性推定モデルの学習結果と，作成した事象語ランキングの結果について考察を行う．

6.1 Twitter ユーザ人物属性推定の結果について

作成した Twitter ユーザ人物属性分類器は，正例に学習の対象とする人物属性（学生，社会人，子育て世代のいずれか）のユーザの Tweet，負例にランダムなユーザの Tweet を入力してファインチューニングを行った．負例として与えているユーザデータはランダムユーザの Tweet であるため，当然，正例側のユーザの人物属性のユーザの Tweet も含まれる．学生ユーザの学習の際には，直感的には，負例として与えたデータの 1/3 から 1/2 は学生ユーザの Tweet であったであろう．そのため，本モデルのファインチューニングの分類結果の正例側の Precision は 1 になり得ない．表 3 が示すように，各人物属性の Precision は 0.5 前後となっている．しかし，いずれの属性においても Recall は 1 と非常に高い値になっている．このように，ラベル付きデータの取りこぼしをほとんどしていないという結果となっていることが示されている．本モデルはラベル付きデータセットの構築を目的に使用するものであるため，Precision より Recall のほうが重要である．このように，本研究で使用する属性推定モデルとして好ましい学習ができていると言える．

6.2 作成したランキングについて

全体として，学生ユーザ，社会人ユーザに関しては，学生の定期的事象，社会人の定期的事象として妥当性を感じる事象語が多く獲得できた．しかし，子育て世代ユーザについては評価の難しい事象語がほとんどであった．これは子育て世代の人物同士の行動傾向の類似性が，学生・社会人の属性の人物同士の行動傾向の類似性に比べて低いためであると考えられる．世の中

表 5　平日・週末のランキング

人物属性	時間区分	事象語				
学生	平日	授業 + 出る	限 + する	学校 + 休む	授業 + 終わる	ニュース + やる
	休日	ラスト + する	街 + 行く	時間 + 並ぶ	今 + 向かう	バス + 行く
社会人	平日	残業 + なる	残業 + 帰る	毎日 + ある	発送 + くる	予約 + 取れる
	休日	撤収 + する	部 + 終わる	高速 + 乗る	会場 + 向かう	MC+ する
子育て世代	平日	会社 + 休む	タオル + する	感 + 持つ	ー + 読む	の + 解る
	休日	久しぶり + 寝る	最高 + 過ぎる	勝利 + する	いこう	参加 + 出来る

表 6　曜日のランキング　火水木の結果のみ

人物属性	時間区分	事象語				
学生	火	弱体 + する	牛 + 食べる	原因 + わかる	ストレス + 感じる	キュンキュン + する
	水	時 + 出す	授業 + 出る	日 + 続ける	プロテイン + 飲む	重要 + 思う
	木	好感 + 持てる	誰 + 持つ	慕う	弾 + する	メン + する
社会人	火	ん + みる	鍛刀 + する	前 + 描く	拾える	刀剣 + する
	水	円 + できる	私 + 入れる	詐欺 + する	意思 + する	強がる
	木	にらめっこ + する	料 + する	アプリ + なる	こと + 始める	動画 + あげる
子育て世代	火	の + 入れる	さん + 居る	の + 迷う	抱きつく	こと + 頑張る
	水	仕事 + 考える	とむ	動画 + みる	血 + 出る	側 + なる
	木	誰 + いる	打ち込む	最近 + 言う	ん + 探す	プラス + する

表 7　朝昼晩のランキング

人物属性	時間区分	事象語				
学生	朝	電車 + 空く	朝 + ある	出社 + する	おきる	電車 + 止まる
	昼	昼飯 + 食う	大学 + 来る	授業 + 出る	今 + いく	駅 + いる
	晩	明日 + 飲む	見入る	明日 + いる	ちゃん + 泣く	久しぶり + 観る
社会人	朝	今日 + 乗り切る	今日 + 頑張れる	朝 + 言う	早朝 + する	冷え込む
	昼	午後 + 頑張る	撤収 + する	買い出し + 行く	会場 + 向かう	お昼 + なる
	晩	付け + 変わる	ため + 寝る	明日 + 観る	日本 + 飲む	BS+ やる
子育て世代	朝	乗り遅れる	浮腫む	用いる	今日 + 楽しむ	社会 + する
	昼	駐車 + する	人 + 似る	明け + する	の + 死ぬ	問い合わせ + する
	晩	バイト + 終わる	申し込み + する	夢 + 見れる	明日 + 備える	エモ + いる

には働きながら子育てをしている人物もいれば，専業主ふとして子育てをしている人物もおり，生活のパターンは大きく異なるであろう．また人物の属性にかかわらずノイズの多い結果となった．事象語の抽出手法やスコアリング手法の改良，Twitter 特有の報告バイアスの考慮などを行ってノイズを取り除き，より高品質なランキングの作成を目指す必要がある．作成した時間区分別のランキング結果については，以下に詳述する．

6.2.1　平日・週末の事象語ランキング

学生の結果には，平日の"授業 + 出る"，"学校 + 休む"，"授業 + 終わる"など学校関連の事象語が獲得できており，学生らしさのある結果となっている．社会人も"残業 + なる"，"残業 + 帰る"など業務関連の事象語が獲得できている．しかし，子育て世代に関しては育児などに関連を感じる事象は獲得できなかった．また学生ユーザと社会人ユーザに関しては，休日にのみ行楽やレジャーに関する事象語が含まれており，平日と休日での行動の差異の抽出にも成功していると言える．

6.2.2　曜日の事象語ランキング

曜日のランキングに関しては，いずれの人物属性においてもその曜日らしさのある事象語は獲得できていない．曜日ごとの行動の差異が平日・休日と比較すると小さいため，本質的に難しい問題になっていると考えられる．

6.2.3　朝昼晩の事象語ランキング

朝昼晩のランキングの結果も表 5 の平日・週末のランキングと傾向が似ており，学生ユーザに関しては大学に関する事象語，社会人に関しては日々の業務に関する事象語が獲得できているが，子育て世代に関しては育児や日々の生活に関する事象語は獲得できなかった．

6.2.4　24 時間の事象語ランキング

全体として各時間らしさのある事象語の獲得ができている．学生ユーザ（表 8）については，16 時の 1 位の事象語である"授業 + 終わる"をはじめとした学校関連の事象語が存在する．また，論文中には掲載していないものの，社会人の結果には 12 時，13 時の 1 位である"午後 + 頑張る"や 7 時の 1 位である"今日 + 頑張れる"，8 時の 1 位である"今日 + 乗り切る"などの，自らを鼓舞する内容の Tweet 中に使用されていたであろう事象語が獲得された．子育て世代の結果に関しては，他の時間区分でのランキングと同様，評価の難しい事象語が多い結果であった．

表 8　24 時間のランキング　学生の結果

時間区分	事象語				
0 時	バイト + 帰る	そこ + 言う	笑い + 止まる	風呂 + 寝る	日 + 祝う
1 時	一番 + 思う	れんく + なる	ねれる	気 + なれる	ヲタク + する
2 時	みんな + 寝る	俺 + 寝る	悶絶 + する	夕方 + 寝る	根 + 持つ
3 時	ん + 寝る	目 + 覚ます	れんく + なる	夢 + 見れる	覚ます
4 時	こと + 寝る	なつく	鼻 + 詰まる	直面 + する	ふえる
5 時	昼夜 + する	時半 + 寝る	普通 + 寝る	ころす	朝 + 来る
6 時	時半 + 寝る	返事 + なる	日 + がんばる	夜勤 + 終わる	帰宅 + 寝る
7 時	電車 + 空く	通勤 + する	日 + 頑張る	ホン + 忘れる	仕事 + がんばる
8 時	電車 + 空く	分 + 遅れる	揺る	朝 + 迎える	尊敬 + できる
9 時	前 + 着く	分 + 着く	昨日 + 終わる	限 + する	バイト + がんばる
10 時	限 + する	限 + ある	大学 + 来る	めくれる	家 + 置く
11 時	色 + 変わる	てらう	絶滅 + する	大学 + 来る	ボス + する
12 時	昼飯 + 食う	かき集める	プール + 行く	書類 + する	払える
13 時	大学 + 来る	影響 + 出る	昼飯 + 食う	系 + ある	方 + 楽しむ
14 時	今日 + 着る	お茶 + 飲む	推奨 + する	洗濯 + 終わる	ネット + ある
15 時	買い + 来る	散々 + する	今更 + する	ドア + する	チャリ + 乗る
16 時	授業 + 終わる	買い + 来る	設ける	もやう	目 + 死ぬ
17 時	退社 + する	ついで + する	ワーク + する	込み上げる	出現 + する
18 時	日 + 当たる	意味 + わく	同意 + する	事務所 + する	ホームラン + 打つ
19 時	ゲー + なる	引き上げる	今日 + 走る	中心 + する	シーズン + する
20 時	時代 + 来る	私 + わかる	要素 + する	いっす	の + 許す
21 時	アルバム + 買う	こめる	人 + 終わる	最初 + 見る	年 + とる
22 時	お送り + する	わたし + いる	お手伝い + する	髪 + 乾かす	定時 + 帰る
23 時	館 + 見る	明日 + いる	牛乳 + 飲む	運命 + 感じる	ちゃん + すぎる

7. おわりに

本研究では学生，社会人，子育て世代のそれぞれの属性において特徴的な定期的事象知識の獲得を行った．まず知識抽出元となる人物属性ラベル付きの Tweet データセットを作成するため，BERT 日本語事前学習モデルを使用して Twitter ユーザの人物属性分類器を作成した．作成した分類器でアノテーションのなされていない Twitter ユーザを分類し，人物属性ラベルを付与することで，1000 万件規模のラベル付き Tweet データセットを構築した．このデータセットを元データとして，Tweet 中に含まれる事象語の抽出を各人物属性ごとに行った．抽出した事象語のうち，本論文が獲得を目指す知識としてふさわしくないものを複数の条件で除外し，残った事象語に対してスコアリングを行った．最終的に時間区分ごとのランキングをスコアを作成することで，各人物属性特有の定期的事象知識の獲得に成功した．今後は，Twitter ユーザ人物属性分類手法の改良，獲得した定期的事象知識評価手法の検討などを行う．

文　献

[1] Katsuma Narisawa, Yotaro Watanabe, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. Is a 204 cm man tall or small? acquisition of numerical common sense from the web. In *Proceedings of ACL*, pp. 382–391, 2013.

[2] Frank F Xu, Bill Yuchen Lin, and Kenny Zhu. Automatic extraction of commonsense locatednear knowledge. In *Proceedings of ACL*, pp. 96–101, 2018.

[3] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, pp. 4444–4451, 2017.

[4] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. In *Semantic Web Journal*, pp. 167–195, 2015.

[5] Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Improving question answering by commonsense-based pre-training. In *Proceedings of NLPCC*, pp. 16–28. Springer, 2019.

[6] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of ICLR*, 2019.

[7] Zhang Zhengyan, Han Xu, Liu Zhiyuan, Jiang Xin, Sun Maosong, and Liu Qun. Ernie: Enhanced language representation with informative entities. In *Proceedings of ACL*, pp. 1441–1451, 2019.

[8] Kohei Yamamoto and Kazutaka Shimada. Acquisition of knowledge with time information from twitter. In *Proceedings of IALP*, 2019.

[9] 山元航平, 嶋田和孝. 人物の属性に着目した時期・時間帯に関連のある事象の獲得. 言語処理学会第 26 回年次大会 発表論文集, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[11] Tao Ge, Lei Cui, Baobao Chang, Zhifang Sui, Furu Wei, and Ming Zhou. Eventwiki: a knowledge base of major events. In *Proceedings of LREC*, pp. 499–503, 2018.

[12] Niket Tandon, Gerard De Melo, Abir De, and Gerhard Weikum. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings of CIKM*, pp. 223–232, 2015.

[13] Wenlin Yao and Ruihong Huang. Temporal event knowledge acquisition via identifying narratives. In *Proceedings of ACL*, pp. 537–547, 2018.

[14] Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. Temporal common sense acquisition with minimal supervision. In *Proceedings of ACL*, pp. 7579–7589, 2020.