Twitterからの自動車の不具合情報抽出に関する研究

†北林 智治　　‡酒井 浩之　　‡増山 繁

†豊橋技術科学大学 知識情報工学課程

‡豊橋技術科学大学　大学院工学研究科　情報・知能工学専攻

kitabayashi@la.cs.tut.ac.jp, sakai@tut.jp, masuyama@tut.jp

1　はじめに

自動車の不具合（以下「**不具合**」とする）は，事故等の原因となり，社会において損失をもたらすため，その情報を探すことは，事故を未然に防ぐために重要である．

不具合情報を探すことに関連する研究として，酒井ら [1] の研究がある．しかしながら，これに倣って新聞など一般のメディアのみを不具合情報の情報源とした場合，情報の発信者がその真偽を確かめるために，情報を公開する時期を遅らせることがあるため，迅速に対応できない可能性がある．また，そもそも一般のメディアには出現しない情報がある場合がある．それに対して，Twitterなどの個人が情報を発信できるメディアは，発信者の身の周りで起きたことを個人的に発信するため，それらの欠点を克服しうる．

Twitterは,「今行っていること」を140文字以内で発信するコミュニケーション・サービスであり，多くのユーザにより大量の情報が発信されている．具体的には，日本人のユーザ数は，2010年9月の時点で1,113万人[1]で，日本人の一ヶ月あたりの総ツイート（発言）数は，2010年8月の時点で2億8千万件[2]である．従って，周知でない個人的な経験に基づく情報が含まれる可能性があると考えられる．例えば，企業がTwitterによるマーケティングを行うことの支援をするサービス[3]が存在する．一方で，周知でない個人的な経験に基づく不具合情報もTwitter上に存在すると考えられるため，本研究では，知識源としてTwitterを用い，不具合情報の抽出を行う．

[1]ITmedia　オルタナティブ・ブログ. mixi, Twitter, Facebook 2010年9月最新ニールセン調査 ～ Twitterが1100万人超、Facebookも200万人超, http://blogs.itmedia.co.jp/saito/2010/10/mixi-twitter-fa.html

[2]MarkeZine（マーケジン）. 8月の総ツイート数は2億8千万件、猛暑やmixiのアクセス不具合、NHKのツイッター特集が話題に, http://markezine.jp/article/detail/11582

[3]ツイッターデータ分析サービス — クチコミ分析・ブログ／ツイッター分析サービス『感゜Report』（かんどれぽーと）, http://kandoreport.jp/twitter/twitter.html

2　提案手法

2.1　提案手法の概要

人手で不具合を示す表現を網羅するのは困難であるが，日常会話で使われる自動車の部品名を網羅するのは容易である．そこで，以下の手順で不具合情報の収集を支援する仕組みを作ることにする．

step 1. 部品名検索:
twitter検索[4]において，自動車の部品名をクエリとして検索する．

step 2. ツイートの保存:
検索結果として得られたツイートから，明らかに不要なものを取り除き，保存する．

step 3. 係り受け解析:
ツイートからTwitter特有の表現を取り除き正規化を行った後，係り受け解析を行う．

step 4. 文節の組の列挙:
部品名を含む文節と，それが係る文節の組み合わせ（以下「**文節の組**」とする）を出現頻度順に列挙する（以下「**文節の組の列挙**」とする）．

step 5. 不具合を表す文節の組の決定:
出現頻度が高く，かつ，人手で見て不具合情報を含むツイートに高確率で含まれる文節の組を，不具合を表す文節の組とする．

step 6. 不具合情報の抽出:
不具合を表す文節を含むツイートで「3.2節に示すような特定の語が出現する，しない」などの条件にあてはまるものを，不具合情報が含まれるツイートとする．□

[4]twitter検索，http://yats-data.com/yats/

2.2 自動車の部品名

本研究において、検索クエリとした自動車の部品名は，車の豆知識.com[5] を参考にし，ハンドル，エンジン，ブレーキ等，50 種類を定めた．一覧を表 1 に示す．

表 1: 検索クエリとした自動車の部品名一覧

ハンドル	エンジン	ブレーキ
タイヤ	ボンネット	バンパー
フロントグリル	ドア	ピラー
スポイラー	ヘッドライト	フォグランプ
ハロゲンランプ	キセノンライト	テールランプ
スタッドレスタイヤ	メーター	シフトレバー
レシプロエンジン	ロータリーエンジン	インパネ
ディスクブレーキ	ドラムブレーキ	ABS
バイアスタイヤ	コンフォートタイヤ	シート
テンパータイヤ	ディスクホイール	

ハザードランプ	コーナーリングランプ
ステアリングホイール	アクティブサスペンション
ディーゼルエンジン	ハイブリットエンジン
DOHC エンジン	SOHC エンジン
スーパーチャージャー	ストラットサスペンション
マルチリンクサスペンション	エアーサスペンション
リーフスプリング	コイルスプリング
パーキングブレーキ	エンジンブレーキ
多気筒エンジン	ラジアルタイヤ
ターボチャージャー	ショックアブソーバー
ダブルウィッシュボーンサスペンション	

2.3 明らかに不要なツイートの除去

不具合と無関係なツイートとして，bot によるツイートや，Twitter ボタンによるツイート，コメントの無いリツイートがある．

bot は，自動的にツイートを発信するプログラムである．bot によるツイートは，以下の二つの理由により除去を行う．

理由 1．同じ表現を用いて多くのツイートを発信することにより，文節の組の列挙の際，少数の文節の組が高頻度で出現してしまうため．

理由 2．事実とは異なることをツイートする可能性が高いため．

bot によるツイートの除去は以下の方法で行う．各ツイートには，その発信者を示す screen_name が付随している．Twitter BOT JAPAN[6] に登録された全ての bot の screen_name を取得し，その screen_name を持つツイートを除去する．しかしながら，Twitter BOT JAPAN に全ての bot が登録されているわけではない．そこで，理由 1 による悪影響を抑えるため，文節の組の列挙の際，同一の screen_name を持つユーザーから取得した全ツイート中で，まったく同じ文節の (部品名，係先が共に一致する) 組が複数回出現した場合，その出現回数は一回とする．

Twitter ボタン (以下「t ボタン」という) によるツイートとは，ニュースサイトにある「tweet」もしくは「t」と書かれたボタンをクリックし，発信するもののことである．この場合，ニュースの概要が，自動的に本文に含まれる．本研究では，ニュースサイトに載るような事象は抽出しない．その理由は，Twitter を使わなくても，他のサイトを利用して抽出できるからである．また，t ボタンによるツイートが複数存在しても，元の事象は一つであるため，多くの事象に使われる表現でなくても，文節の組の列挙の際，上位となる危険性がある．t ボタンを設置するサイトは，自らのサイトを宣伝する目的があるため，本文にサイトの URL を含む可能性が高いと考えられる．そのため，本研究では，URL を本文に含むツイートを除去する．ただし，URL を含むツイートが必ず不要なわけではない．

リツイートとは，あるユーザーのツイートを引用して自分のアカウントから発信することである．引用文を改変して発信されうることと，複数のツイートに対して元の事象が一つであることから，コメント（リツイートする側の意見）が無いものは除去した．

2.4 ツイートの正規化

ツイート中の文節の組を得るには，原文を文節で区切ることと，文節間の係り受け関係の情報が必要になる．本研究では，係り受け解析に CaboCha[2][7] を用いる．しかしながら，Twitter 上には独自の表現があるため，ツイートをそのまま係り受け解析した場合，うまく解析できない．そこで，本研究ではツイートを係り受け解析する前処理として，いくつかの正規化を行う．その内容と例を以下に示す．

- !、?、!、?の「。」への置換
  例 1 きゃー！車のエンジンがかからないおんぃ(¯д ¯;) ノ=3=3=3
  →きゃー。車のエンジンがかからないおんぃ(¯д ¯;) ノ=3=3=3
- w の「。」への置換
  例 2 車のエンジンがかからずコンビニで立ち往生 wwww人生 2 度目の JAF です。。。

[5]車の豆知識.com，http://m3106.com/car/008_00_breake.html
[6]日本の Twitter BOT まとめサイト：Twitter BOT JAPAN, http://bot.cuppat.net/
[7]CaboCha/南瓜: Yet Another Japanese Dependency StructureAnalyzer,http://chasen.org/ taku/software/cabocha/

→車のエンジンがかからずコンビニで立ち往生。人生 2 度目の JAF です。。。

- @(ID 名) の除去
  例 3 @w463a3_ 確実に壊れているのは右後席のドアです。中から開けられなくなりました。あと、エンジンストール。(以下省略)
  →確実に壊れているのは右後席のドアです。中から開けられなくなりました。あと、エンジンストール。(以下省略)

- ハッシュタグの除去
  例 4 雨の日は車の雨漏りが心配です (白目 #hbeat
  →雨の日は車の雨漏りが心配です (白目

@(ID 名) という表現を持つツイートは，書かれた ID を持つユーザーに対して呼びかける意味を持ち，そのユーザーがこの表現を持つツイートを優先して見る使い方ができる．ハッシュタグは，#(ハッシュ) の後にある事柄の名前を書いたものを指す．同じハッシュタグを含むツイートをまとめて見る使い方ができる．このように修正したツイートに対し，文節の組の列挙を行う．

3 評価実験

3.1 文節の組の決定

実験の前処理として，不具合を表す文節の組の検討を，二回行った．一回目は，2010 年 5 月に部品名をクエリとした検索で得た 101,150 件のツイートを係り受け解析した．出現回数が 3 以上の文節の組から，不具合を表しそうな文節の組を人手で判定し，35 種類の文節の組を得た．二回目は，同じクエリで 2010 年 11 月に検索を行い，同 5 月の分と合わせた 297,599 件のツイートから，出現回数 5 以上の文節の組を人手で判定し，新たに 23 種類を得た．一覧を表 2 に示す．

3.2 不具合情報抽出手法

不具合情報の抽出手法を検討するにあたって，不具合を表す係り受け関係になる文節の組を含むツイートを見て，正例（不具合情報を含むツイート）や負例（不具合情報を含まないツイート）であることの手がかりとなる表現を考えた．その理由は，文節の組を含むツイート全てを正例とすると，不具合情報と無関係なツイートが高い確率で抽出されるためである．例えば「車でブレーキ効かなくて事故りまくる夢を見た」というツイートには,「ブレーキ　効かなくて」という不具合を表す係り受け関係になる文節の組が含まれている．しかしながら,「夢で見た」ことを示すツイートであるため，このツイートは，本研究の抽出対象とならない．

表 2: 不具合を表す文節の組一覧

ブレーキ 利かなくなった」	ブレーキ 利かない
ブレーキ 壊れた	エンジン かからない
ブレーキ 効かない	エンジン かからない。
エンジン 故障しており、	ブレーキ 効かなくて
エンジン 壊れて	エンジン 不調となり、
ブレーキ 利かない」	ブレーキ 利かず
ブレーキ 壊れて	ブレーキ きかない
メーター 動かない	エンジン かからん。
ブレーキ 利かなくなった	ブレーキ きかなくて
エンジン かからなくて	ブレーキ 利かなくなった」
ブレーキ 壊れた。	エンジン 壊れた
ブレーキ きかない。	エンジン ダメージを
ブレーキ 効かない。	エンジン 壊れたり、
エンジン かからないんですよ。	ハンドル 壊れた）の
エンジン かからず	エンジン 故障して
ブレーキ 効かないから	ブレーキ 壊れてて
ブレーキ きかないから	エンジン 故障しても
ブレーキ 外れた	エンジン　かからないまま
エンジン　かからない…	エンジン　かからなくなった
エンジン　かかりにくい	エンジン　かからん
エンジン　かからなくなった。	エンジン　かからず。
エンジン　かからないと	エンジン　かからないので、
エンジン　かからなくなって	エンジン　壊れた。
ブレーキ　かからない	ブレーキ　効かず
ブレーキ　効かないし	エンジン　効かなくて
ブレーキ　止まらない	ブレーキ　効かなくなって
ブレーキ　壊れてる。	タイヤ　壊れて
メーター　おかしい。	エンジン　かからないから
エンジン　かからなかった	

本手法では，まず，以下の条件 1 と 2 を共に満たすツイートを抽出した．

条件 1 ：自動車の部品名を二種類以上含む．または，自動車の不具合情報に含まれる可能性がある形態素，即ち正の手がかり表現 (表 3) を登録し，それを一つ以上含む．

条件 2 ：負の手がかり表現，即ちストップワードとして登録した形態素 (表 4) や文字列 (表 5) を一つも含まない．

「形態素を含む」とは，ツイートを語単位で分け，その中で一致するものがあることを指す.「文字列を含む」とは，ツイートを文字列と見て，その中で一致するものがあることを指す.

この条件で抽出されたツイートに対して，以下の二つの手法を適用した．その内容と例を示す．

手法 1 ：不具合を表す文節の組を含み，かつ，それが係り受け関係にあるツイートを抽出する．つまり，係り受け情報を考慮する．例 5 は「エンジン」が「止まる」にかかっており，手法 1 で抽出できる．

例 5　出勤途中なのに、車が故障して動けないなう。

走行中にエンジンが止まるとか勘弁して欲しいなう。

手法２：不具合を表す文節の組を構成する語が共に含まれるツイートを抽出する．つまり，係り受け情報を考慮しない．例６は,「エンジン」が「かからなかった」ではなく「かけよう」にかかっているため，手法１では抽出できないが，手法２ならできる．このように，手法２で抽出するツイートの集合は，手法１のそれを含んでいる．

例６　（省略）車にエンジンかけようとしたら、かからなかった…（以下省略）

表 3: 正の手がかり表現

車	車輪	乗っ	乗れ	焦っ	直っ	直ら
死	死ぬ	不具合				

表 4: ストップワードとして登録した形態素

チャリ	チャリンコ	自転車	バイク	帰還
だけ	バッテリー	原付	原チャ	

表 5: ストップワードとして登録した文字列

はやぶさ	夢	みたいな	のような	『 「

3.3　実験結果

2010 年 12 月 8 日から同 13 日にかけて発信された，不具合を表す文節の組のいずれかを構成する文節を両方含むツイート，即ち例 6 のようなツイート 697 件に対して，抽出を行った．結果を表 6 に示す．

4　考察

手法１では，不具合を表す係り受け関係になる文節の組のいずれかを含むツイートのみが抽出される．一方，手法２では文節の組を構成する語を両方含んでいれば，係り受けの情報は関係ない．例えば，手法２は次のツイートを抽出する．

（省略）とかブレーキ気づかなくて前の車にぶつかりそうだったとか一瞬意識とんでたことがこんなにあるとはうひぃ。（省略）体がいうことききかない…睡眠配分考えどころだな。

このように，文節の組を構成する「ブレーキ」という語と「きかない」という語が別の文に存在するツイートは手法２でのみ抽出される．しかしながら，その内容は不具合を示さない場合がある．手法１は係り受け関係を考慮するため，手法１より精度が高く，良い結果となった．ただし，実験の抽出対象とするツイートの発信期間が短く，充分な数がないため，偶然精度が高くなったとも考えられる．

表 6: 実験結果

	精度（単位：%）	抽出した正例の数	抽出した数
手法１	72.5	58	80
手法２	71.4	65	91

また，不具合なのかどうかの判定が難しいツイートも存在する．例えば,「エンジンがかからない」という本文のツイートは手法１と２で共に抽出されなかった．原因として，文節の組を構成する「エンジン」「かからない」以外に意味のある語が無く，判定ができなかった．このツイートの発信者は，その直後に「いっきにやる気が出てきた、ありがとう」等のツイートを発信しており，自動車に関する発言はしていないため，不具合情報ではなく「やる気がない」というニュアンスのツイートだったことが推測できる．このように，ツイート単体ではなく，その前後のいくつかのツイートを判断材料とした場合，精度の向上に繋がると考えられる．

5　おわりに

不具合情報によく含まれる文節の組み合わせを利用し，自動車の不具合情報の抽出を行った．特に，今回は不要なツイートを除去し，精度を上げることに注力した．

今後の課題として，長期間に渡ってツイートを取得する必要があると考えられる．その理由は，3.3 節での実験は，抽出対象とするツイートを短い期間に発信されたものとし，数が少なくなったため，手法１が２に比べて精度が高くなったことが偶然である可能性があるためである．また，2.3 節について，URL を含んでいるが有益であるツイートが存在し，不具合情報の抽出において除去され，再現率を低下させている可能性がある．そこで，t ボタンを設置してあるサイトのリストを作り，その URL を含むツイートのみを除去することを検討する．また，ツイート単体で不具合情報を含むか否かを判定することは難しい場合があるため，ツイートの発言者が発信した前後のいくつかのツイートも判定材料に入れることを検討する．

参考文献

[1] 酒井浩之，梅村祥之，増山繁: 交通事故事例に含まれる事故原因表現の新聞記事からの抽出，自然言語処理，Vol. 13，No. 4，pp.99-124(2006).

[2] 工藤拓，松本裕治: チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌，Vol. 43，No. 6，pp.1834-1842(2002).