E-003

Twitter のツイートの接続表現を手がかりにした連想関係の抽出
Extraction of Associative Relations Based on Connective Expression on Tweets

斉藤 博文† 山田 剛一† 絹川 博之†
Hirofumi Saito Koichi Yamada Hiroshi Kinukawa

1. はじめに

日常生活において，ある事柄から何かを連想することにより，考えを進めることがある．その連想の幅を広げることで，ものの捉え方，考え方を変えることができる．現在発言されている連想関係を取り出して分析することにより，豊かな連想を行う手助けができるようになると考えられる．Twitter における投稿は tweet と呼ばれ，tweet の文字数は 140 文字以内である．tweet に挙げられる話題には，現代における話題が反映されており，その時々による話題の違いが存在する．ここでは tweet 内で用いられる「といったら」などの接続表現を手がかりに，その前後に現れる話題の語句を抽出する．「といったら」という表現からはその時々に連想される話題が得られる．例として，「寿司といったらマグロですよ♪」という tweet では，「寿司」に対して「マグロ」という話題を得る事ができる．「のくせに」という表現からは非難の話題を取り出すことができる．例えば，「犬のくせにコタツに入る」という tweet からは，「犬」に対して「コタツに入る」という話題を得られる．接続表現の前後に現れる 2 つの話題の組を連想関係とし，現在取り上げられている連想関係を取り出す．連想関係が存在している連想元となる語句が入力された時に，連想関係にある語句を提示することで，豊かな連想を行う手助けを行うことを目標に研究を行った．

2. tweet に現れる連想関係

接続表現を用いて語句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取得を行い，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，連想する語句を取り出す．例えば，語句 A と語句 B の関係が「A といったら B」と表現されたとする．このとき，A と B の組を連想関係にある組みとする．また，「のくせに」から取り出される A と B の組も連想関係として扱う．

2.1 連想関係の種類

連想関係にある語句は，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．「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]．また，「（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．

†東京電機大学大学院 未来科学研究科
Graduate School of Science and Technology for Future Life, Tokyo Denki University

2.2 Twitter の特徴

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．比較のために毎日新聞を例に上げる．WEB 上に挙げら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. 連想関係収集システム

接続表現による tweet の検索を行い検索結果の tweet から，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．その後，連想関係にある語句を取り出し蓄積する．

3.1 連想関係の特定

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 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．連想関係の特定は，以下の方法によって行う．

(1)余分な tweet の削除

手がかり表現を含んだtweetの中で，本システムでは扱わないtweetを以下の方法によって削除する．

・指示語が用いられている場合

指示語が用いられているtweetでは，指示語の指示対象が明らかでないことが多いため，抽出の対象としない．

・文が接続表現で終わっている場合

接続表現で終わっている場合は，連想関係にある語句を得ることができないため，取得を行わない．ただし，倒置表現によって連想関係がある語句が述べられているtweetも存在するが，処理を行う上で判定が困難なため，同様に取得しない．

・連想関係となる語句の品詞

多くの場面で成り立つ(と誤解されている)関係を扱うため，語句 A は名詞句，語句 B は名詞句あるいは動詞句の場合を扱う．

(2)余分な語の削除

接続表現を含む文だけを処理対象にする際に，以下の場合は語の削除を行う．

・記号が使われている場合

接続表現を起点に前方および後方に向けて走査し，記号が現れる直前までを残し，それより先を削除する．これにより文末の記号が削除され，tweetが複数の文からなる場合には接続表現を含む文のみが残る．例えば，「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．また，「ゲイのくせにイケメンww」のように「w」が含まれている場合，「w」は「（笑）」の意味を表わす記号として扱う．

3.2 構文解析による語句の特定

先に挙げた既存研究[1]では抽出する語句内の修飾語句を限定しているが，本研究では，修飾語句を全て含めた形で語句 A および語句 B を取り出す．例として，「カラオケ行きたくなる曲といったらチキンライスが出てくる」という tweet の場合，語句 A「カラオケ行きたくなる曲」，語句 B「チキンライスが出てくる」として取り出す．

3.3 取り出す連想関係の絞り込み

取り出した語句 A および語句 B には，連想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

例として，「無理といったら無理なの！」という tweet の場合では強調の意味が含まれているが，連想の関係ではないので取り出さない．

(2) 語句 A または語句 B が要素として相応しくない場合

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は，多くの場面で成り立たないため取り出さない．

4. 評価

4.1 連想関係の抽出

接続表現を含む tweet を対象とした連想関係の抽出実験を行った．「といったら」と「のくせに」の 2 つの接続表現で tweet 検索を行い，得られた各 200 tweet 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，システムが連想関係として出力したうち，実際に抽出すべき連想関係であった割合である．また，再現率は，抽出すべき全ての連想関係の中で，システムの出力の中に含まれていた連想関係の割合である．

表 1． 連想関係の抽出結果

表現	連想を含む tweet	システムの出力	精度	再現率
のくせに	59	90	48.9%	74%
といったら	60	57	66.7%	63.3%

4.2 連想関係の抽出誤り要因

連想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．取り出した連想関係の誤りのうち，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．なお，連想関係のないtweetから連想関係を取り出してしまうことが約2割存在する．

5. 連想関係の分析

5.1「といったら」から得られた連想関係

「といったら」から得られた連想関係の中には，多少のひねりを効かせた連想関係が約 2 割混在していた．例として，「寿司といったらコーン軍艦だろ」といった tweet などである．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．

5.2「のくせに」から得られた連想関係

「のくせに」から得られた連想関係の中には，品のない表現を含む連想関係が混在していた．比較として， ブログの中でも日常会話に近い文が使われている Ameba ブログ[5]の記事の「のくせに」を含む文と比較を行った．どちら記事の中にも，「男のくせに」，「女のくせに」，「日本人のくせに」などといった文は存在しているが，Twitter 上では，「のくせに」を含む文の中で，品のない表現を含むが占める割合が，約 8％である．これは，Ameba ブログに比べて約 4 倍である．

6. おわりに

接続表現を元に，tweet に現れる連想関係の検出を行った．今回は「といったら」および「のくせに」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，思いがけない連想が得られた．今後の課題として，接続表現による違いを分析し，連想関係にある語句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ｊｐ，http://mainichi.jp/.

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

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

[6] Ameba（アメーバ)ブログ，http://ameblo.jp/.