Twitter 投稿における定性的時間表現の使用時刻分布調査

奥川 智貴
筑波大学情報学群
t.oku@mibel.cs.tsukuba.ac.jp

乾 孝司
筑波大学大学院システム情報工学研究科
inui@cs.tsukuba.ac.jp

概要

人間と自然な対話ができるシステムを実現するためには、時間表現とあいまい表現の理解が不可欠である。本稿では、あいまいさを含む時間表現を定性的時間表現と呼び、定性的時間表現の体系的な整理と、Twitter 投稿における定性的時間表現の使用時刻分布を調査することを目的とする。まず、定性的時間表現を単一型と複合型に大別した上で TimeML の<TIMEX3>の仕様に沿う形で表現の整理をおこなった。次に、複合型定性的時間表現である「もう h 時」および「まだ h 時」を題材にして、これら時間表現を含むテキストの投稿時刻と、当該の時間表現によって言及されている時刻の間の時間ギャップを比較することで、定性的時間表現の使われ方の違いを調査した。

1 はじめに

人間と自然な対話ができるシステムへの需要は高く、その実現のためには時間表現の深い理解が不可欠である。例えば Apple サポート [1] は音声対話システム「Siri」への基本的な発話例を 43 件示しているが、そのうち時間に関するフレーズが含まれる例文は 17 件にも上る。カレンダーやアラーム等の時間管理アプリケーションの設定指示（例：「明日 4 時に起こして」）のみならず、日常的な質問や雑談に付随する表現（例：「今朝のニュースは？」）等、時間表現の利用シーンは多岐にわたる。

また、あいまい表現の処理も対話の自然さを生み出すためには必須である。「コーヒーいれるけど量どうする？」と聞かれた時に我々人間は常に「95mL ください」などと明確な数量を言い伝えるわけではなく、「若干少なめでください」のようにあいまいさを含む表現をしばしば使うだろう。こうしたあいまいさは話者個人の感覚を反映する、人間らしい発話の特徴といえる。

本稿では、時間表現でありかつあいまい表現となる語句を「定性的時間表現」と呼び、定性的時間表現の体系的な整理と、Twitter[1)]投稿における定性的時間表現の使用時刻分布を調査することを目的とする。特に、「もう 4 時」や「まだ 4 時」などの一部の定性的時間表現を対象にした調査をする。Twitter 投稿には投稿者のリアルタイムな感覚が反映されたテキストデータに投稿時刻データが紐づいており、この情報を調査に利用する。具体的には、定性的時間表現を含むテキストが投稿された時刻とその投稿内で言及されている時刻の間の時間のギャップに注目し、対象表現ごとのギャップを比較することで、定性的時間表現の使われ方の違いを調査する。

2 関連研究

2.1 自然言語処理における時間表現の扱い

自然言語処理分野における時間表現に注目した研究では、TimeML[2] の<TIMEX3>に基づいたタグづけ基準を採用したアノテーションが盛んである。TimeML とは時間表現と実世界上のイベントを関連づけるタグづけ基準のことであり、これに沿って作成された<TIMEX3>は様々な言語へのアノテーションに使われている。小西ら [3] は日本語時間表現に対するタグ付け基準を提案しており、対象とする表現を DATE（日付表現）・TIME（時刻表現）・DURATION（持続時間表現）・SET（頻度集合表現）の 4 種類に分類している。DATE・TIME は「昨日」「4 時」等、時点及び時区間の時間軸上の位置を定義するために用いられる表現である。DURATION は「1 時間」等、時間軸上の位置ではなく時間の量を定義するために用いられる表現である。SET は「週一回」等、時間軸上の複数の時区間を定義するために用いられる表現である。

小西らの提案した基準に概ね沿う形で成澤 [4] は時間表現のアノテーションを行ったが、時間表現の前後の文字列が修飾表現であった場合の規格化に関

1） https://twitter.com/

しては大半が<TIMEX3>で定義される@mod 属性に値を付与する処理だと述べている。@mod は時間情報表現のモダリティを表す属性である。例えば文中の「2021 年 12 月 10 日ごろ」は以下のようにタグ付けされる。

```
<TIMEX3 tid="t1" type="DATE" value="2021-12-10" valueFromSurface="2021-12-10" mod="APPROX">2021 年 12 月 10 日ごろ</TIMEX3>
```

ここでは@mod に格納されている “APPROX” という値が「ごろ」という表現に対応している。本研究では@mod を時間表現のあいまい性を捉える属性として注目する。詳細は 3 節で説明する。

2.2 Twitter を活用した時間研究

Twitter を活用した時間情報に焦点をあてた研究も多く存在する。例えば Hürriyetoğlu ら [5] はあるイベントに関する投稿のストリームからイベント発生までの時間を推定する手法を提案した。この手法では、各投稿の文中に含まれる時間表現と投稿時刻を基に演算処理した結果の中央値や平均値が推定に利用されている。しかしこうした既存研究のほとんどはイベントとその発生時間の対応付けが目的で、時間表現が含まれる文の投稿時刻の分布自体には焦点を当てていない。

3 定性的時間表現

3.1 単一型の定性的時間表現

本稿では定性的時間表現を単一型と複合型の 2 種類に分類した上で<TIMEX3>の仕様に沿って各々説明する。

単一型の定性的時間表現とは、時間表現のうち、それ単体でおよその時刻・時間帯や時間量をあらわす語句である。ただし、ある程度定量的な定義が存在する語句でも定性的時間表現とみなせる場合もある。例えば「朝」は気象庁 [6] が「午前 6 時頃から午前 9 時頃まで」の時間帯を指す言葉だと定めているが、話者自身が感覚的に定めた時間区分に基づいて使われている。単一型定性的時間表現の例を表 1 に示す。

3.2 複合型の定性的時間表現

複合型の定性的時間表現とは、時間表現に、@mod、@aspect の 2 属性のいずれかに値を与えるような修飾表現が結びついた語句である。@mod は先述の通り<TIMEX3>で既に採用されている属性で、時間情報のモダリティを表す。「ごろ」や「前」、「過ぎ」等の修飾表現が該当する。一方で@aspect とは今回新たに導入した属性で、話者の感覚や行動の段階や局面を表す。「もう」や「まだ」、「ようやく」等の修飾表現が該当する。複合型定性的時間表現の例を表 2 に示す。ほかにも、@mod にあたる修飾表現と@aspect にあたる修飾表現の両方と結びついた時間表現（例：「まだ 4 時過ぎ」）なども複合型定性的時間表現に該当する。

表 1 単一型の定性的時間表現の例

	例
DATE	「先日」「後日」
TIME	「朝」「深夜」
DURATION	「一瞬」「数日」
SET	「週に数回」

表 2 複合型の定性的時間表現の例

	例（+@mod）	例（+@aspect）
DATE	「10 月半ば」	「もう 10 月」
TIME	「4 時過ぎ」	「まだ 4 時」
DURATION	「1 時間未満」	「ようやく 1 時間」
SET	「週一回以上」	「ちょうど週一回」

4 使用時刻分布調査

4.1 調査対象表現

調査対象として、正時を言及していると考えられる複合型の定性的時間表現である「もう h 時」および「まだ h 時」（$h = 0, 1, ..., 24$、以降の h に関しても同様で、「正時」とも呼ぶ）を選んだ。「もう」および「まだ」は先述の通り @aspect の値を与える修飾表現であり、「h 時」は明確に時刻（正時）を言及するために使われる一般的な時間表現である。

4.2 調査データ

調査データは、2014 年 1 月 1 日から 2014 年 12 月 31 日までに東京都で投稿された日本語ツイートの集合とし、ここから正規表現を用いてテキスト中に調査対象表現が含まれている投稿を抽出して調査に利用した。ここで、例えば「もう 3 時」を抽出する正規表現（Python で記述）は以下のものを用いた。

```
"もう"+"[3 ３三]"+"時[^間半 0-9 ０-９]"
```

各投稿にはメタデータとしてタイムスタンプ、すなわち投稿時刻情報が付与されている。

表 3　*Gap* 値の計算例

テキスト	*Timestamp*	*ReferenceTime*	*Gap*
「もう 2 時かぁ…」	2014-02-01T02:05:00	2014-02-01T02:00:00	-300
「もう二時かよ ww」	2014-02-01T01:57:30	2014-02-01T02:00:00	150
「もう 2 時〜」	2014-02-01T13:50:00	2014-02-01T14:00:00	600

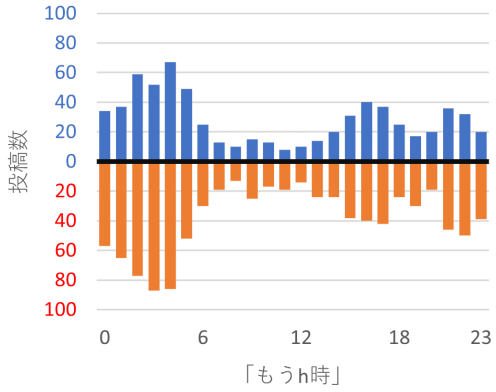

図 1　もう *h* 時：正時別の投稿件数 (上方向の要素は正、下方向の要素は負のギャップをとる投稿件数)

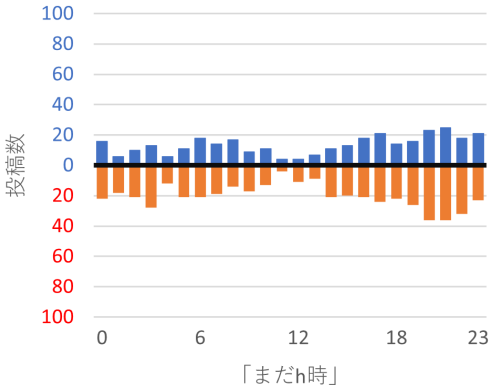

図 2　まだ *h* 時：正時別の投稿件数 (上方向の要素は正、下方向の要素は負のギャップをとる投稿件数)

4.3　ギャップに注目した使用時刻分布調査

本研究の目的を達成するために、ギャップ（*Gap*）という指標を導入する。ギャップとは、ある時間表現を含むテキストが投稿された時刻からその時間表現が表層的に言及している時刻までに経過した秒数と定義する。ある投稿 t に含まれる時間表現 e のもつギャップは以下の式で求められる。

$$Gap(t,e) = ReferenceTime(t,e) - Timestamp(t)$$

ここで、*Timestamp* は対象表現が実際に使用された時刻を表し、ツイートのタイムスタンプの値がそのまま代入される。*ReferenceTime* は対象表現から修飾表現を取り除いた正時時刻であり、*Timestamp* の年月日 + h : 00 : 00 の値が代入される。ただし、$h = 24$ ならば $h = 0$ に変換した上で代入を行う。また、$h = 0, 1, ..., 12$ かつ $-64,800 \leq Gap \leq -21,600$ ならば、午後について言及した投稿だと判断し、*ReferenceTime* の値を 43,200 秒（= 12 時間分）加算して *Gap* を再計算する。例えば 13 時 50 分ちょうどに投稿された「もう 2 時」という投稿の場合、再計算前は $Gap = -42,600$ であるが、再計算後は $Gap = 600$（ = 10 分）となる。$Gap < -64,800$ または $Gap > 21,600$ の場合は、現在時刻とは無関係の投稿だと判断し、集計から除外した。

上記の結果、1,621 件の「もう h 時」と 817 件の「まだ h 時」の表現、およびそれらの *Gap* 値を得た。*Gap* 値の計算例を表 3 に示す。

4.4　調査結果

4.4.1　使用時刻分布の比較　—「もう」と「まだ」—

図 5 が主となる調査結果であるが、その前段階として図 1 から図 4 で正時別の調査結果をまず報告する。正時別のツイート件数を図 1 および図 2 に示す。「もう h 時」を含むツイートは 3 時、4 時をピークにして深夜の時間帯を言及したものが多く、「まだ h 時」を含むツイートは 20 時、21 時を緩やかなピークにして夜の時間帯を言及したものが多く見られた。つまり言及される時間帯によって対象表現の使用頻度に差が生じていることがわかる。*Gap* 値の正負に注目すると、「もう」と「まだ」のどちらの表現とも若干負の件数が多いものの時間帯による正負の違いはあまり見られない結果となった。

次に、正時別のギャップ分布として、*Gap* 値の平均値および中央値を図 3 および図 4 に示す。ここでは平均値を棒グラフで示し、その上に中央値を示す点グラフを重ねている。平均値に関して、部分的に極端な値をとっているが、これは外れ値の影響を受けているものであったため、ここでは中央値に注目する。中央値に基づくとほとんどの時間帯で「もう h 時」は 0 秒に近い *Gap* 値であり、「まだ h 時」は − 600 秒以上 0 秒以下の *Gap* 値であることが概観できる。

図 3 および図 4 の正時別調査結果の要約として、

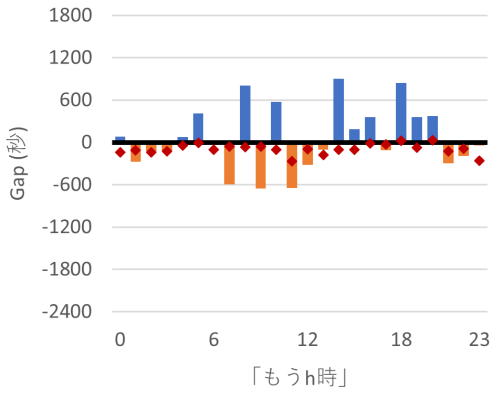

図 3　もう *h* 時：正時別のギャップ平均値・中央値

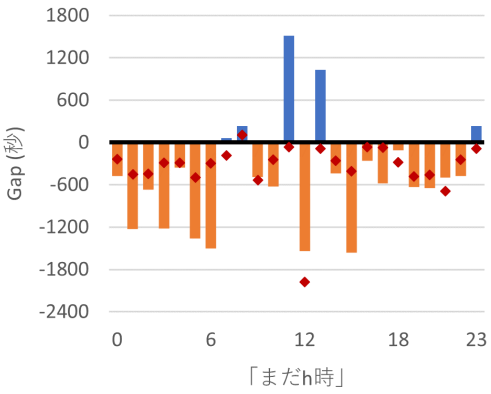

図 4　まだ *h* 時：正時別のギャップ平均値・中央値

@aspect 別のギャップ分布の箱ひげ図を図 5 に示す。*Gap* 値の平均値を三角の点で、中央値を赤線で示している。「もう *h* 時」のツイート全体の *Gap* 値の平均値は 38 秒、中央値は –85 秒 であった。「まだ *h* 時」のツイート全体の平均値は –554 秒、中央値は –285 秒 であった。この結果は、「もう *h* 時」はほぼ *h* 時に使われる一方で、「まだ *h* 時」は約 5 分後 に使われており、@aspect に対応する修飾表現の「もう」と「まだ」では使用時刻に差があることがわかる。

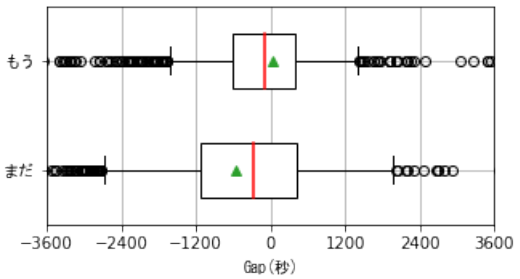

図 5　@aspect 別ギャップ箱ひげ図

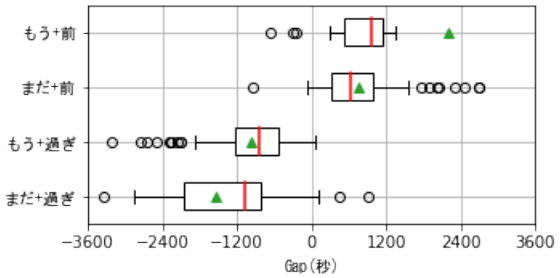

図 6　@aspect+@mod 別ギャップ箱ひげ図

4.4.2　使用時刻分布の比較　―「前」と「過ぎ」―

上記データの中で、さらに@mod 修飾表現の「前」および「過ぎ」を含むデータを抽出し、同様の調査をおこなった。この結果の箱ひげ図を図 6 に示す。「もう *h* 時前」のツイート全体の *Gap* 値の平均値は 2, 189 秒、中央値は 957 秒、「まだ *h* 時前」では平均値は 761 秒、中央値は 627 秒、「もう *h* 時過ぎ」では平均値は – 981 秒、中央値は – 851 秒、「まだ *h* 時過ぎ」では平均値は – 1, 551 秒、中央値は – 1, 075 秒であった。先述の@aspect 別のギャップ分布の結果に基づき、「もう」がリアルタイム性を保証するマーカーで、「まだ」が約 5 分後を示すマーカーだと見なすと、「*h* 時前」は *h* 時の約 15 分前に使われる一方で、「*h* 時過ぎ」は約 15 分後に使われるといえる。

5　おわりに

本稿では、TimeML の<TIMEX3>の仕様に基づいて定性的時間表現の概念を整理した上で、複合型定性的時間表現である「もう *h* 時」および「まだ *h* 時」を対象に時間ギャップを測定し、その分布を比較分析した。今回の調査結果は、たとえば「『もう 9 時だよ』よりも『まだ 9 時だよ』のほうが 9 時からギャップのあるタイミングでシステム発話をしてもよい」や「9 時 15 分に『もう 9 時過ぎ』という発話は自然であるが、9 時 30 分に同じ発話をするのは自然さに欠く」などの知識として自然な対話システム構築に活かし得る。ただし今回の調査結果を十分な知識へと昇華させるためには、対象表現を「もう *h* 時」「まだ *h* 時」以外の定性的時間表現にも拡張した更なる調査が必要である。また、Twitter ユーザが時間を認識してから実際にツイートを投稿するまでに自然と生じるタイムラグの影響も今後考慮するべきである。

参考文献

[1] Apple. Siri ならこんなことも - apple サポート (日本). https://support.apple.com/ja-jp/HT208336, 2021. (参照 2021-12-22).

[2] James Pustejovsky, José M Castano, Robert Ingria, Roser

Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. Timeml: Robust specification of event and temporal expressions in text. **New directions in question answering**, Vol. 3, pp. 28–34, 2003.

[3] 小西光, 浅原正幸, 前川喜久雄. 『現代日本語書き言葉均衡コーパス』 に対する時間情報アノテーション. 自然言語処理, Vol. 20, No. 2, pp. 201–221, 2013.

[4] 成澤克麻. 自然言語処理における数量表現の取り扱い. Master's thesis, 東北大学, 2014.

[5] Ali Hürriyetoglu, Nelleke Oostdijk, and Antal Van den Bosch. Estimating time to event from tweets using temporal expressions. In **Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)**, pp. 8–16, 04 2014.

[6] 気象庁. 予報用語 時に関する用語. https://www.jma.go.jp/jma/kishou/know/yougo_hp/toki.html. (参照 2021-12-22).