E-010

Tweet に出現する接続表現を手がかりにした関連話題の抽出
Extraction of Related Subject based on Connectives Expression in Tweets

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

1. はじめに

会話や文章の中では，一般に自然なことや常識的なことは発言されにくい．例えば，休日には会社に行かないことや，夜は暗いことなどである．本研究では，Twitter を活用し，このような自然なことを取得することで，知識の獲得に活用していくことを目標に研究を行っている．知識獲得には順接および逆接を表す接続表現を用いて行う．「のくせに」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．「のくせに」という表現からは予期されることに反することを表す話題を取り出すことができる．例えば，「男のくせに料理できるんだ」という文からは，「男」に対して「料理できる」という話題を得られる． このような予期されることに反することを表す接続表現を元にして，一般に自然な関係にある語句の取得を行う．Twitter における投稿は tweet と呼ばれる．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[1]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 1,000 tweets 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．

2．tweet に現れる関連話題

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[2]がある．この研究の手法は，「にもかかわらず」の前文(P)と後文(Q)から P→¬Q の形の知識を述語論理形式で獲得するものである．接続詞「なのに」においても同様に変換処理を適用することで因果知識を得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，関連の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 の組を関連話題とする．接続表現ごとに関連話題の A と B の関連性は変化し，一般的に思われている，あるいは一般的には考え難いような関連話題が得られる．

2.1　関連話題の種類

接続表現の種類は多く，関連話題は接続表現によって A と B の関連性が変化する．典型性や一般性を明確に表している接続表現および予期されることに反することを表す接続表現を用い，知識の獲得を行う．

2.2　順接からの知識獲得

典型性や一般性を明確に表している接続表現から得られた関連話題は，知識として扱う．順接の例を以下に挙げる．

(1)「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[3].

2.3　逆接からの知識獲得

予期されることに反することを表す接続表現から得られた関連話題の語句 A を，対となる語句 P に変換した場合，一般に自然な関連の語句が成立すると考えられる．逆接の例を以下に挙げる．

(1)「（の）くせに」，「くせして」は，語句 A から予期されることに反する事柄が語句 B として起こることを，語句 A 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[3]．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[3]

3. 関連話題収集システム

接続表現による tweet の検索を行い，検索結果の tweet から，不要な 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，関連話題を取り出し蓄積する．

3.1　関連話題の特定

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．関連性を持った語句を得るため，接続表現を含んだ tweet の中で，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．

3.2　取り出す関連話題の絞り込み

取り出した語句 A および語句 B には，関連話題の要素として扱うべきではないものが含まれている．取り出さない場合を示す．

(1) 強調の場合

語句 B に語句 A が含まれている場合において，例えば，「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で繰り返されており，関連話題ではない．

(2)自立語がない場合

語句 B に自立語がない場合は，語句 B のみでは意味をなさないため，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，指示対象の特定ができない場合が多いため扱わない．

†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

4. 評価

関連話題の抽出結果および，逆接からの知識獲得の結果についての評価を行った．

4.1 関連話題の特定

接続表現を含む tweet を対象とした関連話題の抽出実験を行った．「のくせに」，「にしては」，「といったら」3 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．接続表現別の精度・再現率を**表 1** に示す．ここで， 精度は，システムが関連話題として出力したうち，実際に抽出すべき関連話題であった割合である．また，再現率は，抽出すべき全ての関連話題の中で，システムの出力の中に含まれていた関連話題の割合である．なお, 関連話題にある語句が出力に含まれていれば，必要のない語が付随していても正解としている．

表 1． 関連話題の抽出結果

接続表現	関連話題を含む tweet 数	システムの全出力数	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%

4.2 逆接からの知識獲得

予期されることに反することを表す接続表現から得られた関連話題の語句 A を，対となる語句 P に変換した場合，一般に自然な関連の語句が成立することが言える場合が多いことがわかった．例えば，「だけど」の関連話題で「雨」-「ドライブしてくる」がある時に，語句 A の対義語 P である「晴れ」とすることで，「晴れ」-「ドライブしてくる」が一般に自然な関連話題として成立する．その結果を**表 2**，**表 3** に示す．ここでは，接続表現ごとの語句 A は対にすることができ，得られた語句の数が多い語句を選定している．

表 2. 関連話題の例

語句A	接続表現a	語句B
「雨」	だけど	花火大会やってる，屋上で一杯
「平日」	なのに	混んでる，お休みだ
「男」	のくせに	化粧してる，料理出来る

表 3. 語句 A を対語句 P に変換した時に自然である確率

語句A	接続表現 a	A+aの出現頻度	Aの対となる語P	P→Bが自然である確率
「雨」	だけど	6 2	「晴れ」	60/62
「平日」	なのに	4 3	「休日」	43/43
「男」	のくせに	4 1	「女」	39/41

4.3 関連話題の抽出誤り要因

関連話題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．取り出した関連話題の誤りのうち，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ことがある．Twitterの特徴により，砕けた文章が多いためで，この場合，構文解析をうまく行えずに話題の抽出を誤ってしまう．

5. 得られた関連話題

関連話題を取得する中で，特徴的な傾向のある接続表現が見られた．そのため，接続表現から得られた関連話題に，どのような関係が得られたかを調査した．

5.1「にしては」から得られた関連話題

「にしては」から得られた関連話題は，結果が予想や標準を上回ることを表しているものが 146/200 であった．その反対の，下回る事を表しているものは 54/200 であり，前件の使われ方のほう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，「前に述べたようにして（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」から得られた関連話題

「といったら」から得られた関連話題の中には，典型でないとわかっていながら，あえて違った関係を発言している tweet 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．ただし，関連話題の数が多い場合には，同じ語句 A 同士を比較し，典型性を見極めることが出来るかもしれないが，得られる関連話題の数は少なく，この検出は困難である．

6. おわりに

関連話題の取得および，一般に自然な関係をもつ関連話題の取得を行った．関連話題が得られた場合は，語句 A を対の語句に変換することで一般に自然な関連話題を得やすいことがわかった．ただし，関連話題の抽出の漏れが多く，得られる関連話題の数が少なくなってしまっている点は今後の改善すべき点である．

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] 毎日新聞，http://mainichi.jp/.

[2] 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).

[3] 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).

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

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