3Q-03

Twitterを用いた皮膚障害を引き起こす製品名等の抽出

阿部健一†　　吉田博哉†

神戸情報大学院大学†

1. はじめに

近年，美白成分ロドデノールが入った化粧品によって皮膚に白斑が生じた問題[1]や，小麦抽出成分の入った石鹸を長期間使用したことによる重度小麦アレルギーの発生[2]など，身近な製品による皮膚障害が発生している．これらの被害は各地の消費生活センターを通じて国民生活センターの消費者情報ネットワーク「PIO-NET」に収集されているが，当局に被害が認知され分析が開始されるまでに時間がかかっている．

このような情報収集と分析の遅れに対し，インターネット上で直近の情報を機械的に収集し，分析することで時間を短縮し，被害の拡大を防ぐ手法が考えられる．本研究では Twitter を用いた情報収集と既存の不具合情報を取集する研究[3]とを組み合わせ，皮膚障害を引き起こす製品名やその商標・企業名（以下「製品名等」）を抽出し，その有効性を検証する．

2. コーパスの作成

2.1. コーパスの必要性

国民生活センターから皮膚障害を起こすと公表されている製品名を元に tweet を収集したところ，消費者被害の訴え（以下「被害情報」）と製品名等が複数の tweet に分かれるケースが存在したため，被害を受けたユーザごとの tweet を収集したコーパス（以下「皮膚障害コーパス」）を作成する．

2.2. コーパス作成システム

皮膚障害コーパスを作成する「コーパス作成システム」は，**図 1** に示す通り「危険表現取得サブシステム」と「ユーザ選定サブシステム」という２つのサブシステムで構成される．

本システムを開発するにあたり，国民生活センターが公表している皮膚障害に関する被害報告を分析したところ「皮膚」という単語が直接現れることは少なく，「太股が痛い」「腕がヒリヒリする」など皮膚が存在する「身体部位」とその被害に関する動詞・形容詞が組み合わされていることが分かった．このため，危険表現

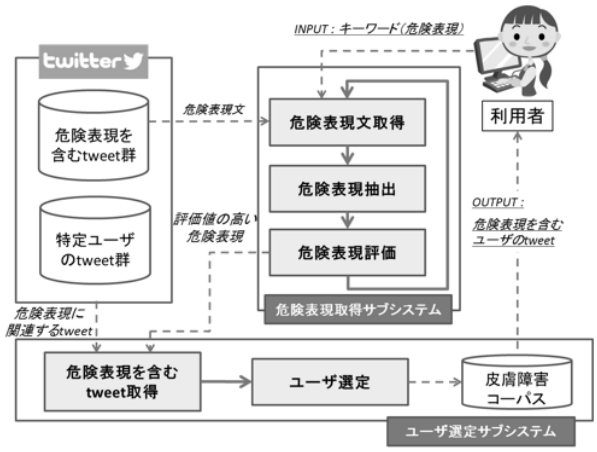

図 1　コーパス作成システム

取得サブシステムでは不具合情報の研究で用いられてきた「信頼度スコアリング[3]」を用いて被害報告に特徴的な皮膚の存在する身体部位を「外部身体部位表現」として機械的に収集し，得られた身体部位と「被害を訴える動詞・形容詞」を用いて危機表現を抽出する．

続いて，ユーザ選定サブシステムではユーザ自身が皮膚障害を訴えている可能性の高い tweet を人手で少数抽出し，残りの tweet を TF-IDF を元にロジスティック回帰で分類し，更にそこから人手で tweet を選びユーザを決定したのち皮膚障害コーパスを作成する．

2.3. 対立コーパスの作成

このように作成したコーパスから特徴的な単語を抽出する場合，もともと出現頻度の高い単語を分析対象から外さねばならない．このため，比較のためのコーパス（以下「対立コーパス」）を構築し語彙を比較することで皮膚障害コーパスに特徴的な製品名等を抽出する．対立コーパスは無作為に選んだ日本語使用者のユーザ ID を用いて同程度の tweet 数を収集する．

3. 製品名等の抽出とコーパスの比較

皮膚障害コーパス及び対立コーパスから製品名等を抽出するために，まず固有表現を抽出する．Twitter 上で使用されている特徴的な商標や企業名を含む固有表現を抽出するため，汎用的な辞書ではなくネット上で使用されている単語を大量に収集した大規模辞書 mecab-ipadic-

Extraction Information of Products Which Cause Skin Disorders by Twitter
Kenichi Abe, Hiroya Yoshida
Kobe Institute of Computing

NEologd[4]を用いた．両コーパスに対し，形態素解析エンジン MeCab で形態素解析された単語のうち，品詞が「名詞」，品詞細目が「固有名詞」のものを分析の対象とする．

製品名等を特徴語として抽出するため，対立コーパスに該当の単語が存在しなかった単語ほど評価を高くし，対立コーパスに出現する頻度が多かった単語ほど評価が低くなるよう，単語の出現数を抽出された固有名詞の総数で割ることで正規化を行ったものを評価値とし，対立コーパスに該当の単語が存在しなかった場合に 0 の評価値を与え，減算したものを用いる．

4. 評価実験

本研究の有効性を確認するために，一定期間収集した情報をもとに分析を行った．なお，収集した tweet 数，ユーザ数を**表 1** に示す．

表 1 収集データ

	皮膚障害コーパス	対立コーパス
tweet 数	10,458,393	10,442,634
ユーザ数	5,178	6,978

これらの収集データのうち，出現頻度の少ない単語は特徴語ではないので評価から外す．今回は皮膚障害コーパスにおいて 10 回以上出現した固有名詞が上位約 20%にすぎないことから，10 回以上出現した固有名詞のみを評価の対象とした．皮膚障害コーパスにおいて 10 回以上出現した固有名詞は 73,278 語であった．

そして，皮膚障害コーパスで 10 回以上出現した固有名詞の評価値に対し，対立コーパスにおける同じ固有名詞の評価値を引いて得られた値を元に降順で一覧を作成し，その後人手で製品名等の抽出を行った．抽出された製品名等のうち，上位 20 位を**表 2** に示す．

5. 考察

製品名等を抽出したが，20 位以内には製品の商標・企業名は入らなかった．

抽出された製品名等を元の tweet で確認したところ，1 位の「スマホ（スマートフォン）」はスマホの熱による低温火傷の報告があった．2 位の「お酒」は飲酒による痒み・湿疹の報告があった．「ウィッグ」使用時のかゆみや，「カラコン（カラーコンタクト）」使用時の目の痛みなども tweet されている．また，「ネイル」「DVD」「CD」は「ウィッグ」「カラコン」と共に「コス」「レイヤー」といった単語と共起していることが判明した．Twitter 上で皮膚障害を報告しているユーザには「コスプレ（コスチュームプレイ）」と呼ばれるアニメやゲームなどの登場人物に扮する行為を好む人がおり，コスプレの過程で皮膚障害を発症している可能性が高いと考えられる．一方で，3 位の「チャリ（自転車）」は，自転車使用後の痒み・湿疹の tweet が確認できたものの，こちらは製品による被害とは言い難い内容であると言える．

表 2 抽出された製品名等

1	スマホ	11	缶バッジ
2	お酒	12	洗濯機
3	チャリ	13	掃除機
4	DVD	14	いちご
5	ネイル	15	ノート
6	ウィッグ	16	唐揚げ
7	カラコン	17	アラーム
8	イヤホン	18	日焼け止め
9	CD	19	化粧品
10	ケータイ	20	ティッシュ

6. まとめ

本研究では，消費者被害，特に皮膚障害の被害拡大を防ぐため，Twitter とテキストマイニングを用いたシステムを開発した．具体的には，Twitter から皮膚の障害を訴えるユーザを特定し，その一連の tweet を収集した皮膚障害コーパスと，比較を行うための対立コーパスを作成した．そして，皮膚障害コーパスと対立コーパスから固有表現を抽出して比較し，製品名等を抽出するシステムを提案した．

実験の結果，実際に皮膚障害を引き起こしている製品名が得られたが，直接的な障害を与えていない製品名も取得されてしまっている．

今後は PIO-NET 上の情報との比較や，機械学習を用いた製品名等の抽出を行っていきたい．

参考文献

[1] 平郡，他：加水分解コムギ含有石鹸の使用後に発症した小麦依存性運動誘発アナフィラキシーとその経過について，アレルギー，vol.60, no.12，pp. 1630-1640，2011.
[2] 最上：化学物質による白斑―職業性白斑の機序とロドデノール白斑―，国立医薬品食品衛生研究所報告，vol.133，pp.13-20，2015.
[3] 栗原，嶋田：ブートストラップ法を用いた Twitter からの不具合文抽出，言語処理学会第 21 回年次大会発表論文集，pp.341-344，2015.
[4] https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md