社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

ルールと機械学習を用いたTwitterからの不具合情報の抽出

栗原　光平†　　嶋田　和孝†

† 九州工業大学　〒 820–8502 福岡県飯塚市川津 680–4
E-mail: †{k_kurihara,shimada}@pluto.ai.kyutech.ac.jp

あらまし　製品の不具合に関する情報を収集することは，事故や社会的損失を未然に防ぐためにも重要である．本研究では，マイクロブログサービス Twitter を情報源とし，不具合情報を自動的に抽出する手法を提案する．まず，情報抽出の分野において一般的である，機械学習による手法を用いた実験を行い，機械学習による手法の問題点や，Twitter 上の不具合情報抽出における難しさを考察する．さらに，機械学習とは異なるアプローチとして，人手で作成したルールベースの抽出手法を提案する．

キーワード　情報抽出，不具合情報，Twitter

Trouble information extraction based on rules and machine learning from Twitter

Kohei KURIHARA† and Kazutaka SHIMADA†

† Kyushu Institute of Technology　680–4, Kawazu, Fukuoka,Japan
E-mail: †{k_kurihara,shimada}@pluto.ai.kyutech.ac.jp

Abstract　In this paper, we propose a method for extracting trouble information from Twitter. One useful approach is based on machine learning techniques such as SVMs. However, trouble information is a fraction of a percent of all tweets on Twitter. In general, imbalanced training data generates a weak classifier on machine learning techniques. Therefore, we utilize another approach; a rule based method. We consider the advantages and disadvantages of the two approaches through experiments.

Key words　Infomation extraction, Trouble infomation, Twitter

1. はじめに

自動車のリコールなどに代表されるように，製品の不具合は大きな社会的損失に結びつき，時には重大な事故等につながることもある．メーカーや企業は不具合の発生を防ぐため，過去の不具合事例等の情報を製品製造に取り入れ，信頼性の向上に活用している [1]．安全な製品の開発を支援するためにも，不具合に関する情報を多く収集することは重要である．

製品の不具合情報収集に関連する研究として，新聞を対象に交通事故の記事から事故の原因となる表現や関連情報を抽出する研究 [2], [3] や，不具合事例文から製品・部品を示す語を抽出する研究 [4] などが行われている．しかしながら，新聞などの一般メディアを対象とした場合，不具合の発生から記事として公開されるまでに時差がある，一般メディアには出現しない不具合事例が多く存在する可能性がある，などの問題がある．また，その他の情報源として，公的組織が独自に不具合情報を収集し，不具合事例集として情報を保持している場合がある．それらを用いれば不具合について詳細な情報を得ることができるものの，公的組織の詳細な調査に基づき作成・公開されているものであることから，データ数に限りがあり，追加収集も困難であるという問題がある．

そこで，本論文ではそれらの欠点を補うために，個人が自由に情報を発信することができる CGM (Consumer Generated Media) に着目し，その中でも情報源に Twitter を用いた情報抽出手法を試みる．Twitter は，今していることや感じたことを 140 文字以内で投稿する「マイクロブログ」と呼ばれ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．現在，国内のユーザ数は 3000 万人以上，一ヶ月の日本人の総ツイート数は 2012 年 6 月の時点で 1 億件を超えており，一般のメディアには登場しない個人の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．

本研究では，Twitter から製品の不具合情報を抽出することを目的とし，機械学習による手法を Twitter に適用した場合の問題点について考察するとともに，機械学習とは異なるア

1	2013年06月30日 男性　福井 HP	ホンダ　CRF250L 2013年06月　1,500 km JBK-MD38　MD38E	動力伝達　2013年06月29日発見 エキゾーストマニフォールドのエンジンから出た曲線部分と、チェーンの内側全体に錆が発生している。またエキゾーストマニフォールドには穴をふさいだような溶接跡がある	

図 1　国土交通省の不具合事例.

プローチとして，ルールベースによる抽出手法を提案する．まず，情報抽出の分野で一般的な手法である機械学習を用いて，Twitter からの不具合情報の抽出を行う．得られた結果についてエラー分析を行い，Twitter というサービス自体の特性や，Twitter 上の不具合情報の持つ特徴から，機械学習によるアプローチの問題点について考察する．また，それらの分析を基に，機械学習とは異なるアプローチとしてルールベースの抽出手法を提案する．機械学習による手法についての考察で得た，Twitter 上の不具合情報の特性をもとに，抽出ルールを人手で作成し，実際に抽出実験を行うことで精度を検証をする．

2.　本研究で抽出対象とする不具合情報

Twitter 上に出現する不具合情報は，一般メディアのものとは大きく異なる特徴を持つ．ここでは，一般メディアの不具合情報と比較しながら，Twitter 上の不具合情報の持つ特徴について述べ，本研究で扱う不具合情報について定義する．

2.1　一般メディアの不具合情報との比較

新聞記事や不具合事例をまとめたサイトの不具合情報は，多くの場合，不具合対象とその症状について詳しく明記しており，標準的な日本語で記述されている（図 1 参照）．それに対して，Twitter 上で見られる不具合情報には，具体的な症状の記述の省略や，Web 特有の表現の使用など，Twitter ならではの特徴が強く反映されている．ここで，Twitter 上の不具合情報の特長について，具体的な事例を示しながら説明する．

2.1.1　Web 特有の表現

Twitter 上のテキストにはネットスラングや顔文字といった Web 特有の表現が頻繁に出現する．Web 特有の表現が用いられている例を次に示す.

- あと遂に車があかんわ www ブレーキ踏んだら轟音が wwwwww 怖すぎ笑えん (´; ω ; ‘)

Twitter 上の不具合情報では，特に悲しんでいる顔文字や驚いた顔文字，落ち込んでいる感情を表す記号などが特によく用いられる．また，文全体としてネガティブな極性を持ちやすい不具合情報だが，笑いを意味する「w」という記号が用いられることもある．

2.1.2　比 喩 表 現

Twitter 上の不具合情報では，不具合の症状の記述に比喩表現が用いられることがある．症状の記述に比喩表現が用いられている例を次に示す.

- こんな時に車のバッテリーが死ぬなんて

この例では，本来「バッテリーがあがる」と書くべきところに，「死ぬ」という比喩表現が用いられている．同様な意味で「逝った」や「終わった」などの比喩表現が用いられることもある．

2.1.3　具体的な症状の記述の省略

Twitter 上の不具合情報では，具体的に症状について述べている部分そのものが省略されている場合も存在する．例を次に示す.

- あぁ〜あ、俺の車が (。¿Д¡)

この例では，具体的に車がどうなったのかについては書かれていないものの，困ったような顔文字が出現していることから不具合や事故などが起きたのではないかと推測できる．また，顔文字と同様の使い方で「…」が使われていたり，症状の記述を省略して，不具合を示す写真の URL が添付されていたりする場合などもある．省略が用いられている場合は，顔文字やその他の記号などから文の極性を測ったり，前後のツイートを見るなどして，不具合らしさを推測する必要がある．

2.2　対象とする不具合情報の定義

Twitter 上の不具合情報は一般メディアのものに比べて非常に多様であり，曖昧な表現をされる場合もある．よって，ここでは本論文で扱う不具合情報を次のように定義する．

[条件 1]　不具合対象が同ツイート内に存在している

Twitter では一度のつぶやきが 140 文字以内という制限があり，短い文を気軽に投稿できるという特徴がある．そのため，時に一連の話題が複数ツイートに分けて投稿される場合がある．例えば不具合対象についての記述と，症状についての記述が複数ツイートに分けて投稿される場合がある．今回はそのようなケースは対象外とし，少なくとも同ツイート内に不具合対象が出現しているものを対象とする．

[条件 2]　症状の記述の省略を認める

具体的な症状の記述が省略されている不具合情報は，情報の信頼度や確かさといった点では疑問があるものの，Twitter ならではの興味深い表現であるといえる．少なくとも顔文字や「…」など不具合を連想させるようなその他の付加要素がある場合に限り，具体的な不具合の症状の記述が省略されている場合でも，不具合情報として扱う．

3.　機械学習による不具合情報抽出

ここでは，機械学習を用いた不具合情報の抽出方法について述べる．機械学習により作成した分類器に対し，実際の Twitter 上のリアルなデータを適用し，その結果について考察を行う．

3.1　手　　法

機械学習器には SVM [5] を，実装のためのツールには SVM^{light} [6] を使用する．我々は，先行研究において顔文字やネットスラングといった Twitter 特有の特徴を考慮した複数の素性を組み合わせた分類器を作成したが，BoW(Bag of Words) だけを用いた場合に対する有意性は見られなかった [7]．よって，今回の実験では素性として BoW のみを用いるものとする．

3.2　先行研究の実験設定とその結果

分類器の訓練データには，Twitter から人手で収集した不具

合情報（以降，正例）と，Twitter から機械でランダムに収集した不具合情報ではないツイート（以降，負例）をそれぞれ 450 件ずつ，計 900 件を用いた．また，検定方法には，leave-one-out 法による交差検定を用いた．先行研究で作成した分類器の，訓練データに対する再現率と適合率，F 値を表 1 に示す．素性として BoW のみを用いたシンプルな分類器でも，F 値が約 90%と非常に高い精度になっているのがわかる．

表 1　先行研究で作成した分類器の精度.

再現率	適合率	F 値
0.88	0.98	0.93

3.3 追　実　験

先行研究において作成した分類器は，訓練データに対して非常に高い精度で不具合情報を分類することに成功した．しかし，訓練データに強く依存した分類モデルになってしまっている可能性があるため，Twitter 上のリアルなテストデータを適用することで，再度精度を検証する．

3.3.1 実 験 設 定

テストデータとして，Twitter から自動で収集したツイート 30000 件を用いる．テストデータはアノテーションしていない完全未知なデータであるため，分類器が抽出した結果について人手で不具合かどうかのアノテーションを行い，適合率のみで評価を行う．

3.3.2 実 験 結 果

分類器に，未知のテストデータを適用した場合の結果を表 2 に示す．訓練データに対する精度は非常に高い値であったのに対し，未知のテストデータを適用した場合には大きく適合率が低下している．

表 2　分類器にテストデータを適用した場合の精度.

抽出数	正解数	適合率
3742	720	0.19

3.4 機械学習による手法の問題点

テストデータによる検証で，適合率が非常に低い値となってしまった大きな原因として，訓練データの単語のカバレッジが不足していることが考えられる．テストデータ数が 30000 件であるのに対し，訓練データ数が合計 900 件と非常に少なく，訓練データから得られる単語のカバレッジが低くなってしまい，テストデータを適切に分類することができなかったと考えられる．

また，Twitter 上における実際のデータのバランスを考慮すると，訓練データにおける正例と負例の比率が等しくなっているのも適切ではない．Twitter 上において不具合情報は，全体のごくわずかな割合でしか存在しておらず，Twitter 上のほとんどの情報は不具合とは無関係な情報である．よって，訓練データにおいても，実際の Twitter 上の環境と同じように，負例のデータ数が正例よりも多いほうが望ましい．または，機械学習において正例に対して学習コストを付加することも有効であると考えられる．しかし，訓練データの負例の割合を増加させたり，正例に学習コストを付加した場合には，再現率が低下してしまう問題がある．

機械学習による手法において，総合的に精度を向上させるには，正例・負例ともにデータ数を増加させることが最も重要であるが，人手による正例の収集には大きなコストがかかり現実的ではない．また，膨大なノイズ（無関係なデータ）の中から，ごく一部の不具合情報を取り出すという問題設定に対し，機械学習によって二値分類問題として解こうとするアプローチは適切ではない可能性がある．訓練データの集めづらさや，データのバランスを考慮すると，今回のタスクにはパターンマッチング等を用いた抽出問題としてのアプローチが適していると考えられる．

4. ルールベースの手法による不具合情報抽出

ここでは，ルールベースの手法による不具合情報の抽出方法について述べる．ルールベースの手法では，人手で不具合情報の抽出ルールや抽出パターンを作成することで，不具合情報を抽出する．まず，不具合の対象となるドメインごとに不具合の症状や製品の異常を表す表現の収集を行い，不具合表現辞書を作成する．次に，ドメインごとに不具合情報の特徴を分析し，不具合らしさをより強める働きを持った副詞や修飾子を不具合ブースト表現として収集する．さらに，ドメインの不具合情報ではないその他のツイートから，正常な動作を表す表現をストップワードとして収集する．最後に，不具合表現と不具合ブースト表現，ストップワードをそれぞれ組み合わせ，抽出ルールを作成する．

4.1 不具合表現の収集

不具合情報において，不具合の症状や製品の異常を表す表現を不具合表現と定義する．不具合情報の抽出において，不具合表現を認識することは重要であるが，不具合を表す表現は非常に多様であり，それらを網羅的に収集するのは困難である．また，通常は不具合らしさを持たないが，特定のドメインと共に出現した場合にのみ，不具合らしさを持つような表現も存在する．例えば，図 2 のような例では，A の文章は不具合情報であり,「落ちる」が不具合表現となるが，B の文章は不具合情報ではなく,「落ちる」も不具合表現ではない．

A.「パソコンがいきなり落ちたんだけど !!???」
B.「マウスが床に落ちて電池飛んでった」

図 2　携帯電話の不具合対象要素の例

このように，特定のドメインに強く依存した不具合表現も存在するため，単純に不具合を示す可能性のある表現のみを収集しても不十分であり，ドメインとの関係についての知識も必要となる場合がある．

そこで，本手法では不具合の対象となるドメインごとに不具合表現を収集，辞書を作成する．不具合表現は非常に多様であるものの，ドメインによっては出現する不具合表現が偏っていたり，限定的である場合がある．よって，1 つのドメインごとに不具合表現の分析・収集を行うことで，収集がしやすくなる

と考えられる．また，ドメインと不具合表現を関連づけて扱うことにより，ドメインに対して無効な不具合表現によって誤った情報を抽出してしまうことを防止できる．具体的には，次の手順で不具合表現の収集を行う．

（1） 不具合の検索対象となる製品の決定.

（2） 不具合対象要素の収集.

（3） 不具合表現の獲得.

まず,「携帯電話」や「自動車」など，不具合の対象となる製品を決定する．次に，その製品の中で不具合の対象となり得る要素（以降，不具合対象要素）を収集する．1 つの製品でも，通常は様々な部品や部位から構成されており，それぞれで発生する不具合にも違いがある．よって，不具合表現もその製品単体ではなく，製品の持つ様々な部分的要素ごとに関連付けて考える必要がある．不具合対象要素の具体的な収集方法として，製品に関連のあるツイートを収集し，正規表現を用いて製品名の直後にくる要素を抜き出し，その中から不具合と関連のあると思われるものを人手で選別する．例えば，スマートフォンにおける不具合対象要素の例としては次の図 3 のようなものがある．

電源，充電，画面，バッテリー，ボタン，調子

図 3　スマートフォンの不具合対象要素の例.

不具合対象要素を収集したら，次にそれぞれの要素を含むツイートを収集・分析し，人手で不具合表現を抽出する．製品とその部分要素を含むツイートという条件を設けることで，出現する不具合表現も比較的限定的になり，収集が行い易くなる．スマートフォンの各不具合対象要素ごとの不具合表現の例を表 3 に示す.

表 3　携帯電話の不具合対象要素と不具合表現の例

不具合対象要素	不具合表現
充電	”出来ない ”，”すぐになくなる ”など
電源	”落ちる ”，”消える ”など
画面	”映らない ”，”真っ暗になる ”など
ボタン	”反応しない ”，”押せない ”など

4.2　不具合ブースト表現とストップワードの収集

不具合情報には，ドメインとそれに関連のある不具合表現の知識だけでは解けない事例が存在する．例えば，次の図 4 のような場合である．

C.「スマホの電源が勝手に落ちるのなんでー？」
D.「スマホ充電切れて電源落ちたー (--)」

図 4　不具合表現だけでは解けない例.

この例ではどちらの文章にも「電源」という不具合対象要素と,「落ちる」という不具合表現が出現している．しかし，C の文章は不具合情報であるが，D の文章はあくまで正常な事象であり，不具合情報ではない．このように，不具合対象要素と不具合表現の有無に加え，文章の意味的関係を考慮しなければ解けない事例も存在する．しかし，Twitter の文章は口語的な文体や Web 特有の表現の影響で非常に解析しづらく，意味的関係の認識を行うのはコストが大きい．そこで，ここでは不具合らしさを強調，または減少させている要素に着目し，それらの有無を判別することでこの問題を解決する．

村上らの研究 [8] では,「突然」や「頻繁に」といった，不具合や想定外の事態を示唆するような働きを持った連用修飾表現を収集し，その修飾関係から不具合を表す動詞を収集するというアプローチをとっている．そこで，本研究では「突然」などのような，不具合情報に多く出現する修飾表現やフレーズを，不具合ブースト表現として収集し，不具合表現と合わせて不具合情報の判別に利用する．例えば先程の図 4 の C の文章では,「勝手に」という副詞が不具合表現「落ちる」を修飾していることで，文章全体の不具合らしさが強調されている．よって，不具合表現「落ちる」にとって「勝手に」という副詞は不具合ブースト表現であるといえる．

さらに，不具合ブースト表現とは逆に，不具合らしさを減少させたり，無くしたりさせる表現も存在する．それらの表現はストップワードとして収集し，不具合ではない情報の誤抽出を防止するために用いる．例えば，先程の例の D の文章では,「充電切れて」という部分が，不具合らしさを無くさせる働きをしている．よって，不具合表現「落ちる」にとって,「充電切れて」はストップワードとなる．

不具合対象要素と不具合表現の関係に加え，不具合ブースト要素とストップワードの有無も考慮することで，単純に不具合表現の有無だけでは解決できないような事例でも適切に対応することが可能になると考えられる．不具合ブースト表現とストップワードは，不具合表現を収集する際に人手で獲得し，不具合要素と不具合表現と関連付けて保持しておく．

4.3　抽出ルールの作成

不具合表現，不具合ブースト表現，ストップワードをそれぞれ収集したら，入力となるツイートに次の手順で処理を行うことで不具合情報を抽出する．

（1） 不具合対象要素の決定.

（2） 不具合表現のマッチング.

（3） 不具合ブーストワード・ストップワードのマッチング.

まず，不具合の対象となる要素が何なのかを決定する．これは，不具合対象となる製品を決定した際に収集した不具合対象要素のリストをもとに，パターンマッチングによって決定する．不具合対象要素が何であるかがわかったら，次に，その不具合対象要素に関連のある不具合表現が出現しているかどうかを判定する．さらに，不具合対象要素と不具合表現に，何らかの不具合ブースト表現またはストップワードが関連づいている場合には，それらの有無についてもマッチングを行い，ブースト表現が出現している場合には不具合情報として抽出し，ストップワードが出現している場合は不具合ではないとして抽出しない．不具合ブースト表現やストップワードが関連づいていない場合には，不具合表現が出現した時点で不具合情報として抽出する．

例として，4.2 節の図 4 にある文章 C を不具合情報として抽出する際のイメージ図を図 5 に示す．まずスマートフォンにおける不具合対象要素として「電源」がマッチし，さらに「電

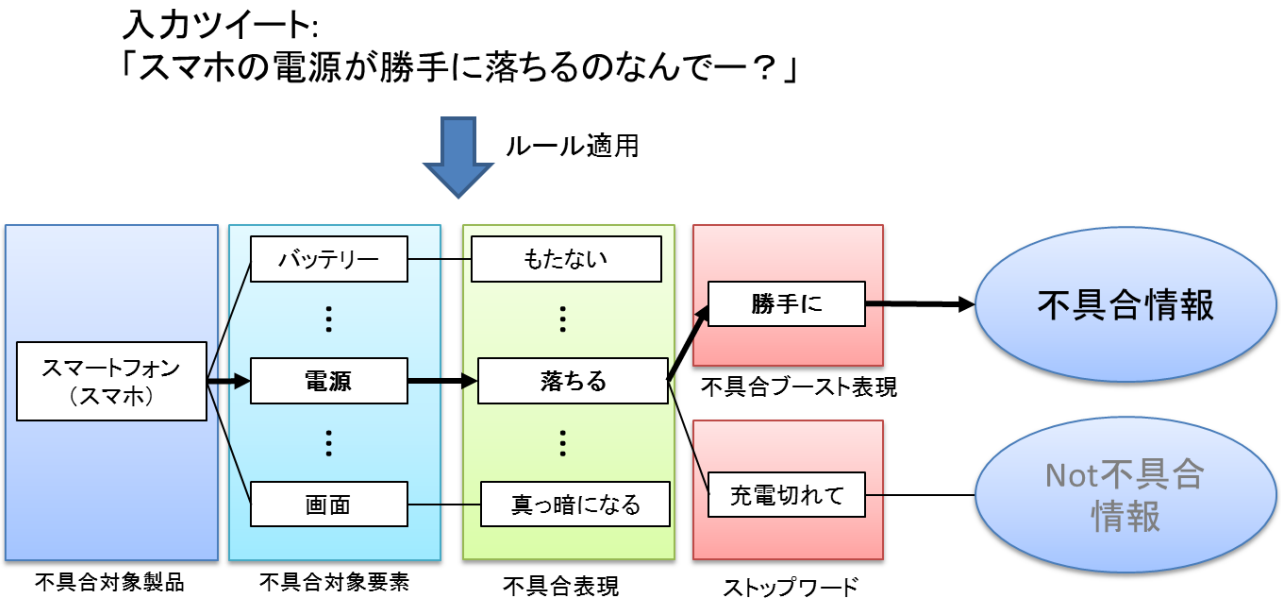

図 5　ルールによる抽出のイメージ図.

源」に関連のある不具合表現の中から「落ちる」がマッチする．さらに,「電源」と「落ちる」に依存する不具合ブースト表現として「勝手に」がマッチし，最終的に不具合情報として判断される．

4.4　実　　験

4.4.1　実 験 設 定

作成した抽出ルールで，実際に Twitter 上で収集したデータから不具合情報の抽出を行い，有効性を検証する．機械学習による手法と比較するため，不具合の対象となるドメインやテストデータなどは同じものを用いる．不具合の対象は,「車」「パソコン」「携帯電話（スマートフォン）」の 3 種類とし，テストデータには Twitter から収集した未知のツイート 30000 件を用いる．

4.4.2　実験結果と考察

ルールによる手法で，テストデータから不具合情報を抽出した結果を表 4 に示す．抽出数こそ機械学習による手法に比べ少ないものの，非常に高い適合率で不具合情報を抽出できることが確認できた．ただし，30000 件のテストデータに対し，抽出数が 474 件と少なく，再現率においてはまだ改善の余地があるといえる．今後は不具合表現の収集とルールの拡張を続け，再現率の向上を目指す．

表 4　ルールによる手法の実験結果.

	抽出数	正解数	適合率
提案手法	474	444	0.94

5.　考　　察

機械学習による手法では，テストデータを適用すると大幅に適合率が低下したのに対し，ルールによる手法ではテストデータを適用しても高い適合率で抽出することができた．また，今回のルールによる手法では，特定のデータに依存してルールを作成するのではなく，特定のドメイン単位でルールを作成するため，適用するテストデータを変更しても精度が大幅に低下しづらいと思われる．さらに，それぞれの手法の抽出結果を比較してみると，ドメインに強く依存した不具合表現が用いられている不具合情報は，機械学習では抽出に失敗している場合が多かったが，ルールによる抽出ではそういった事例も正しく抽出ができていた．

ルールによる手法の課題としては，抽出数が少ない点があげられる．未アノテーションのテストデータのため正確な再現率は不明だが，機械学習による手法よりも抽出数が少なく，30000 件というテストデータの数から考えても，まだ抽出に失敗している不具合情報が多数存在していると考えられる．しかし，今後もエラー分析を重ね，不具合表現や不具合ブースト表現などのさらなる獲得を続けることで，確実に再現率は向上していくと思われる．また，ルールによる手法の高い適合率を利用し，ルールによる手法を大量のデータに適用し，獲得した不具合情報を機械学習の訓練データとして利用することで，同時に機械学習による手法の制度改善にもつながっていくと考えられる．現在はルールによる手法と機械学習による手法を独立させて用いているが，今後は 2 つの手法を組み合わせた手法に関しても検討していきたい．

さらに，応用タスクとして Web 特有の表現を不具合表現として活用することが挙げられる．Twitter 上の不具合情報には，不具合表現さえ省略され，顔文字や記号，長音化などの現象により置き換えられているような事例が存在する．それらの情報は，前後のツイートを見なければ，それ単体で不具合表現かどうかを判断することができないが，顔文字などの種類によっては不具合を暗に示しているような文章も存在する．そういった情報を「不具合候補」という第 2 のクラスとして抽出することができれば，より Twitter の特性が現れた不具合情報の獲得が期待できる．

6. まとめと今後の展望

本論文では，抽出ルールを用いて，Twitter から自動的に不具合情報を抽出する手法を論じた．情報抽出において一般的な，機械学習による手法を Twitter 上のデータに適用した場合の問題点について考察し，ドメインごとの不具合表現に着目したルールベースの抽出手法を提案した．結果として，高い適合率で不具合情報を抽出することができたが，再現率については改善の必要があることがわかった．

今後はさらに不具合表現の収集と抽出ルールの拡張を行い，再現率の向上を目指すとともに，機械学習とルールによる手法を組み合わせた手法についても検討していく．さらに応用タスクとして，顔文字等の Web 特有の要素に着目し，不具合表現が省略されているような不具合情報の抽出も試みる．

文　　献

[1] 粟納裕貴, 馬強, 吉川正俊,”失敗知識データベースを用いた失敗事象の原因分析”, DEIM2012，E2-5，(2012).

[2] 酒井浩之，梅村洋之，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.12，No.2 pp.99-123 (2006).

[3] 野畑周, 佐田いち子, 井佐原均, “ 新聞記事中の事故・事件名の自動抽出 ”，情報処理学会，研究報告 2005-NL-167，pp.125-130 (2005).

[4] 大森信行, 森辰則, “ 不具合事例文からの製品・部品を示す語の抽出－語の実体性による分類－ ”，電子情報通信学会論文誌 D, Vol.J95-D No.3, pp.697-706 (2012).

[5] Vladimir N. Vapnik：”The Nature of Statistical Learning Theory”，Springer,New York (1995).

[6] Joachims T：”Making large-Scale SVM Learning Practical. In Advances in Ker-nel Methods-Support Vector Learning”, chapter 11, MIT Press (1999).

[7] 栗原光平, 嶋田和孝, “ Twitter からの不具合情報の抽出抽出 ”, 電気学会情報システム研究会 IS-13, pp.63-68 (2013).

[8] 村上拓真, 那須川哲哉, “ 特長的な記述を利用した問題発見手法の実現 ”，電子情報通信学会技術研究報告，言語理解とコミュニケーション 111(119), pp.31-35, (2011).