6P-5

Tweetに出現する接続表現を手がかりにした特異な関係の抽出とその評価

斉藤 博文† 山田 剛一† 絹川 博之†
東京電機大学大学院 未来科学研究科†

1. はじめに

Twitter には人々の日々の発言が溢れている。これらを分析することにより、いまの時代の人々が、物事をどう捉えているのかを明らかにすることができる。本研究では、モノとモノとの関係が人々にどう捉えられているかに注目し、特にその関係が接続表現によって表される場合について分析をする。例えば、一昔前は「韓国といえばドラマ」であったが、今では異なる関係が tweet されている。

接続表現には「ならば」「といったら」「なのに」など多くの種類があり、それぞれ表す関係が異なる。ここでは、因果関係のような基本的な関係に対し、特異である関係を表す接続表現を取り上げる。例えば、「のくせに」という接続表現には非難の気持ちが込められている。「小学生のくせにワックスつけてる」という tweet からは,「小学生」が「ワックスつけてる」ことが反発の対象になっていることがわかる。

特定の接続表現の前後に現れる語句の組を抽出し，現在取り上げられている関係を収集するシステムを構築し、収集実験を行った。

2. Tweetに現れる語句間の関係

接続表現を用いて語句間の関係を分析する研究として，特定の用法の接続詞を含む文からの因果知識の自動取得の方法を検討した研究[1]がある．この研究では，接続詞「にもかかわらず」を利用した因果知識の獲得を行い，接続詞「なのに」においても同様の変換処理を適用する手法が提案されている．本研究では特定の接続表現を含む tweet を手がかりに，因果関係のような基本的な関係ではなく、特異な関係にある語句を取り出す．例えば，語句 A と語句 B の関係が「A にしては B」と表現されたとする．このとき，A と B の組の関係を抽出し、通常では得られにくい特異な関係を得る．

2.1 特異な関係の種類

特異な関係にある語句は，接続表現により結び付けられていることが多い．接続表現の種類は多く，それにより様々な関係が表現される．今回は，特徴的な傾向が現れた「のくせに」，「にしては」，「といったら」，「だけあって」に関し抽出する．

(1)「（の）くせに」，「くせして」は，前件から予期されることに反する事柄が後件として起こることを，前件の主体に対する非難や反発の気持ちをこめて示すものである[2]．特に，「のくせに」という接続表現は，「なのに」と比較して，非難する気持ちが強い．

(2)「にしては」は，条件と結果を比較し，結果が予想や標準を上回るか下回るかしたことを表すものである[2]

(3)「といったら」,「といえば」，「というと」は，その場の誰かが既に話題にしていたり，自分が心の中で思い浮かべていたりした事柄を積極的に自分から引き取って題目化し，それをきっかけに関連事項を述べていくといった表現である[2].

(4)「だけあって」は，後件の事実が生じると思われる理由付けを前件に求める表現である[2]．今回は、この意味で用いられることが 9 割以上である「なだけあって」に着目する．

2.2 Twitter の特性

Twitter では日常会話が行われているため，ブログなどに比べて推敲されておらず，軽い気持ちで投稿されている．「のくせに」という表現がブログなどで使われることは少ないが，Twitter 上では多く使われている．他方、新聞記事においては「のくせに」はほとんど使われていない。比較のために毎日新聞を例に上げる．Web 上で公開されている毎日新聞の記事[3]の中で，「のくせに」が含まれている記事は 1 週間の中で 1 個程しか存在していない．それに対して Twitter では，「のくせに」を含む tweet の数は，一日に約 1 万 tweet 程投稿されている．Twitter を使用することで，日常会話でしか使わないであろう接続表現を含む文を，簡単に得ることができる．

3. 関係収集システム

あらかじめ定めた接続表現を含む tweet を検索し、組となる語句を抽出・蓄積するシステムを構築した。

3.1 関係の要素である語句の特定

接続表現による tweet の検索を行い，検索結果の tweet から，対象外 tweet の削除，および，接続表現を含む文だけを処理対象とする処理を行う．

(1) 対象外 tweet の判定

接続表現を含んだtweetの中で，本システムでは扱わないtweetを以下の条件によって判定する．

・指示語が用いられている場合.
・文が接続表現で終わっている場合.
・語句 A が名詞句でない場合.
・語句 B が名詞句あるいは動詞句でない場合.

(2) 接続表現を含む文の特定

Tweet の文字列を、文末などに使われる記号を手がかりに文に分割し、接続表現を含む文だけを残し、他の部分を削除する。「あれ？ゴミのくせにカワイイ…！？」というtweetの場合は，接続表現の前には「ゴミ」が残り，接続表現より後には「カワイイ」が残ることになる．

これらの処理の後，構文解析を行うことによって，接続表現に係っている語句 A，接続表現の係り先である語句

Extraction and Evaluation of Singular Relations Based on Connection Expression in Tweets
†Hirofumi Saito, †Koichi Yamada, †Hiroshi Kinukawa
†Graduate School of Science and Technology for Future Life, Tokyo Denki University

B を得る．形態素解析に MeCab（和布蕪）[4]を用い，構文解析に CaboCha(南瓜) [5]を用いる．

3.2 取り出す関係の絞り込み

取り出した語句 A および語句 B には，関係の要素として扱うべきではないものが含まれている．

(1) 語句 B に語句 A が含まれている場合，例えば，「無理といったら無理なの！」という tweet の場合はトートロジーであり、取り出すべき関係ではない．

(2) 語句 B に自立語がない場合は，語句 B のみでは意味を表さないため，取り出す語句としてふさわしくない．また，語句 A または語句 B が代名詞である場合，一般的に成り立つ関係ではないことが多いため取り出さない．

4. 評価

4.1 関係の抽出

接続表現を含む tweet を対象とした関係の抽出実験を行った．「のくせに」，「にしては」，「といったら」，「なだけあって」の 4 つの接続表現で tweet 検索を行い，得られた各 500 tweets を対象とした．ただし，「なだけあって」については Twitter の検索の仕様上，得られた tweet 数が少ないため，100 tweets を対象とした．接続表現別の精度・再現率を**表 1** に示す．ここで， 精度は，システムが関係として出力したうち，実際に抽出すべき関係であった割合である．また，再現率は，抽出すべき全ての関係の中で，システムの出力の中に含まれていた関係の割合である．なお, 抽出すべき関係にある語句が出力に含まれていれば, 余分な語まで取り出していても正解としている．

表 1． 関係の抽出結果

表現	関係を含む tweet	システムの出力	精度	再現率
のくせに	321	179	79.3%	44.2%
にしては	119	114	79.2%	33.2%
といったら	119	114	72.4%	32.2%
なだけあって	94	47	74.5%	37.2%

4.2 関係の抽出誤り要因

関係を得られない要因として，形態素解析の誤りが挙げられる．取り出した関係の誤りのうち，約2割が形態素解析の誤りである． また，固有名詞が形態素解析の辞書にないことも問題に挙げられる．他の要因としては，tweetをうまく文ごとに切り分けることができていない場合がある．Twitterの特徴により，砕けた文章が多く，うまく文ごとに切り分けられていないのが現状である．この場合，構文解析をうまく行えずに誤ってしまう．なお，特異な関係のないtweetから関係を取り出してしまうことが約2割存在する．

5. 得られた関係

接続表現から得られた関係について，どのような関係が得られたかを調査した．また，接続表現別の関係の例を**表 2** に示す．

5.1「にしては」から得られた関係

「にしては」から得られた関係は，結果が予想や標準を上回ることを表しているものが約 6 割であった．その反対の，下回る事を表しているものは約 2 割であり，前者の使われ方が多いことが分かる．また，残りの約 2 割は，「涙を無駄にしてはいけない」のように，「～にして（させて）はいけない」といった使われ方をしている．

5.2「といったら」から得られた関係

「といったら」から得られた関係の中には，ストレートな連想ではなく多少のひねりを効かせた関係が約 2 割混在していた．「寿司といったらコーン軍艦だろ」といった，アイロニーとも解釈できるような tweet があった．これは，Twitter の特徴として，ウケを狙ったネタを取り入れた投稿を行うことがあるからであると推測できる．これを踏まえると，「といったら」を含む tweet からは，よく連想されがちな連想が得られるとは一概には言うことができない．

表 2． 得られた関係の例

のくせに	赤道付近 - 氷河がある
	おっさん - 泣きたい
にしては	チワワ - 胴が長い
	貴族 - 気品に欠ける
といったら	セーラー服 - ツインテール
	下北 - 893
なだけあって	最年長 – 料理が得意
	初出勤 – 精神的に疲れた

6. おわりに

接続表現を用いて表される語句間の関係の抽出を行った。今回は「のくせに」，「にしては」，「といったら」，および「なだけあって」という接続表現を使用した．「といったら」を含む tweet では主として，よく連想されがちな連想が得られた．「のくせに」を含む tweet では，通常では考えがたいような，思いがけない関係が得られた．今後の課題として，接続表現の持つ代表的な意味ではなく例外的な意味を持つ関係を検出する手法を構築することが挙げられる．

謝辞

本研究で使用した MeCab，CaboCha を開発された方々に深く感謝いたします．

参考文献

[1] 今給黎勇佑, 石川勉 ， “特定の接続詞の意味特性を利用した電子化文書からの因果知識の獲得方法”， 情報科学技術フォーラム講演論文集，8(2), 539-540 (2009).

[2] 森田良行，松木正恵，“日本語表現文型 用例中心・複合辞の意味と用法”，株式会社アルク(1989).

[3] 毎日新聞，http://mainichi.jp/.

[4] MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

[5] CaboCha, https://code.google.com/p/cabocha/