人物の属性を考慮した周期的事象の獲得

山元航平
九州工業大学大学院 情報工学府
k_yamamoto@pluto.ai.kyutech.ac.jp

嶋田和孝
九州工業大学 情報工学部
shimada@pluto.ai.kyutech.ac.jp

1 はじめに

昨今，常識的知識獲得 [1] や常識的知識推論 [2] を始めとして，常識的知識についての研究が盛んに行われている．大規模な知識ベースも多数提案されており [2, 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．

我々は以前の研究 [8] において，学生や社会人などの人物属性を定義し，これらの人物属性に強い関連を持つ周期的事象 (Periodic Event) の獲得を試みた．周期的事象とは，人間の行動などの事象の中でも，特定の時期や時間帯に発生する傾向のある事象のことである．周期的事象は{“事象”, “時期・時間帯”}のように事象と時期・時間帯の 2 つ組で表記する．具体例としては{雪が降る, 冬}や{湯船につかる, 夜}などが挙げられる．この先行研究では人物属性分類の手法を導入することで，学生の{大学に来る, 13 時}や社会人の{残業になる, 平日}など，人物属性らしさを持つ周期的事象の獲得には一定の成功を収めた．しかし周期的事象と人物属性との関連の強さの評価や，獲得した周期的事象の評価などの点において問題が残っている．

本研究では我々の以前 [8] の研究を踏まえ，人物属性との関連性評価手法および周期的事象の有用性評価手法を導入することで，Twitter データからのより高品質な人物属性に特徴的な周期的事象知識の獲得を目指す．

2 関連研究

時間的な常識知識に着目した研究としては Zhou ら [9] の研究が挙げられる．Zhou らは時間的常識知識をモデルに与えることを目的として，時間的常識知識の収集および時間的常識推論モデルの作成を行った．目的の異なる研究ではあるものの，我々が獲得を目指す周期的事象も Zhou らの定義した複数の時間的常識知識に含まれている．しかし，Znou らの研究では行動主体の人物の属性が考慮されていない．本研究では人物の属性を考慮し，人物の属性ごとに特徴的な定期的事象知識の獲得を目指す．

3 データセット

本研究では人物属性ごとの周期的事象知識の抽出元として，属性ごとに分割された Tweet データセット (以下，属性別データセット) を作成する必要がある．以下，属性別データセットに含まれる人物属性の定義 (3.1 節)，作成に使用した 2 種類のデータセット (3.2 節) および先行研究 [8] の手法による属性別データセットの作成 (3.3 節) について述べる．

3.1 人物属性の定義

本研究では学生と社会人の 2 種類の人物属性を設定し，以下のように定義した．

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．

3.2 2 種類のデータセット

属性別データセットの作成のために，2 種類の Twitter ユーザデータセットを用意した．各データセットには複数の Twitter ユーザデータが含まれ，各 Twitter ユーザデータにはユーザ名，一定期間分の投稿 Tweet，自己紹介欄テキストなどのデータが紐づいている．

3.2.1 オリジナル大規模データセット

ランダムに収集した 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[1)]

1） https://developer.twitter.com/en/docs/twitter-api

表 1　属性別データセットに含まれるユーザ数と 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

を使用した．なおフィルタリングを行い，Tweet 投稿数の少なすぎるアカウントや商業目的のアカウントのデータは除外した．本研究では収集したデータの内，20 万ユーザ分のデータ (約 1.2 億 Tweet) を使用した．

3.2.2　ラベル付き小規模データセット

オリジナルデータセット (3.2.1 節) 中の一部の Twitter ユーザデータに，人手で人物属性ラベルを付与し作成したデータセットである．データ数は学生ユーザが 1032 ユーザ分 (約 100 万 Tweet)，社会人ユーザが 1216 ユーザ分 (約 120 万 Tweet) である．

3.3　属性別データセット

知識獲得においては知識の抽出元のデータセットの規模は大きいことが望ましいが，人手で大規模なラベル付きデータセットを作成することは難しい．そこで本研究では先行研究 [8] の手法に基づきラベル付き小規模データセット (3.2.2 節) から Twitter ユーザ人物属性分類モデルを作成する．具体的には BERT[10] 日本語事前学習モデル[2] を用いて，Tweet 単位での人物属性分類タスクのファインチューニングを行い，あるユーザの各 Tweet における人物属性分類の結果から当該ユーザの人物属性の推定するモデルである．オリジナル大規模データセット (3.2.1 節) 中の Twitter ユーザに用いることで擬似的に人物属性ラベルの付与を行い，属性別データセットを作成した．作成した属性別データセットに含まれるユーザ数と Tweet 数を表 1 に示す．

4　提案手法

本研究では人物属性において特徴的な周期的事象の獲得を目指す．まず，属性別データセット (3 節) 中の Tweet から事象を抽出する (4.1 節)．抽出した各事象に対し，時間区分における周期性の評価と人物属性との関連性の強さの評価を行い，スコアを付与し (4.2 節)，ランキングの作成を行う (4.3 節)．図 1 に提案手法の概要図を示す．以下，各手順について詳しく説明する．

2)　https://github.com/cl-tohoku/bert-japanese

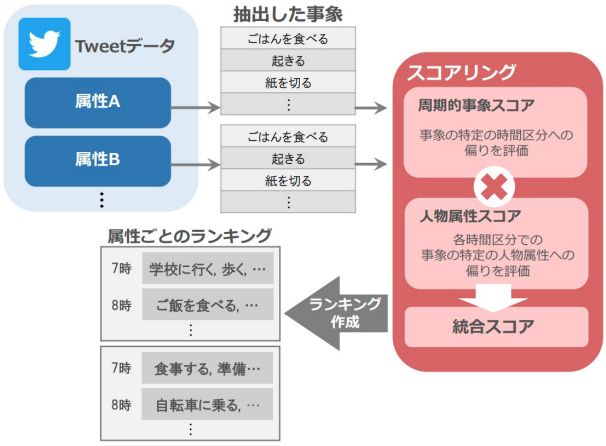

図 1　提案手法の概要図

4.1　事象の抽出

人間の行動を始め，様々な事象は動詞で表現されることが多い．そこで Tweet 中から動詞を抽出する．さらに，係り受け関係に名詞が存在する場合には，動詞単体に加えて名詞と動詞のペアも抽出する．抽出した名詞と動詞のペアは，“名詞+動詞” と表記する．本研究では，抽出した動詞単体および名詞+動詞をまとめて「事象語」と呼ぶ．

4.2　スコアリング

本研究では周期的事象の性質を持ち，かつ各人物属性において特徴的な事象の獲得を目指すため，周期的事象の性質を持つかを評価する周期的事象スコア (4.2.1 節) と人物属性において特徴的かを評価する人物属性スコア (4.2.2 節) の 2 種類のスコアを使用する．各スコアの詳細について以下に詳述する．

4.2.1　周期的事象スコア

平日・休日，曜日，朝昼晩，24 時間の 4 種類，計 36 個の時間区分を設定する．ある時間区分 t に対して，t 自身を除く t と同じ種類の任意の時間区分を t_{other} とする．ある事象語 w の時間区分 t における頻度を $f(w,t)$ とし，ある事象語の総頻度を $F(w)$ とする．また以下の 2 つの条件を設定する．

C1: 少数のユーザに集中的に使用されている
C2: 365 日の各日の頻度の分散が高い

このとき，事象語 w のある時間区分 t における周期的事象スコア $Score_p(w,t)$ をロジットを用いて以下のように定義する．

$$Score_P(w,t) = \begin{cases} 0 & (C1\ or\ C2\ or\ <0) \\ E\left(\log \dfrac{f(w,t)}{f(w,t_{other}) - f(w,t)}\right) & (otherwise) \end{cases}$$

4.2.2 人物属性スコア

属性 a のデータセット中での事象語 w の時間区分 t における頻度を $f_a(a,w,t)$ とする．属性 a 以外の任意の属性を a_{other} とする．このとき，属性 a のデータセット中での事象語 w の時間区分 t における人物属性スコア $Score_A(a,t,w)$ を，ロジットを用いて以下のように定義する．

$$Score_A(a,t,w) = \begin{cases} 0 & (<0) \\ E\left(\log \dfrac{f_a(a,w,t)}{f_a(a_{other},w,t) - f_a(a,w,t)}\right) & (otherwise) \end{cases}$$

4.3 ランキング作成

各事象語の最低頻度の閾値を m とする．属性 a のデータセットにおける時間区分 t の事象語 w の統合スコア $Score(a,t,w)$ を，周期的事象スコア (4.2.1 節) と人物属性スコア (4.2.2 節) の統合スコアを用いて以下のように定義する．

$$Score(a,t,w) = \begin{cases} 0 & (F(w) < m) \\ Score_P(w,t) \times Score_A(a,t,w) & (F(w) \geq m) \end{cases}$$

統合スコアを用いて，ある人物属性の各時間区分における周期的事象語のランキングを作成する．

5 実験

本研究では 3.1 節で定義した「学生」「社会人」の 2 種類の人物属性を周期的事象獲得の対象とする．周期的事象スコア (4.2.1 節) の条件 1 は抽出元 Tweet の 1 割以上を占める単一ユーザがいるか，条件 2 の分散は $5e-5$，ランキング作成 (4.3 節) の閾値 m は 500 に設定した．作成したランキングのうち，24 時間のランキングの 10 時から 12 時と平日休日の上位 5 件の結果を表 2 に示す．

6 評価実験

我々の以前の研究 [8] における手法をベースライン手法とする．具体的には，4.2.1 節の周期的事象スコアのみによる手法である．本手法によって獲得した事象語に対し，ベースライン手法との定性的評価および定量的評価を行う．

6.1 定性的評価

比較のためにベースライン手法による事象語のランキングを作成した．24 時間のランキングの 10 時から 12 時と平日休日の上位 5 件の結果を表 3 に示す．最低頻度などのパラメータは 5 節の設定の通りである．

6.2 定量的評価

提案手法により，目的とする知識の獲得に成功したとすると，作成したランキング中には各人物属性に特徴的な周期的事象語が多数含まれているはずである．そのため，作成したランキング中の語は人物の属性分類の素性として有効であることが予想できる．そこで各ユーザに対し，獲得した周期的事象語を踏まえたベクトル (以降，周期的事象ベクトルとする) を作成し，周期的事象ベクトルを素性として用いた人物属性分類を行うことで，間接的に提案手法の定量的評価を行う．

6.2.1 定量的評価手法

あるユーザ u の周期的事象ベクトルはランキング上位の単語群で構成される．各ベクトルの値は，ユーザ u の Tweet 群中における各単語の出現頻度を，Tweet 群中の全事象語数で正規化した値 $n^u_{(a,t,i)}$ である．次元数は 属性数 (a) × 時間区分 (t) × 単語数 (i) となる．

$$v_u = \{n^u_{(\text{社, 平, 1})}, n^u_{(\text{社, 平, 2})}, \cdots, n^u_{(\text{社, 休, 1})}, \cdots\}$$

本論文では，a = { 社会人, 学生 } の 2 次元，t = { 平日, 休日, 朝, 昼, $\cdots$, 22 時, 23 時 } の 32 次元，i = { ランキング 1 位, 2 位, $\cdots$, 10 位 } の 10 次元に設定した．

6.2.2 定量的評価結果

周期的事象ベクトルを用いた人物属性分類を行った．分類器はナイーブベイズを使用した．データは属性別データセット中の 9725 ユーザデータ (学生 4862 ユーザ，社会人 4863 ユーザ) を使用し，5 分割交差検定を行って F1 値の平均値を確認した．結果を表 4 に示す．

7 考察

表 2 より，学生のランキングには{大学+行く, 11 時}や{勉強+やる, 平日}などのように学校や勉学関連の事象語が含まれている．また，{寝坊+する, 10

表 2　提案手法によるランキング　10-12 時，平日休日の結果

属性	時間区分	事象語				
学生	10 時	飽きる	遅刻+する	打つ	寝坊+する	学校+休む
	11 時	テスト+終わる	指示+する	大学+行く	学校+終わる	布団+出る
	12 時	お腹+空く	捧げる	空く	学校+終わる	テスト+終わる
	平日	勉強+やる	テスト+終わる	学校+終わる	学校+休む	凍る
	休日	お腹+減る	自分+する	溶ける	中+有る	夢+見る
社会人	10 時	預ける	緩む	お迎え+する	一+成る	洗濯+する
	11 時	コス+する	昼+休む	スマホ+触る	散歩+する	発信+する
	12 時	教わる	茶+飲む	見守る	昼+休む	職場+する
	平日	方法+教える	スマホ+触る	茶+飲む	私+教える	スマホ+出来る
	休日	捏ねる	自分+する	行ける	移動+する	盗む

表 3　ベースライン手法のランキング　10-12 時，平日休日の結果

属性	時間区分	事象語				
学生	10 時	飽きる	寝坊+する	遅刻+する	二+寝る	布団+出る
	11 時	昼+食べる	布団+出る	寝坊+する	二+寝る	昼+する
	12 時	お腹+空く	昼+食べる	空く	捧げる	昼+する
	平日	学校+終わる	学校+休む	テスト+終わる	凍る	授業+する
	休日	お腹+減る	テレビ+する	会話+する	知らせる	聴く
社会人	10 時	朝+食べる	今日+休む	二+寝る	朝+言う	預ける
	11 時	昼+休む	昼+食べる	昼+する	買い物+行く	飯+する
	12 時	昼+休む	昼+食べる	昼+する	教わる	飯+する
	平日	昼+休む	残業+する	就く	しご+終わる	学校+行く
	休日	捏ねる	テレビ+する	渡る	習う	駅+する

表 4　手法の定量的評価のための人物属性分類結果

手法	F1
ベースライン	0.648
提案手法	0.676

時}や{布団+出る, 11 時}などのように，社会人では考えにくい生活リズムを示す事象語も確認できることから，学生の周期的事象の獲得には一定の成功を収めていると言える．社会人のランキングにおいては，{洗濯+する, 10 時}や{職場+する, 12 時}のように家事や勤務に関する事象語が獲得できているものの少数である．全体としては，どのような事象についての事象語か理解の難しいものが多い．本研究における社会人の定義には会社員と主ふの両方が含まれることから，会社員にのみ特徴的な語や主ふにのみ特徴的な語のスコアが低くなる傾向にあるためだと考えられる．

ベースライン手法によるランキング中には，{昼+食べる, 11 時}や社会人の{昼+休む, 12 時}などのように，その時間区分に強く関連するものの特定の人物属性に紐づかない事象語が存在している．また社会人の{学校+行く, 平日}のように明らかに当該人物属性に関連しない事象語も存在する．これらの事象語は提案手法のランキング (表 2) からは除去できており，提案手法の有効性が確認できた．

次に定量的評価について考察する．表 4 より，提案手法における周期的事象ベクトルでの分類のほうが高精度であるため，提案手法で獲得した周期的事象知識のほうが各人物属性に特徴的な周期的事象として適切であることが改めて確認できる．

定性的評価，定量的評価の両方において提案手法の有効性が確認できたことから，ベースラインと比べてより高品質な属性ごとの周期的事象の獲得に成功したと言える．しかしノイズ的な事象語もいまだに多く存在するため，事象語に付与した 2 種類のスコアへの重みの追加なども含め，知識獲得手法の改良は必要である．

8　おわりに

本研究では，人物属性ごとに特徴的な周期的事象の獲得を目指して Tweet データからの知識獲得を試みた．属性ごとのばらつきはあるものの，いずれの属性においても特徴的な周期的事象の獲得に成功した．またベースライン手法との比較により，定性的評価と定量的評価の両方において提案手法の優位性を確認した．今後はより高品質な周期的事象知識獲得を目指し，手法の改良に取り組む．

参考文献

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

[2] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of AAAI*, Vol. 33, pp. 3027–3035, 2019.

[3] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of ACL*, pp. 463–473, 2018.

[4] 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.

[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 periodic events with person attributes. In *Proceedings of IALP*, pp. 229–234, 2020.

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

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.