Twitterを対象とした不具合情報の抽出

栗原　光平*　　嶋田　和孝（九州工業大学）

Trouble information extraction from Twitter

Kohei Kurihara*, Kazutaka Shimada, (Kyushu Institute of Technology)

Abstract

In this paper, we propose a method of trouble information extraction from the Web. We focus on Twitter as the information resource. We apply some dictionaries that we construct, such as wikipedia and net slang, to the analysis method. The features for the extraction are bag-of-words, modality, emoticon and so on. We classify each tweet into trouble information or not by using SVMs. We obtained approximately 80% on the F-measure.

キーワード：情報抽出，不具合情報，Twitter
(Information extraction, Trouble information, 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[1] 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「ミニブログ」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，Twitter の特性を考慮した抽出手法の検討を行う．Twitter 上の文書には，Web 特有の表現や流行の語などが多く出現するため，従来までの言語処理技術をそのまま適用しても不十分である(5)．そこで，Twitter 上のテキストを解析する事前準備として，ネットスラング辞書や Wikipedia タイトル辞書などの合計 4 つの解析器用辞書を新たに作成し，形態素解析器を Twitter 上のテキスト解析が行い易くなるよう拡張する．次に，Twitter から不具合情報を抽出するため，機械学習による手法を実装し，精度を求める．得られた結果をもとに Twitter における不具合情報抽出の難しさについて考察する．

2. 本研究で対象とする不具合情報

Twitter の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時例文と比較しながら，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

[1] https://twitter.com/

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時例文

Fig. 1.　Trouble information from a public agency.

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．

a.　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．

b.　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．

c.　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.　解析器用辞書の拡張

前節で述べたように, Twitter 上のテキストには, Web 特有の表現や顔文字，記号などが出現し，一般メディアに比べ言語表現が非常に多様である．そのため，これまで研究開発されてきた言語処理技術の多くを，そのまま Twitter 上のつぶやき（以降，ツイート）に対し適用しても十分な精度は得られない．そこで，ツイートを解析しやすくするために解析器用辞書を拡張し, Twitter 特有の表現等を適切に解析できるようにする．

具体的に解析を困難にしている要因には様々なものがあるが，語彙知識の充実で解決できるものと，そうではないものに大きく分けることができる．例えば顔文字やネットスラング，流行語などは語彙知識として保持していれば認識できるが，単語の音声置換('before' →'b4')や長音化現象('cool' →'cooooool') [6]などは語彙知識の充実だけでは解決できない．

ここでは，語彙知識で対応できる問題にのみ焦点を当て，既存の形態素解析器用の辞書を新たに作成することで，Twitter 特有の表現や固有名詞を認識できるようにする. 具体的には次の 4 つの辞書を作成する．

（1） ネットスラング辞書
ネット上でよく用いられる表現や俗語，慣用表現や記号表現などを登録する．情報源として，ネットスラングや IT 用語などをまとめたサイト「ネット用語辞典ネット王子」[2]を用いた．2ch，Twitter，ニコニコ動画，ゲーム用語，アニメ用語，若者言葉など様々なジャンルの語を手動で収集し，合計 2,155 語を登録した．

（2） 顔文字辞書
顔文字辞書には，一般的に使われている顔文字を合計 5,645 種類登録した．情報源には顔文字辞書サイト[3]の他，実際に Twitter のテキストから抽出した顔文字や日本語入力ソフトに登録されている顔文字などを登録した．しかし，顔文字に関して全て網羅することは不可能であり，機種依存文字や特殊文字を用いたもの，ユーザが自作した顔文字には対応できない．今回は特殊なものは対象とせず，あくまで一般的に使用されている顔文字を認識することを目的とする．

（3） Wikipedia タイトル辞書
主に名詞への語彙知識を充実させるために，日本語版 Wikipedia[4]の全タイトル 1,124,106 語を登録した．これにより地名や人名，作品名やサービス名など様々な名詞を認識することができる．

（4） はてなキーワード辞書
はてなキーワード[5]とは，「株式会社はてな」が運営する共有辞書サービスであり，ユーザが自由に編集可能であることが特長である．Wikipedia のように辞書的な定義のほか，ネットで話題になった語や比較的長い説明文など，Wikipedia に比べて多様なキーワードが登録されており，Wikipedia 辞書よりも時勢や流行に関連深い語彙知識が得られると期待している．今回ははてなキーワードに登録されている語から 342,097 語を登録した．

以上 4 つの辞書を適用し,ツイートの解析を行うことで，従来の解析器では単なる記号列や未知語として認識していた文字列を，意味のある文字列として認識することが可能となる．

4. 機械学習による不具合情報抽出

ここでは，機械学習による不具合情報抽出の方法について述べる．機械学習には，Bag of words などの一般的な素性に加え，顔文字やネットスラングといった Twitter 特有の特徴も用いる．学習器にはサポートベクターマシン(Support Vector Machine：SVM) を用い，2 値分類により，そのツイートが不具合情報かどうかを判別する．

〈4･1〉 手法

情報抽出において基本となる，機械学習による手法を実装する．機械学習器には SVM[(7)] を，機械学習を行うツールには SVM[light(8)] を使用する．SVM は，1995 年に，AT&T の V．Vapnik によって統計的学習理論の枠組みで提案された 2 クラスのパターン認識手法のことである．SVM では，2 種類のクラスのデータと，分離超平面との間の距離(マージンと呼ぶ) が最大になるような分離超平面が，最も汎化能力の高い超平面になるということを利用している．クラスの特徴ベクトルを非線形変換して，その空間で線形の識別を行う「カーネルトリック」と呼ばれている方法を，「マージン最大化」という基準で行うため，SVM は高い汎化性能で識別を行うことができる．

素性には，機械学習では一般的である Bag of words と文長のほか，モダリティや Web 特有の表現に関するものを用いる．

〈4･2〉 素性

ここでは機械学習に用いる素性を説明する．モダリティに着目したものや，Twitter 特有の特長を利用したものなど計個 7 個の素性を用いる．

（1）Bag of words
機械学習には一般的な Bag of words を素性の一つとして用いる．Twitter 上の文章は 1 文が短いため，単語が文全体に与える影響は強くなると考えられる．不具合情報に頻出し，強い不具合らしさを持っているような単語や表現の出現の有無を測る．

（2）文長
不具合の多くは予期せぬタイミングで発生し，意外性を伴うものであると考えられるので，不具合が発生したことをツイートする場合，その文章は動揺や驚きから比較的短く簡潔なものになると予想される．Twitter 上における不具合情報と文章の長さにおける関係性を測るため，文長を素性の一つとして加える．

（3）感嘆詞
不具合の発生についてのツイートには，「えっ」や「はっ」などの感嘆詞が一緒に出現することが多い．これは，不具合の発生による動揺や驚きが感嘆詞と表れているものと考えられる．よって感嘆詞と不具合情報には強い共起性があるものと考え，素性の一つとして加える．

（4）認識のモダリティ
Twitter 上の不具合情報の例を分析した結果，「かもしれない」，「ようだ」，「かも」といった認識のモダリティが多く出現している傾向が見られた．これは，突如発生した不具合についての，発言者自身の判断（可能性，証拠性）が不安の気持ちと共に表れているものであると考えられる．不具合と認識のモダリティの関連性について測るため，素性に追加する．

[2] http://netyougo.com/
[3] http://matsucon.net/material/dic/
[4] http://ja.wikipedia.org/wiki/
[5] http://d.hatena.ne.jp/keyword/

表 1　実験結果

Table 1.　The experiment result.

素性	辞書の拡張有り			辞書の拡張なし		
	適合率 [%]	再現率 [%]	F 値 [%]	適合率 [%]	再現率 [%]	F 値 [%]
(1) BOW （ベースライン）	69.33	81.25	74.82	**75.33**	83.70	**79.29**
(2) BOW+文長	68.67	79.23	73.57	66.67	83.33	74.07
(3) BOW+感動詞	68.00	**82.26**	74.45	72.67	**83.85**	77.86
(4) BOW＋認識のモダリティ	68.67	79.84	73.83	72.67	83.21	77.58
(5) BOW+疑問形	69.33	81.89	**75.09**	72.67	82.58	77.30
(6) BOW＋顔文字	**72.00**	77.14	74.48	—	—	—
(7) BOW＋ネットスラング	69.33	81.89	**75.09**	—	—	—
(8) ALL	62.00	78.81	69.40	63.33	83.33	71.97

(5) 疑問形

Twitter 上の不具合情報の例を分析した結果, 文章の最後が疑問形で終わっているものが多い傾向にあった．これは，不具合の発生に関して「信じられない」といったような心理が表れているものと考えられる．しかし疑問形で終わるすべての文章が不具合情報となるわけではないのは明らかである．不具合情報と疑問形の関係をより明確にするために，疑問形の出現を素性に追加する．

(6) 顔文字

Twitter 上のテキストには, 発言者の感情などを表した顔文字が多く出現する．これは新聞などの一般メディアには登場しない Web 特有の要素であり, 情報源として Twitter を用いるにあたって無視できない重要な文の要素であると考えられる．不具合情報にはネガティブな意味を持つ顔文字が出現しやすいという仮定のもと，顔文字と不具合情報の関連性を測るためにも，顔文字を素性の一つとして追加する．

(7) ネットスラング

Twitter 上のテキストには落ち込んでいる姿を模した「orz」などの記号列や,「終わった」を言い換えた「オワタ」などの表現のように，様々なネットスラングが出現する．これも顔文字と同じように一般メディアには出現しない Twitter 特有の要素であり重要な特長であると考えられる．不具合情報とこれらネットスラングの関連性を測るため，素性の一つとして追加する．

5. 実験

実際に Twitter 上のつぶやきを収集し，機械学習による分類を行う．得られた結果を分析し，Twitter における不具合情報抽出の難しさや問題点を明確にすることで，より具体的な手法を検討する．

〈5･1〉　実験設定

実験データには, Twitter から人手で収集した不具合情報と（以降，正例），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 300 件ずつ，計 600 件を用いた．Twitter には，特定の相手に向けてメッセージを発信する「リプライ」機能や，他人のつぶやきを取り上げ自分のフォロワーに紹介することができる「リツイート(RT)」機能など，様々な機能が存在する．これら Twitter 特有の機能に関する文字列は，ツイートの解析に影響を及ぼす可能性があるため，今回の実験では予め除去している．また, bot と呼ばれる自動的にツイートを送信するプログラムによるツイートや，宣伝目的のツイートなども明らかに不要であるため，それらのツイートは実験データから除外している．

この 600 件のデータについて，以下に示す条件で leave-one-out 法による交差検定を実施した．素性に BOW のみ用いるものをベースラインとする．

(1) BOW （ベースライン）
(2) BOW + 文長
(3) BOW + 感嘆詞
(4) BOW + 認識のモダリティ
(5) BOW + 疑問形
(6) BOW + 顔文字
(7) BOW + ネットスラング
(8) ALL

形態素解析器には mecab を利用した．また，今回作成した解析器用辞書の効果を検証するために，解析器用辞書を拡張した場合としなかった場合の 2 パターンで実験を行う．なお，顔文字とネットスラングの素性に関しては解析器用辞書の拡張が前提であるため，解析器用辞書を拡張しない実験では除外する．

〈5･2〉　実験結果

実験結果として得られた再現率と適合率，F 値を表 1 に示す．太字は再現率と適合率，F 値それぞれの列において，(1) ～ (7) の中で最良の値である．

まず解析器用辞書を拡張した場合の結果を見ると，F 値については疑問形とネットスラングがともに最良，再現率では感動詞，適合率では顔文字が最良という結果になった．ベースラインの結果と比較してみると，適合率については若干上昇しているものの，再現率と F 値についてはほとんど差が見られない．

次に解析器用辞書の拡張をしなかった場合の結果を見ると，再現率では感動詞が最良，適合率と F 値ではベースラインが最良という結果になった．唯一再現率の向上が見られた感動詞の素性も，ベースラインとの差はごくわずかであり，BOW 以外の素性が効果を発揮していないことがわかる．

解析器用辞書を拡張した場合としなかった場合の結果をそれぞれ見比べてみると，解析器用辞書を拡張したほうが，全体的に数値が低下していることがわかる．対して，解析器用辞書を拡張しなかった場合は，数値は高いものの素性ごとの値のばらつきが小さく，BOW 以外の素性が効果を発揮していない．

6. 考察

得られた結果について考察を行う．解析器用辞書の拡張の有無による差異や，各素性の働きについて分析し，実際に分類に失敗した事例を示しながら Twitter 上の不具合情報を機械学習によって抽出する難しさについて議論する．

〈6·1〉 解析器用辞書拡張の影響

一般メディアとは大きく異なる Twitter 上のテキストに，既存の言語解析器を適用させるため，4 つの辞書を作成・追加した．また，辞書の追加により顔文字やネットスラングといった Web 特有の要素そのものを認識し，素性として用いることができるようになる．

しかし，結果を見てみると，辞書ありのほうが辞書なしの結果に比べ全体的に数値が低下している．よって，単純に語彙知識を拡張しただけでは精度向上には直接影響しないということがわかる．実際に失敗した事例について見てみると，辞書ありのほうが辞書なしよりも顔文字やネットスラングが出現している事例を優先的に不具合に分類する傾向があったものの，不具合事例ではないがネットスラングが出現している事例などを不具合だと誤分類している例も見られた．

〈6·2〉 各素性の効果

解析器用辞書を拡張した場合，適合率で最も良い性能だったのは顔文字の素性を加えた場合である．Twitter 上の不具合情報には筆者の感情が表現されている場合があり，特に泣いた顔文字や悲しんでいる顔文字などがよく見られたため素性に追加した．若干ながら適合率が上昇していることから，不具合情報と顔文字の間にはなんらかの関係があると考えられる．しかし，今回は顔文字の種類や極性までは細分化していないため，負例に出現する顔文字も同じく顔文字と認識している．再現率が低下しているのはそのためであると考えられる．より顔文字を素性として重要視する場合は，顔文字をさらに感情や極性（ポジティブ・ネガティブ）といったクラスに分けて扱う方法が考えられる．しかし，顔文字の種類は非常に多様であり最近では非常に複雑な顔文字も出現している．また感情や極性が一意に決まらない場合も多く，網羅的に辞書化をするのは困難である．

次に，再現率で最も良い性能だったのが感動詞の有無である．これは，事前に Twitter 上の不具合情報を分析した際に比較的よく見られたため素性として追加した．不具合の発生による驚きや落胆といった感情が表れているものと考えられ，不具合の発生を示す強い手がかりになるのではと期待した．解析器用辞書を拡張した場合だけでなく，辞書を拡張しなかった場合でも再現率で最良の結果となっていることから，不具合情報の判別において感動詞・感嘆詞を考慮することは有効であると考えられる．また，今回は解析器による解析結果を基に感動詞かどうかを判定したが，Twitter 上ではさらに強い感情の表れとして，長音化を用いた「叫び」のような表現が出現することもあり，それらを認識することができればより強い特徴として用いられることができると考えられる．

解析器用辞書の拡張をした場合の F 値の値が最も高かったのは，疑問形とネットスラングの素性である．不具合情報における疑問形には，筆者の驚きや不安といった感情が込められていると考えられ，これも比較的よく出現していた．今回の結果では良い性能であるものの，疑問形は不具合情報にかぎらず幅広く出現するため，より不具合情報らしさの特徴として扱うならば，感嘆詞などと組み合わせてより限定的な特徴とする必要がある．

ネットスラングは Web 特有の特徴であり，これは不具合情報であるかに関わらず Twitter 上のテキストには幅広く出現する．特に不具合情報には，がっかりした様子を模した記号である「orz」や，「終わった」を言い換えた「オワタ」などのネットスラングがよく出現する傾向にある．顔文字と同様に幅広く出現する特徴ではあるが，顔文字に比べると不具合情報において出現するネットスラングにはかなり偏り，すなわち不具合情報で頻出する傾向があり，そのため，より有効な特徴として機能したと考えられる．適合率の値がやや低いのは，顔文字と同様，ネットスラングという大きなくくりで素性として扱ったためであると考えられる．ネットスラングは，種類が多様であるのに加え，日々新たな言葉が生まれていくため，顔文字と同様網羅的に収集しづらいという問題がある．しかし，不具合情報に出現するものにだけ着目すると，その種類や傾向は比較的限定されているため，不具合情報らしさを測る重要な特徴として利用可能であると考えられる．

文長とモダリティの素性に関しては，解析器用辞書の拡張の有無にかかわらずあまり効果は見られなかった．この結果から，ツイートの長さは不具合情報らしさに影響しないということがわかる．また，モダリティに関しては目立った効果はないものの，実際の不具合事例では比較的よく

出現する特徴であり，他の素性と組み合わせてみるなどして組み込み方を工夫する必要があると考える．

〈6･3〉 エラー分析

失敗した事例から，Twitter 上の不具合情報を抽出する難しさがどこにあるのかを考察する．ここでは，ドメイン知識の不足に依存した問題と，機械学習だけでは解けない意味的関係認識に依存した問題の 2 つに分けて議論する．

〈6･3･1〉ドメイン知識の不足に関する

ドメイン知識の不足に依存した問題とは，不具合の症状を意味する表現を，それがその製品にとって不具合であると認識できない問題である．例えば次のような例がある．

d. スマホが急に固まってしまった

人間が見ればすぐに不具合であるとわかる事例であるが，これを不具合だと認識するにはスマートフォンというドメインにとって「急に固まること」が異常であるという前提知識を持っている必要がある． しかし，「固まる」という表現が常にいつも不具合を示すわけではない．このように，特定の製品や特定の状況でのみ不具合を示す表現となりうる場合が存在するため，それらを正しく分類するためには，事前に製品と不具合を示す表現の組み合わせについて知識を持っておく必要がある．

〈6･3･2〉意味的関係の認識が必要な問題

意味的関係認識とは，文内の各単語や文節間にある意味的な関係を認識するタスクである[9][10]．意味的関係認識が必要になる事例は，単純に機械学習を用いただけでは解決が困難である．例えば，次のような例が考えられる．

e. この携帯は買ったばかりなのにすぐ電源が落ちる
f. この携帯はもう古いからすぐ電源が落ちる

上の例は，e は不具合情報であるが f は不具合情報ではない．しかし，互いに「すぐ電源が落ちる」という不具合情報らしさを持つ記述が含まれている． e を不具合情報と判断できる要因は「買ったばかりである」という前提条件が付加されているためである．このように，文の各要素や単語だけでなく，文全体の意味的な関係を考慮して初めて不具合情報だと断定できる事例も存在する．この問題は単純に機械学習を適用するだけでは解決が困難であり，文の要素の意味的関係の認識や，文内の因果関係を認識する技術[11]が必要になる．

7. まとめと今後の展望

本研究では，Twitter からの不具合情報抽出を目的とし，Twitter ならではの素性を考慮した機械学習による抽出実験を行った．また，Twitter 上のテキストを解析するにあたって，解析器用の辞書を作成し，解析器を Twitter 用に調整した．実験の結果としては 8 割ほどの精度で抽出ができたものの，今回加えた各素性の影響はまだ小さく，素性の組み込み方には一考の余地が見られた．また，解析器の調整による変化もごくわずかであった．

失敗した事例を基に，Twitter からの不具合情報の抽出における難しさについて考察を行った．大きな問題として，ドメイン知識の不足に関する問題と，意味的関係の認識が必要な問題の 2 つが挙げられる．ドメイン知識に関する問題は，知識を拡張させることで対処可能だが，意味的関係の認識が必要な問題に関しては，単純に機械学習を用いるだけでは解決できない．機械学習による手法に加え，文内の意味的関係や因果関係を認識する技術を用いる必要があると分かった．

今後は，顔文字やネットスラングなどの Twitter ならではの素性をより有効に活用できるよう素性への組み込み方を検討するとともに，意味的関係認識が必要な問題を解決する方法について考えて行きたい．

文　　献

（1） 栗納裕貴，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） 奥村学：「マイクロブログマイニングの現在」，電子情報通信学会技術研究報告,NLC，言語理解とコミュニケーション，111.427: 19-24 (2012)
（6） Brody, Samuel, and Nicholas Diakopoulos："Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!：using word lengthening to detect sentiment in microblogs.", Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 562-570 (2011)
（7） Vladimir N. Vapnik : "The Nature of Statistical Learning Theory", Springer,New York (1995)
（8） Joachims T : "Making large-Scale SVM Learning Practical. In Advances in Ker-nel Methods-Support Vector Learning", chapter 11, MIT Press (1999)
（9） 村上浩司，水野淳太，後藤隼人，大木環美，松吉俊，乾健太郎，松本裕治：「文間意味的関係認識による言論マップ生成」，言語処理学会第 16 回年次大会， PA2-22 (2010)
(10) 古塩貴行，徳久雅人，村上仁一，池原悟：「テキストを対象とした情緒推定のための事象間関係の解析に向けて」，言語処理学会第 12 回年次大会発表論文集 (2006)
(11) 乾孝司，乾健太郎，松本裕治：「接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得」，情報処理学会論文誌，45.3: 919-933 (2004)