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
NLC2018-38 (2019-02)

Twitterからの事象パターン知識の獲得

山元　航平†　　嶋田　和孝†

† 九州工業大学情報工学部　〒820–0067 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，雑談対話システムに有効な知識獲得の手法を提案する．雑談対話システムでは，適切な発話を行うために幅広い知識が必要となる．しかし，そのような知識の人手での作成は高コストである．また，どのような要素を知識化するかは自明でなく，網羅性の問題が生じる．この問題を解決するため，本論文では Twitter に着目し，Tweet 内容とその Tweet の投稿時間の関係を利用して，自動で時間情報を保持した知識の獲得を行う．具体的には，収集された Tweet から事象語（動詞もしくは名詞と動詞のペア）を抽出し，5 つの時間区分における頻度分布（事象パターン知識）を獲得する．さらに，この頻度分布から，どの時間にどのような事象が生じているのかという知識も自動的に獲得する．実験結果より，夜に “寝る” のような一般的な知識のみならず，4 月に “チャレンジする” などの興味深い事例も得られた．

キーワード　知識獲得，テキストマイニング，マイクロブログ

Knowledge Acquisition about Event Information from Twitter

Kouhei YAMAMOTO† and Kazutaka SHIMADA†

† Department of Artificial Intelligence, Kyushu Institute of Technology
680–4 Kawazu Iizuka-shi Fukuoka, 820–8502, Japan
E-mail: †{k_yamamoto,shimada}@pluto.ai.kyutech.ac.jp

Abstract　In this paper, we propose a knowledge acquisition method for non-task-oriented dialogue systems, Such dialogue systems need a wide variety of knowledge for generating appropriate and sophisticated responses. However, constructing such knowledge is costly. To solve this problem, we focus on a relation about each tweet and the posted time. First, we extract event words, such as verbs, from tweets. Then, we generate frequency distribution for five different time divisions, e.g., a monthly basis. We checked high ranked event words in each time division. As a result, we obtained not only common-sense things such as “sleep” in night but also interesting events such as “challenge” in April (April is the starting month in Japan.)

Key words　knowledge aqcuisition, text mining, microblogging

1. はじめに

情報技術の発展に伴い，対話そのものを楽しむことを目的とした，非タスク指向型対話システム（以後，雑談対話システムと呼ぶ）が一般的になりつつある．代表的なシステムとしては，Microsoft 社のりんな (注1)などが挙げられる．また，広く利用されるようになった Apple 社の Siri(注2)や Softbank 社の Pepper(注3)なども，特定のタスクの達成を目的とするタスク指向型のシステムという側面が強いものの，対話そのものを楽しむことを目的とした非タスク指向型の機能も持ち合わせている．

このように身近な存在になりつつある雑談対話システムであるが，研究は古くから行われており [1]，これまでに様々な手法が提案されてきた．ここで次のような例を考える．

対話例 1 (発話日時：2 月 7 日 13:00)
システム: もう昼食は食べましたか？
ユーザ: はい、もう食べました。

対話例 2 (発話日時：2 月 7 日 15:00)
ユーザ: 昨日は海水浴に行ってきました。
システム: 2 月に海水浴とは珍しいですね。

対話例 3 (発話日時：3 月 15 日 12:00)
システム: そろそろ桜が咲く時期ですね．
ユーザ: そうですね．桜が咲いたらお花見に行きたいです．

(注1)：https://www.rinna.jp/
(注2)：https://www.apple.com/jp/siri/
(注3)：https://www.apple.com/jp/siri/

対話例 1〜3 は，様々な事象と時期・時間帯の関係を考慮してシステムが発話を行っている対話例である．対話例 1 では，対話時の時間が 13 時であることを踏まえて，ユーザがすでに昼食をとったかシステムが問いかけている．対話例 2 では，対話時が 2 月であることを踏まえ，ユーザの海水浴に行くという行動が一般的でな行動ではないことをシステムが指摘している．また，対話例 3 では対話時が桜の開花が始まる時期であることを踏まえ，桜に関係する発話をシステムが行っている．

このような対話を単純に実現するには，ELIZA [2] に代表されるルールベース型システムのように，様々な事象と時期・時間帯についての発話ルールを人手で作成すればよい．例えば，昼食をとったかを尋ねる発話を 12 時から 14 時に行うというルールを作れば，対話例 1 は実現できる．しかし，実際の雑談対話システムでは，様々な事象に対応する必要がある．昼食をとること以外にも事象は膨大に存在する上に，事象ごとに適切に時間の設定をする必要がある．そのため，幅広い事象と時期・時間帯の関係を考慮した発話が可能なシステムを人手で作成するには，非常に高いコストがかかる．さらに，人手でのルールの改良や拡張が，ルールベース型システムの性能向上につながりにくいことがすでに報告されている [3]．

雑談対話システムの発話内容や対応可能なテーマを増やすための手法としては，Web テキストを活用する手法が多数提案されている．具体的には，Web ニュースを利用する手法 [4] や Web ニュースと Wikipedia を利用する手法 [5]，Wikipedia と複数の概念辞書を利用する手法 [6]，Wikipedia からトリビア文を取得して発話に使用する手法 [7]，Twitter を利用する手法 [8] などが提案され，一定の成果を上げている．

そこで本研究では，雑談対話システムでの使用を想定し，Web 上からの事象パターン知識の自動獲得手法を検討する．事象パターン知識とは，様々な事象がどのような時期・タイミングに発生する傾向にあるのかという，事象の頻度分布である．具体的には，人々の行動や自然現象・社会現象がリアルタイムに反映されている SNS である，Twitter からの事象パターン知識の自動獲得を試みる．

2. 関 連 研 究

雑談対話時の時間を考慮したシステムの研究としては，佐藤ら [9] の非明示的な発話状況を考慮したシステムの研究がある．佐藤らは，発話状況の中でも特に，発話者のドメインや対話時の時間を考慮しているが，時間に関しては季節（春夏秋冬）で区分するのみである．本研究では，複数の時間区分を考慮する．

対話システムでの使用を想定した知識獲得の研究は幅広く行われている [10]〜[13]．下川ら [10] は，対話文内での話題把握のために Web 上からの話題語の抽出を行っている．Narisawa ら [11] は，対話システムへの応用が可能な含意関係認識のタスクにおいて，Web 上からの人間の数量感覚知識の自動獲得を行っている．さらに，どちらも人手での収集ではあるが，光田ら [12] は雑談対話には言外の情報が重要であるとして，対話における言外の情報の知識の収集を行い，町田ら [13] は関連語の知識をゲーミフィケーションを利用して収集している．

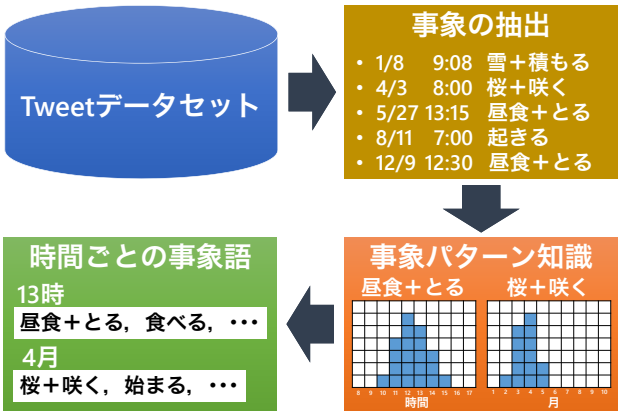

図 1　提案手法の概略図．

また，人々の行動という事象に着目し，Twitter などからの知識獲得を行っている研究も複数存在する．馬緤ら [14] は Tweet の本文と Twitter ユーザのアカウント情報から，職業ごとの行動に関する知識の収集を行っている．加藤ら [15] は Twitter の投稿内容からユーザ属性と習慣行動の推定を行っているが，その際に人々の行動知識の抽出を行い，その知識を習慣行動の推定に利用している．これら 2 つの研究はいずれも行動知識の獲得を行っているが，馬緤ら [14] は行動の主体である職業を，加藤ら [15] は行動の内容をあらかじめ設定しており，設定に沿う知識についてトップダウン的に収集を行っている．一方本研究では，より広範な知識が対話システムの性能向上につながることを期待し，ボトムアップ的に知識の獲得を行う．

3. 提 案 手 法

本研究では，事象パターン知識の獲得を目指す．事象パターン知識とは，人間・動物の行動，自然現象や社会現象などの様々な事象と，その事象がどの時期・タイミングで発生する傾向にあるかという事象ごとの頻度分布の組の情報であるとする．知識獲得の手順としては，まず大規模な Tweet データからの事象の獲得を行う．そして獲得できた事象ごとに頻度分布（事象パターン知識）を作成する．その後，事象パターン知識の獲得ができているかの確認を行う．提案手法の概略図を図 1 に示す．以下，各手順についての詳しい説明を行う．

3.1 事象の抽出

Tweet にはユーザ自身の行動をはじめとして，ユーザの身の回りで起きたことや時事に関する事象が含まれている．様々な事象は動詞で表現されることが多いため，まず Tweet 本文から動詞を抽出する．また “行く” と “学校に行く”，“咲く” と “桜が咲く” のように同じ動詞で表現されるが，抽象度合いの異なる事象も存在する．本研究では幅広い事象パターン知識の獲得を目指しているため，このような抽象度合いのことなる事象も区別して収集を行いたい．そこで，抽出した動詞の直前に名詞が存在する場合は，動詞単体とは別に名詞と動詞の組の抽出も行う．以後，抽出を行う動詞および動詞と直前の名詞の組をまとめて事象語と呼ぶ．

表 1　設定した時間区分

時間区分	集計の幅
1 年間	1ヶ月ごと
平日・週末	平日（月〜金曜日），週末（土曜日，日曜日）
1 週間	曜日ごと
朝・日中・夜	朝（3〜8 時），日中（9〜17 時），夜（18〜2 時）
24 時間	1 時間ごと

Tweet の解析には形態素解析機の MeCab(注4)を使用する．また，SNS 上で使用される新語や固有表現に対応するため，MeCab 用のシステム辞書である mecab-ipadic-NEologd(注5)を使用する．

3.2　事象の頻度分布の作成

抽出した事象語を抽出元の Tweet の投稿時間をもとにして集計する．本研究では幅広い事象を解析の対象としている．例えば，四季が関係するような自然現象は 1 年間の頻度分布に強い特徴を，人間の日常的な行動は 24 時間や 1 週間の頻度分布に強い特徴を持つのではないかと予想できる．そこで，様々な事象の出現パターンに対応できるよう，5 つの時間区分を設定する．設定した時間区分の詳細を表 1 に示す．なお，朝・日中・夜の時間の幅に関しては，気象庁の定義(注6)を参考にした．設定した時間区分ごとに集計を行うことで，1 つの事象語につき 5 種類の頻度分布を作成する．事象語とその事象語の 5 種類の頻度分布をまとめて事象パターン知識とする．

3.3　時間ごとの事象語の獲得

世の中のには様々な事象存在するが，その中には一定の期間に集中して発生する傾向のある事象が存在する．例えば，季節ごとの自然現象や人間の日常行動などである．そして，特定の期間に集中して発生する事象は，その期間らしさを強く持つ事象であると考えられる．そのため，本手法によって獲得した事象語のうち特定の期間に頻出する事象語が，その期間らしさを持つ事象語であれば事象パターン知識の獲得に成功していると考えられる．そこで，獲得した事象パターン知識の確認のため，各時間区分に集中して出現している事象語のランキングを作成する．具体的には，各事象語の時間区分ごとの出現回数を，その事象語のその時間区分におけるスコアとし，獲得した全事象語のすべての時間区分におけるランキングを作成する．

4.　実験・考察

4.1　データセット

本研究で使用したデータセットについて説明する。Tweet データの収集には Twitter 社から提供されている TwitterAPI を使用し、ランダムに指定したユーザのタイムラインを遡る方法で Tweet を収集した．この方法で Tweet が取得できたユーザのうち，過去 2 年以上の幅で Tweet の取得ができた約 12,000 ユーザの Tweet から各ユーザにつき任意の 1 年間分の Tweet を解析対象とした．その結果，約 270 万 Tweet が解析対象となった．

4.2　実験結果

対象となる 270 万 Tweet から，約 125 万個の事象語を抽出した．抽出したすべての事象語を集計したところ，約 32 万件の事象パターン知識が得られた．なお獲得した事象パターン知識のうち，データセット中に 2 回以上出現した事象語の事象パターン知識は約 8 万件であった．

獲得した事象パターン知識の確認のために，各時間区分ごとのランキングを作成した．事象語の閾値による結果の違いを確認するため，ランキングは各時間区分につき 2 つ（データセット中での最低出現回数 100 件のものと 1000 件のもの）を作成している．作成したランキングを表に示す．作成したランキングと表の対応は以下である．なお，いずれの表もランキングの上位 5 件の結果である．

・1 年間の事象パターン　閾値 100（表 2），閾値 1000（表 3）
・平日・週末の事象パターン　閾値 100（表 4），閾値 1000（表 5）
・1 週間の事象パターン　閾値 100（表 6），閾値 1000（表 7）
・朝・日中・夜の事象パターン　閾値 100（表 8），閾値 1000（表 9）
・24 時間の事象パターン　閾値 100（表 10），閾値 1000（表 11）

以下，各時間区分での結果，事象語の閾値間の比較結果，時間区分間の比較結果を節に分けてに示す．

4.2.1　1 年間（表 2，表 3）

表 2 より，1 年間の時間区分での出現回数 100 回以上の事象語のランキングの上位には，1 月・2 月に “積もる”，“雪＋降る”，“凍る” などの冬の自然現象の事象語や，3 月・4 月に桜に関係する事象語の “咲く”，8 月のセミに関係する事象語の “鳴く” など，自然現象に関係する事象語が見られる．また，8 月の “帰省＋する” や 12 月の “明ける” など季節のイベントに関する事象語も存在する．

しかし，同じく 1 年間の時間区分のランキングであっても，出現回数 1000 回以上の事象語のランキングである表 3 には，表 2 でみられたような，季節的な自然現象や季節的なイベントに関係する事象語は見られなかった．同様の設定で作成したランキング上位 30 件の事象語の確認も行ったが，その期間らしさを持つ事象語は存在しなかった．これは，データセット中に出現回数の多い事象語は動詞単体のものが多いために，意味が抽象的でその期間らしさを感じることができなかったためであると考える．

4.2.2　平日・週末（表 4，表 5）

表 3，表 4 より，平日・週末の時間区分でのランキングでは，平日に “病院＋行く” や仕事に関係する事象語である “残業＋する”，“仕事＋する” が見られる．また，休日に “勝つ”，“負ける”，“楽しむ”，“観る” など趣味やレジャーなどに関係すると思われる事象語が存在する．

4.2.3　1 週間（表 6，表 7）

表 6，表 7 より，1 週間の時間区分でのランキングでは，平日らしさ，休日らしさという観点ではいくつかの事象語が見られるものの，その曜日らしさのある事象語は確認できなかった．これは，例えば “平日としての水曜日らしさ” ではなく，“他の

(注4) : http://taku910.github.io/mecab/
(注5) : https://github.com/neologd/mecab-ipadic-neologd
(注6) : http://www.jma.go.jp/jma/kishou/know/yougo_hp/saibun.html

表 2　1 年間の事象パターン　閾値 100.

月	行動語				
1 月	積もる	雪+降る	凍る	MONG HANG+やる	滑る
2 月	斬る	MONG HANG+する	雪+降る	MONG HANG+やる	滑る
3 月	咲く	受かる	選べる	ふむ	MONG HANG+する
4 月	咲く	チャレンジ+する	うつ	病む	音楽+聴く
5 月	黙る	ここ+来る	傷つく	デート+する	今日+終わる
6 月	あたる	再生+する	うる	測る	文句+言う
7 月	シャワー+浴びる	殺る	アイス+食べる	ぬく	晴れる
8 月	帰省+する	鳴く	ぬく	映画+見る	掘る
9 月	ぬく	焼く	おなか+すく	募る	怪我+する
10 月	決闘+やる	投票+する	お誘い+待つ	募る	嫌+なる
11 月	決闘+やる	お誘い+待つ	待てる	うねる	ささくれる
12 月	納める	決闘+やる	お誘い+待つ	明ける	こう

表 3　1 年間の事象パターン　閾値 1000.

月	行動語				
1 月	降る	引く	覚える	始まる	困る
2 月	合う	作る	止まる	流れる	似る
3 月	届く	ちる	取る	言える	増える
4 月	着る	聴く	取れる	似る	合う
5 月	下さる	がんばる	当たる	おる	探す
6 月	勝つ	決める	描く	話す	降る
7 月	着る	書く	間に合う	描く	推す
8 月	居る	落ちる	遊ぶ	戻る	続く
9 月	上げる	撮る	決まる	続く	当たる
10 月	歩く	取る	続く	売る	付ける
11 月	似る	気+なる	見れる	困る	流れる
12 月	歌う	推す	お願い+する	引く	待つ

表 4　平日・週末の事象パターン　閾値 100.

期間	行動語				
平日	うねる	ついていける	病院+行く	残業+する	よぶ
週末	引き取る	打ち上げる	つづく	嘘+つく	並ぶ

表 5　平日・週末の事象パターン　閾値 1000.

期間	行動語				
平日	休む	仕事+する	届く	働く	売る
週末	勝つ	負ける	楽しむ	観る	すげる

すべての曜日と比較した際の水曜日らしさ” など，本質的なその曜日らしさがというものが，他の時間区分のその時間区分らしさと比較して非常に小さく，本研究で使用したデータセットの規模ではその特徴を抽出することができなかったためではないかと考えられる．

4.2.4　朝・日中・夜（表 8，表 9）

表 8，表 9 より，朝・日中・夜の時間区分でのランキングでは，朝に “今日＋頑張る”，“寝坊＋する”，“起きる”，日中に “昼寝＋する”，夜に “寝る” など，その時間区分に関する事象語が確認できる．

4.2.5　24 時間（表 10，表 11）

表 10，表 11 より，24 時間の時間区分でのランキングでは，0〜3 時にかけて “眠れる”，“ねれる” など睡眠に関する事象語が見られる．また，5〜8 時にかけては起床に関する事象語が確認できる．

4.2.6　事象語の閾値間の比較結果（表 2〜表 11）

全体的な傾向として出現回数が 100 回以上のランキングには名詞＋動詞の事象語が多く，出現回数が 1000 件以上のランキングには動詞単体など抽象的な事象語が多く確認できる．これは，提案手法において，名詞＋動詞のペアを事象語として Tweet から抽出した場合には，同様に動詞単体のものも別の事象語として抽出を行っているため，当然の結果であるといえる．

4.2.7　時間区分間の比較結果（表 2〜表 11）

1 年間での事象パターンでは季節らしさのある事象語が，24 時間の事象パターンでは人間の日常行動に関係する事象語が確認できるなど，時間区分ごとに確認できる事象語の種類が大きく異なっている．本手法では，様々な事象の出現パターンに対応できるよう 5 つの時間区分を設定したが，その設定が想定通りうまく働いたといえる．

4.3　考　　察

本研究の手法によって事象パターン知識の獲得に成功しているかを確認するために，事象語のランキングを作成したが，表 2，表 4，表 5，表 8，表 9，表 10，表 11 の 7 つのランキングではその期間らしさを持つ事象語が確認できる．このことから，曜日以外の時間区分での頻度分布に特徴を持つ事象パターン知識の獲得は，一定の成功を納めていると判断できる．加えて，表 2 の 4 月の欄に “チャレンジ＋する” という興味深い事象語が見られる．4 月に “チャレンジ＋する” という事象語が頻出するという知識は，言われてみれば確からしさのある知識ではあるが，4 月に関連する事象を人手で集めるという場面において

表 6　1 週間の事象パターン　閾値 100.

曜日	行動語				
月曜	待てる	怪我+する	ついていける	積もる	崩れる
火曜	黙る	再生+する	たくさん+いる	奢る	おなか+すく
水曜	解る	祝う	勉強+なる	修正+する	通じる
木曜	コーヒー+飲む	完了+する	そう	炊く	盛る
金曜	いらっしゃる	カット+する	今日+終わる	今日+頑張る	好む
土曜	ささくれる	うつ	こう	引き取る	つづく
日曜	号泣+する	放つ	打ち上げる	あたる	声+聞こえる

表 7　1 週間の事象パターン　閾値 1000.

曜日	行動語				
月曜	休む	いただく	こと+する	生きる	お願い+する
火曜	困る	上げる	教える	気+なる	こと+する
水曜	上げる	取る	減る	似る	おく
木曜	決まる	届く	仕事+する	売る	怒る
金曜	間に合う	働く	の+思う	仕事+する	食う
土曜	楽しむ	勝つ	観る	歌う	飲む
日曜	負ける	勝つ	楽しむ	いただく	遊ぶ

表 8　平日・週末の事象パターン　閾値 100.

時間帯	行動語				
朝	尾羽+いる	うねる	今日+頑張る	寝坊+する	曇る
昼	昼寝+する	待てる	つづく	混む	当選+する
夜	たま+絡める	フォロー+困る	リプライ+くださる	絡める	ねれる

表 9　平日・週末の事象パターン　閾値 1000.

時間帯	行動語				
朝	降る	起きる	休む	乗る	頑張る
昼	売る	届く	帰る	買う	当たる
夜	寝る	歌う	描く	聴く	見れる

表 10　24 時間の事象パターン　閾値 100.

時間	行動語				
0 時	ねれる	えむ	傷つく	眠れる	挙げる
1 時	ねれる	病む	眠れる	時間+起きる	寝落ち+する
2 時	時間+起きる	腹立つ	ねれる	目+覚める	覚める
3 時	時間+起きる	ねれる	今+寝る	眠れる	目+覚める
4 時	曇る	朝+なる	成る	目+覚める	覚める
5 時	うねる	曇る	朝+なる	家+出る	成る
6 時	うねる	朝+なる	曇る	今日+頑張る	尾羽+いる
7 時	尾羽+いる	よぶ	今日+頑張る	今日+終わる	寝坊+する
8 時	尾羽+いる	寝坊+する	お待ち+する	測る	座れる
9 時	寝坊+する	売り切れる	離す	遅刻+する	つかう
10 時	奢る	買い物+行く	掛かる	今日+頑張る	なくす
11 時	今日+帰る	飯+食う	着せる	調整+する	再生+する
12 時	買い物+行く	影響+する	ニュース+見る	はじまる	こと+思い出す
13 時	外+出る	強化+する	昼寝+する	浸る	ストレス+溜まる
14 時	おなか+すく	MONG HANG+する	昼寝+する	乾く	引っ越す
15 時	待てる	成る	昼寝+する	注意+する	表現+する
16 時	黙る	つづく	成る	つながる	表す
17 時	当選+する	絶対+行く	お腹+減る	ご飯+作る	申し込む
18 時	晴れる	絶対+行く	収まる	残業+する	巡る
19 時	募集+する	うる	映画+見る	デート+する	縛る
20 時	こう	号泣+する	挑む	燃える	再生+する
21 時	いらっしゃる	ぼる	響く	ありえる	紹介+する
22 時	日付+変わる	受かる	酔っ払う	度+寝る	喧嘩+する
23 時	たま+絡める	フォロー+困る	リプライ+くださる	絡める	日付+変わる

は，思いつきにくい知識であると考えられる．詳しい検証が必要ではあるが，今回の手法で，人間には思いつきづらい事象のパターン知識を獲得できている可能性がある．

本研究は雑談対話システムでの使用を想定して行っている．全体として長期の時間区分での結果には季節らしさの強い事象語が，短期の時間区分での結果には日常生活らしさの強い事象語が見られた．このため，システムが発話を行う際には，参照する時間区分を変えることで季節らしさのある発話や日常生活らしさのある発話など，発話の種類を任意に変更することが可能になるのではないかと考える．

また，今回行った頻度によるランキング作成の手法は，非常に素朴な手法であった．そのため，その時間区分らしさを持つ

表 11　24 時間の事象パターン　閾値 1000.

時間	行動語				
0 時	寝る	生きる	く	すげる	ん+思う
1 時	寝る	起きる	決める	続ける	生きる
2 時	寝る	起きる	負ける	怒る	勝つ
3 時	起きる	寝る	決める	気づく	戻る
4 時	起きる	寝る	降る	頑張る	減る
5 時	起きる	降る	寝る	休む	働く
6 時	降る	起きる	頑張る	付ける	寝る
7 時	降る	がんばる	ちる	休む	起きる
8 時	休む	くださる	乗る	おる	付ける
9 時	いただく	受ける	取れる	着る	歩く
10 時	売る	乗る	降る	探す	送る
11 時	かかる	売る	決める	取れる	食う
12 時	売る	探す	取れる	ひる	取る
13 時	付ける	食う	置く	買う	届く
14 時	取れる	勝つ	送る	当たる	売る
15 時	とる	歩く	似る	いただく	帰る
16 時	お願い+する	引く	付ける	ちる	歩く
17 時	届く	間に合う	食う	当たる	帰る
18 時	当たる	売る	届く	減る	食う
19 時	下さる	歌う	推す	見れる	待つ
20 時	見れる	歌う	勝つ	行ける	観る
21 時	描く	撮る	すげる	合う	見れる
22 時	聴く	会う	観る	気+なる	撮る
23 時	困る	くださる	寝る	歌う	笑う

事象語ではないが，別の要因によってその時間区分に極端に集中して出現した事象語も，ランキングの上位に来てしまっている．例えば，表 10 の “MONG HANG +する” や “当選+する” などである．前者は 2018 年 1 月における人気ゲームシリーズである MONSTER HUNTER WORLD の発売，後者は 2017 年 10 月における衆議院議員総選挙の影響で，関連する内容の Tweet が特定の期間に多くなされたために，ランキング上位に来てしまっている．同一時間区分内での分散を利用するなどして，バースト的に発生した事象語を除き，普遍的なその時間区分らしさを持つ事象語の獲得に努めたい．

5. おわりに

本研究では，雑談対話システムでの使用を目的として，Tweet データからの事象パターン知識の獲得手法を提案した．Tweet に含まれる様々な事象と Tweet の投稿時間データに着目した素朴な手法ではあるが，複数の事象パターン知識の獲得に成功した．また，人手での獲得は難しいのではないかと思われる，興味深い知識の獲得にも成功した．今後は，事象語の評価手法を改良などを行い，その時間区分らしさを強く持つ事象語の獲得を行う．さらに，最終的な目標とする雑談対話システムにおける，事象パターン知識の有効的な活用手法の検討に取り組む．

文　献

[1] 河原達也 : “音声対話システムの進化と淘汰: 歴史と最近の技術動向 (<特集> 音声対話システムの実用化に向けて)”, 人工知能学会誌, **28**, 1, pp. 45–51 (2013).

[2] J. Weizenbaum: “Computer power and human reason: From judgment to calculation.” (1976).

[3] R. Higashinaka, T. Meguro, H. Sugiyama, T. Makino and Y. Matsuo: “On the difficulty of improving hand-crafted rules in chat-oriented dialogue systems”, APSIPAIEEE, pp. 1014–1018 (2015).

[4] 水野淳太, 乾健太郎, 松本裕治 : “ウェブニュースを利用した雑談対話システム”, 人工知能学会言語・音声理解と対話処理研究会資料, **55**, pp. 1–6 (2009).

[5] 江頭勇佑, 柴田知秀, 黒橋禎夫 : “Web から獲得した知識に基づく雑談対話システム”, 情報処理学会関西支部 支部大会 講演論文集 (2011).

[6] 吉田裕介, 萩原将文 : “複数の言語資源を用いたユーモアを含む対話システム”, 知能と情報, **26**, 2, pp. 627–636 (2014).

[7] K. Niina and K. Shimada: “Trivia score and ranking estimation using support vector regression and ranknet”, Proceedings of PACLIC, **32**, (2018).

[8] 稲葉通将, 神園彩香, 高橋健一 : “Twitter を用いた非タスク指向型対話システムのための発話候補文獲得”, 人工知能学会論文誌, **29**, 1, pp. 21–31 (2014).

[9] 佐藤翔悦, 吉永直樹, 豊田正史, 喜連川優 : “非明示的な発話状況を考慮したニューラル対話モデルの検討”, 第 31 回人工知能学会全国大会, pp. 1B1OS25a1–1B1OS25a1 (2017).

[10] 下川尚亮, 荒木健治 : “対話文生成のための web を用いた話題語の抽出”, 研究報告自然言語処理 (NL), 2, pp. 121–126 (2009).

[11] K. Narisawa, Y. Watanabe, J. Mizuno, N. Okazaki and K. Inui: “Is a 204 cm man tall or small? acquisition of numerical common sense from the web”, Proceedings of ACL, Vol. 1, pp. 382–391 (2013).

[12] 光田航, 東中竜一郎, 牧野俊朗, 松尾義博 : “雑談対話における言外の情報を推定するためのデータ収集と分析”, 第 30 回人工知能学会全国大会, pp. 2P14in1–2P14in1 (2016).

[13] 町田雄一郎, 河原大輔, 黒橋禎夫, 颯々野学 : “関連語知識獲得のための対話システム上の連想ゲームのデザイン”, 情報処理学会論文誌, **57**, 3, pp. 1058–1068 (2016).

[14] 馬緤美穂, 笹野遼平, 高村大也, 奥村学 : “職業ごとの行動に関する知識の収集”, 研究報告自然言語処理 (NL), pp. 1–8 (2015).

[15] 加藤諒, 中村健二, 山本雄平, 田中成典, 坂本一磨 : “マイクロブログにおけるユーザの属性と習慣行動の推定に関する研究”, 情報処理学会論文誌, **57**, 5, pp. 1421–1435 (2016).