Twitterにおける誤情報の拡散収束過程の可視化

渡邉研斗† 鍋島啓太† 水野淳太† 岡崎直観†‡ 乾健太郎†
東北大学† 科学技術振興機構 さきがけ‡
{kento.w, nabeshima, junta-m, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

東日本大震災では，電話やメールよりも大量の情報を即座に得られるTwitterが情報提供・収集に大きく貢献した．しかし，必要以上の情報が行き交い，「コスモ石油の爆発で有害な雨が降る」などの誤情報が拡散し，人々の混乱の要因となった [1, 2]．これまで震災時の誤情報を収集する研究 [3, 4] が報告されているが，誤情報の拡散を防ぐための方策にまで踏み込んだ研究は少ない．そこで，本研究では以下の3つの課題に取り組む．

- 誤情報の発生拡散収束のケース・スタディを行い，誤情報を早く収束させる要因を分析する．
- 誤情報の発生から訂正情報が流れるまでの時間や，誤情報の拡散を食い止めるまでの時間を計測し，集合知に基づく誤情報の自動検出手法の実用性を探る．
- 誤情報の自動検出手法として，訂正表現を用いたシンプルな手法を提案し，その性能を評価する．

2 誤情報はどのように拡散し，収束したのか？

まず，東日本大震災で拡散した誤情報の発生，拡散，収束のケース・スタディを行った．分析では，東日本大震災ビッグデータワークショップでTwitter Japanより配布された震災直後一週間分の全ツイートを対象に，鍋島ら [3] の手法で獲得した14件の誤情報を用いた．各誤情報（例えば「コスモ石油の爆発で有害な雨が降る」）に対し，適切に検索クエリ（例えば「コスモ石油 AND 雨」）を選ぶことにより，誤情報を拡散するツイート，訂正するツイートの両方を区別せずに収集した．それらのツイートを誤情報（誤情報を拡散・支持する情報），訂正情報（誤情報を訂正・阻止する情報），その他（誤情報に言及していない情報）に手作業で分類した (表1)．全部で5195件のツイートを対象とし，2462件の誤情報ツイート，2376件の訂正情報ツイート，357件のその他のツイートを同定した．

このように用意したコーパスを用いて，各誤情報の発生から訂正情報が出始めるまでの時間 (TTC: Time To Correction) や，誤情報の数が減り始めるまでの時間 (TTS: Time To Suppress) を計測した．表1のコスモ石油の例のように，長い時間拡散し続ける誤情報は社会に大きな損害をもたらす．しかし，誤情報の発生から最初の訂正情報が出始めるまでの時間は，概ね数時間である．よって，訂正情報をうまく捉えることで，誤情報の検出と注意喚起を行うことは十分可能である．

更に，誤情報の個別のケースを詳しく調べるため，誤情報の「拡散」ツイートと「訂正」ツイートの数を，それぞれ一定時間おきに折れ線グラフにプロットし，誤情報の拡散状況を可視化するシステムを開発した．このシステムでは，各時点でどのようなツイートが拡散していたのか，ツイート本文を閲覧できるようになっている．なお，グラフにプロットするツイートの数はリツイート数も考慮し，ツイート空間上での情報の拡散状況を表している．14件の誤情報に対して，本システムで可視化したグラフを観察すると，誤情報の拡散状況は，主に訂正ツイートの量と収束までの時間で特徴づけられることが分かった（図1）．

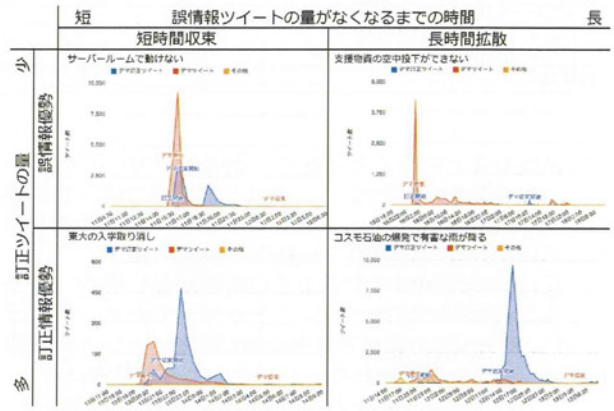

図 1: 4種類に分けられる拡散状況

これらの2つの要素の組み合わせにより，大きく4種類の拡散収束状況に分類できる．例えば，誤情報ツイート数が訂正ツイート数を上回り，かつ誤ツイート量が0になるまでの時間が1日未満なら，訂正情報劣勢・短時間収束型である．以降では，訂正情報の数や収束までの時間を決定づける要因について考察する．

訂正情報の量を決める要因: 誤情報より訂正情報の量が少ない場合，訂正情報の信憑性・影響力が小さいことが考えられる．例えば「支援物資の空中投下は法律で認められていない」という誤情報において，「許可があれば可能」という訂正情報が流れたが，決定的な証拠や公式発表がなかった．このため，危機意識に駆られた人々が誤情報をどんどん拡散し，訂正情報が浸透しなかった．
逆に訂正情報の量が誤情報より多い場合，訂正情報の信憑性・影響力が強いことが考えられる．例えば「被災者の合格者が期限までに書類を提出できないと東大の入学が取り消される」という誤情報に対し，東大がウェブサイト上で「合格者本人の意志を確認せずに入学の資格を取り消すようなことはありません」と発表した．人々の不安を取り除くに十分な訂正情報だったため，誤情報よりも訂正情報の量が多くなり，誤情報を効果的に抑制することができた．

収束までの時間を決める要因： 誤情報が収束するまでの時間が短いということは，有効な訂正情報が早期

表 1: 誤情報の拡散収束過程の分析 (TTC: 訂正情報が出始めるまでの時間, TTS: 誤情報の数が減り始めるまでの時間)

誤情報	誤情報拡散からの時間 (h)		正解データの比率			自動分類の評価		
	TTC	TTS	誤情報	訂正情報	その他	再現率	適合率	F1
トルコが 100 億円支援	1.5	3.5	100	43	17	0.914	1.000	0.955
ONEPEACE 作者尾田栄一郎が 15 億円寄付	2	1	170	134	7	0.902	1.000	0.949
コスモ石油の爆発で有害な雨が降る	2	16	382	499	98	0.871	0.995	0.929
阪神大震災では三時間後に最大の揺れが来る	0.5	2	506	84	20	0.797	1.000	0.887
阪神大震災でレイプが多発した	2	1.5	69	82	1	0.841	0.932	0.884
ポケモンクリエーターの田尻智が死去	0.5	0	2	36	3	0.750	1.000	0.857
支援物資の空中投下が認められていない	0.5	32.5	38	58	69	0.741	0.977	0.843
サーバーラックが倒れて動けない	0	1	742	401	12	0.678	0.996	0.807
フジテレビの募金は日本ユニセフに行く	0	4	82	64	7	0.578	0.948	0.718
放射線対策にイソジン（うがい薬）が利く	9.5	32.5	162	700	63	0.490	0.985	0.654
東大が合格者の入学取り消し	0.5	1.5	140	81	28	0.419	0.918	0.576
天皇陛下が京都御所へ避難	12	0	25	129	17	0.108	0.875	0.193
福島第一原発が核爆発の恐れ	0	0.5	16	45	13	0.111	0.625	0.188
辻本補佐官が米軍の救助活動に抗議	2	0	28	16	2	1.000	0.380	0.551
平均	2.3	6.9				0.657	0.902	0.714

に現れたということである．例えば「サーバールームで身動きが取れない」という誤情報では，この情報の発信者がジョークとしてつぶやいていたことが早期に発覚したため，短期間で収束した．

逆に誤情報が収束するまでの時間が長い場合，有効な訂正情報の出現が遅いことが考えられる．例えば「コスモ石油の爆発で有害な雨が降る」という誤情報は，コスモ石油本社からの「そのような事実はない」という発表が遅れたため，長時間にわたり拡散した．

このように，真偽の検証に必要な情報の信憑性・入手性により，誤情報の拡散と訂正の過程が変化する．

3 誤情報の拡散・訂正ツイートの自動分類

前節では，ツイートを誤情報，訂正情報，その他に人手で分類した．しかし，情報の混乱が発生している際に，大量のツイートを手作業で分類することは現実的ではない．そこで，誤情報に関するツイート群を，誤情報の「拡散」もしくは「訂正」に自動的に分類した．本研究では，「デマ」や「風説」などの訂正表現を含むツイートを「訂正情報」とし，含まないものを「訂正情報ではない」ツイートとした．訂正表現は震災時のツイートを読みながら，121 個を用意した．

表 1 に，本手法が訂正情報を認識する精度（再現率・適合率・F1 スコア）を示した．この評価では，リツイートは削除し，オリジナルのツイートのみ用いている．表 1 によると，訂正表現に対するマッチングという単純な手法でも，高い適合率が得られることが分かる．つまり誤情報に関するツイートにおいて，「デマ」などの訂正表現を含むものは，かなりの確度で訂正情報と見なせる．「デマ」という語を伴って誤情報の拡散を行うことは，通常では考えにくいので，これは直感的に理解できる．

しかし，量は少ないものの，訂正表現を含む誤情報拡散ツイートも見受けられた．

> 万が一原発から放射能が漏れ出した際，被爆しない為にイソジンを１５cc飲んでおいて下さい！原液です！ガセネタではありません．お医者さんからの情報です．これは RT ではないので信じてください！

このツイートでは，「ガセ」という訂正表現を含んでいるが，「ガセ」をさらに否定しているので，二重否定により誤情報の拡散ツイートと解釈できる．さらに，訂正表現を用いずに誤情報を否定するツイートも存在する．

> 厚生労働省です不特定多数の方に送信されている，コスモ石油千葉製油所における火災関連のメールについては，厚生労働省からの発表情報ではありませんのでご留意願います

このツイートでは，「デマ」「嘘」などの訂正表現は一切使われておらず，また誤情報の内容（「コスモ石油の火災により有害物質の雨が降る」）も説明していないが，内容から誤情報を訂正するツイートであると判断できる．このようなツイートを訂正ツイートと認識するためには，深い言語解析（例えば「火災関連のメール」を「火災により有害物質の雨が降るというチェーンメール」と解釈する）や，ツイートやユーザ間の関係（例えば，厚生労働省はこの誤情報に関連して別のツイートで訂正表現を用いて打ち消しを行った，等の手がかり）を用いる必要がある．

4 おわりに

本研究では東日本大震災時の内容を誤情報，訂正情報，その他に手作業で分類し，誤情報の発生・拡散・収束の過程を調査した．また，訂正表現に基づく誤情報の自動検出の可能性を，実用面，技術面から検証した．さらに，訂正表現のマッチングにより，誤情報と訂正情報の自動分類を行う手法を提案し，評価を行った．この手法は非常にシンプルであるが，高い精度を達成することができた．今後は誤情報・訂正情報の分類のリアルタイム化や自動分類の精度向上などに取り組む予定である．

謝辞

本研究は，文部科学省科研費（23240018），文部科学省科研費（23700159），および JST 戦略的創造研究推進事業さきがけの一環として行われた．貴重なデータを提供して頂いた Twitter Japan 株式会社に感謝いたします．

参考文献

[1] 萩上チキ. 検証 東日本大震災の流言・デマ. 光文社, 2011.

[2] 情報支援プロボノ・プラットフォーム（iSPP）. 3.11　被災地の証言ー東日本大震災　情報行動調査で検証するデジタル大国・日本の盲点ー. インプレスジャパン, 2012.

[3] 鍋島啓太, 水野淳太, 岡崎直観, 乾健太郎. 訂正パターンに基づく誤情報の抽出と集約. 情報処理学会 第 75 回全国大会 発表論文集, 2013.

[4] 宮部真衣, 梅島彩奈, 灘本明代, 荒牧英治. 流言情報クラウド：人間の発信した訂正情報の抽出による流言収集. 言語処理学会 第 18 回年次大会 発表論文集, pp. 891–894, 2012.

マイクロブログユーザからの現地被災者抽出の技術的支援

水野淳太[†]　岡崎直観[†‡]　乾健太郎[†]
東北大学情報科学研究科[†]　科学技術推進機構さきがけ[‡]
{junta-m, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

東日本大震災における情報伝達手段として，ツイッターをはじめとするマイクロブログは大きな役割を果たした．被災地で発生した問題や要望などは，今後の災害への対策に有用であると考えられている [1]．そこで本研究では，震災当時のツイッターから，被災したユーザを抽出し，そのツイートの収集に取り組む．情報伝達手段や安全上の制約があったにも関わらず，震災直後に被災地から情報を発信していたユーザは少なくない．震災による被災は，火災や津波など様々であるが，本研究では津波による被災者を抽出の対象とする．すなわち，震災当時に東北3県の沿岸部に滞在していたユーザを抽出することが本研究の目的である．以下では，抽出対象となるユーザを「津波被災者」と呼ぶ．

ツイッターには，緯度経度情報をツイートに付与する機能が存在するが，この機能の利用者は少数であるため，本研究では取り扱わない．ツイート本文をもとにして発信場所推定手法についても研究が進められているが [2, 3]，いずれの推定精度も限定的であり，東北3県の沿岸部という狭い範囲の推定において，有効であるとは考えにくい．そこで本研究では，ツイート本文に含まれる住所情報，画像データを手がかりとすることで，津波被災者を効率よく見つけられることを示す．

2 津波被災者の抽出

本研究で抽出の対象とするツイートデータは，東日本大震災ビッグデータワークショップ[1]において Twitter Japan から提供された，2011年3月11日の午前9時から3月18日の午前9時までの全ツイート（179,286,297件）である．

津波で特に大きな被害を受けたのは岩手・宮城・福島の3県である．人口の比率を考えると，日本全体のツイッターユーザに対して，この3県のユーザが占める割合は小さい．さらに，この3県では停電やネットワーク障害が長期間にわたって発生しており，津波被災者からの情報発信が滞っていた可能性もある．このような理由から，ツイートデータの中から単語の頻度や共起頻度を測定し，統計的に顕著な部分に着目したとしても，津波被災者のツイートを発見するのは難しいと想像される．

津波被災者のツイートを発見することの難しさを示す一例として，ツイートデータ全体に対して，「津波」を本文に含むツイートを検索し[2]，検索された1,545,910ツイートの中でリツイート数の多いツイート100件をまとめたものを表1に示す．このツイート群の中では，注意喚起のツイート，情報提供のツイートが7割以上を占めており，津波に関する注意や情報を積極的に拡散していることが分かる．しかしながら，津波の被害を自分の体験として報告しているツイートは，この100件の中には見つからなかった．この結果から，津波被災者のツイートは，よくリツイートされるとは限らないことが分かる．そこで，ツイート本文に含まれる住所情報，画像データを利用した抽出を試みる．

表1: RT数top100のツイートの分類結果

ツイートタイプ	ツイート数
注意喚起	39
情報提供	32
賞賛	11
意見	7
救援要請	5
非難	4
ジョーク	2
合計	100

2.1 住所情報に基づく抽出

ツイッターユーザが津波被災者であるかは，そのユーザが被災地域に滞在していたかによって判断することができる．ユーザのプロフィール情報を閲覧すると，そのユーザがどの地域に住んでいるかを判断することができるが，本ワークショップにはプロフィール情報は含まれていない．そこで，ユーザのツイート内容を基にプロファイリングを行い，ユーザの滞在地を推定することが考えられる．

東北3県の沿岸部についてよく言及しているユーザは，その地域に居住あるいは滞在している可能性が高いという仮説に基づき，以下の手順によって津波被災者の抽出を行った．

1. 宮城県の主要な沿岸部（南三陸町など）を，町名の粒度で人手で15箇所を選択する．
2. 各ユーザのツイート集合に対して，15箇所の地名の本文中での出現頻度を計る．
3. 15箇所の地名のうち，20回以上言及していた地名があるユーザは，その地域に滞在していたと判断する．20回以上言及していた地名が複数ある場合は，より多く言及していた地域に滞在していたと判断する．
4. 抽出されたユーザのツイート本文を読み，沿岸部に滞在していたかを人手で判断する．

3までで，723人のユーザを抽出することができた．それらのユーザに対して，4で人手で判断したところ，15人が滞在していたと判断できた．本手法は，4でかかるコストが問題となる．723人から15人を抽出するのに約12時間かかっており，多大な労力を要する．そこで，次節ではツイートに含まれる画像データに着目した抽出手法について述べる．

[1] http://sites.google.com/site/prj311/

[2] 全文検索エンジンには Apache Solr (http://lucene.apache.org/solr/) を用い，全ツイートの本文を，文字 bi-gram で索引付けした．

図 1: 「津波」を本文に含む，画像付きツイート

図 2: 横須賀で撮影された写真付きツイート

図 3: 仙台港で撮影された写真付きツイート

2.2 画像データに基づく抽出

津波被災者のツイートを効率よく選び出す方法として，我々はツイート本文中に含まれる画像データへのリンクに着目した．震災当時のツイッター上では，被災状況や安否不明者のリストなどが，画像データとして拡散していた．そこで，津波の状況が添付されているツイートに着目することで，津波被災者の選別が出来るのではないかと考えた．

東日本大震災ビッグデータワークショップのツイートデータの中で，「津波」を本文に含み，かつ画像付きのツイートは 19,696 件であった．その一部を図 1 に示す．なお，ツイートに画像が添付されているかどうかは，本文に含まれる URL が代表的な画像投稿サービス（Twitpic や yfrog など）のものであるかによって判別した．図 1 を見ると，テレビでの報道の一部を撮影して投稿されたツイートが目立つが，津波の被害状況を撮影した写真も少なからず存在する．これらの画像は，以下のように大別できる．

被災地の写真 津波の到達前・到達時・到達後の様子，津波による被害などを撮影したもの

テレビの報道 テレビの報道番組の画面を撮影したもの

無関係の画像 被災地の応援を目的としたイラストや，津波とは無関係の写真など

このうち，テレビの報道は画面の映り込みや回転，L 字型画面，テロップなどを手がかりに，容易に判別可能である．無関係の画像は，津波以外の被害状況の写真やイラストなどが該当する．これらも人間には容易に判別

図 4: 人手で抽出した結果

可能である．このように，ツイートに添付されている画像を人間が目視確認することで，津波被災者が津波の状況を撮影した写真かどうか，迅速に判定できる．

ただし，今回の東日本大震災では広範囲の沿岸部に津波が襲来したため，津波被害を撮影した写真かどうかを判別するだけでは，東北 3 県のツイートに限定することはできない．例えば，図 2 のツイートは，「横須賀」で津波を撮影したものである．一方で，図 3 のツイートは，「仙台港」であることが明記されている．そこで，ツイートに添付されている画像に加えて，本文に含まれる地名を手がかりとし，人手で 19,696 件のツイートをチェックした．約 3 時間の作業時間で，全てのツイートに対するチェックを行うことができ，津波被災者が津波の被害状況を撮影したと思われる 28 件のツイート（28 ユーザ）を抽出できた．その一部を図 4 に示した．

本手法は，画像データを投稿したユーザのみに限定した抽出しか行えないが，迅速に判断できるというメリットがある．住所情報に基づく手法で抽出された 15 ユーザと，画像データに基づく手法で抽出された 28 ユーザに重複はなかったことから，その他の情報に着目することによって，新たな津波被災者を抽出できる可能性が示唆される．

3 おわりに

本稿では，東日本大震災当時のツイートデータに対して，本文中の住所情報や画像データを利用することで，津波被災者を抽出するための技術的支援手法について述べた．これらの支援技術により，合計で 43 名の津波被災者を抽出することができた．今後は，抽出されたユーザのツイートを用いて，新たな津波被災者をマイニングしていくことが考えられる．

謝辞

本研究は，文部科学省科研費（23240018，23700159），および JST 戦略的創造研究推進事業さきがけの一環として行われた．

参考文献

[1] 今村文彦, 佐藤翔輔, 柴山明寛. みちのく震録伝：産学官民の力を結集して東日本大震災のアーカイブに挑む. 情報管理, Vol. 55, No. 4, pp. 241–252, 2012.

[2] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. of CIKM 2010*, pp. 759–768, 2010.

[3] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. Location inference using microblog messages. In *Proc. of WWW 2012*, pp. 687–690, 2012.

マイクロブログからの誤情報の発見と集約

鍋島 啓太[†]　水野 淳太[†]　岡崎 直観[†‡]　乾 健太郎[†]
東北大学大学院 情報科学研究科[†]　科学技術振興機構 さきがけ[‡]
{nabeshima, junta-m, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

2011年3月に発生した東日本大震災において，ソーシャルメディアは有益な情報源として活躍した [1]．震災に関する情報源として，ソーシャルメディアを挙げたネットユーザーは18.3%で，インターネットの新聞社（18.6%），インターネットの政府・自治体のサイト（23.1%）と同程度の影響力を示した．

一方で，「コスモ石油のコンビナート火災に伴う有害物質の雨」に代表されるように，インターネットやソーシャルメディアがいわゆるデマ情報の流通を加速させたという指摘もある．東日本大震災とそれに関連する福島第一原子力発電所の事故では，多くの国民の生命が脅かされる事態となったため，人間の安全・危険に関する誤情報（例えば「放射性物質から甲状腺を守るにはイソジンを飲め」）が拡散した．東日本大震災に関するデマをまとめたツイート[1]では，2012年1月時点でも月に十数件のペースでデマ情報が掲載されている．このように，Twitter上の情報の信憑性の確保は，災害発生時だけではなく，平時においても急務であり，誤情報に対する注意喚起を低コストで実現する仕組みが必要である．

本論文では「○○というのはデマ」などの誤情報を訂正する表現（以下，訂正パターン）に着目し，ツイート集合から誤情報を自動的に収集する手法を提案する．提案手法を東日本大震災後1周間のツイートに適用したところ，既存のまとめサイトに収録されている60件の誤情報の約半数を再現でき，まとめサイトに収録されていない22件の誤情報を獲得することができた．

2 関連研究

ツイッターを対象とした研究は数多くあるが，本節ではツイートで発信される情報の真偽性や信憑性に関連する研究を紹介する．

Qazvinianら [2] は，誤情報に関連するツイート群（例えば「バラク・オバマ」と「ムスリム」を含むツイート群）から，誤情報に言及しているツイート（例えば「バラク・オバマはムスリムである」）と，誤情報に言及していないツイート（例えば「バラク・オバマがムスリムのリーダーと面会した」）を分類し，さらに誤情報に関して言及しているツイート群を，誤情報を支持するツイートと否定するツイートに分類する手法を提案した．Qazvinianらの研究は，誤情報に関連するツイート群（もしくはクエリ）が与えられることを想定しており，本研究のように大規模なツイートデータから誤情報をマイニングすることは，研究対象の範囲外である．

日本では，東日本大震災時にツイッター上で誤情報を拡散したという問題意識から，関連する研究が多く発表されている．藤川ら [3] は，ツイートに対して疑っているユーザがどの程度いるのか，根拠付きで流言であると反論されているか等，情報に対するユーザの反応を分類することで，情報の真偽判断を支援する手法を提案した．鳥海ら [4] は，あるツイートの内容がデマかどうかを判別するため，ツイートの内容語と「デマ」「嘘」「誤報」などの反論を表す語の共起度合いを調べる手法を提案した．梅島ら [5] は，東日本大震災時のツイッターにおけるデマと，デマ訂正の拡散の傾向を分析することを目標とし，「URLを含むリツイートはデマである可能性が低い」「デマは行動を促す内容，ネガティブな内容，不安を煽る内容が多い」「この3つのいずれかの特徴を持つツイートはリツイートされやすい」等の仮説を検証した．彼女らのグループはその後の研究 [6, 7] で，誤情報のデータベースを構築するために，「デマ」や「間違い」といった訂正を明示する表現を用いることで，訂正ツイートの認識に有用であることを示した．さらに彼女らは，訂正を明示する表現を含むツイートを収集し，各ツイートが特定の情報を訂正しているか，訂正していないのか[2]を識別する二値分類器を構築した．

これらの先行研究は，ツイートの本文を単位とし，誤情報を含むか，もしくは特定の情報を訂正しているかどうかを認識することに注力しており，ツイート本文中から誤情報の箇所をピンポイントで特定しているわけではない．したがって，大規模なツイートデータから誤情報を網羅的に収集する研究は，我々の知る限り本研究が最初の試みである．

3 提案手法

図1に提案手法の流れを示す．手順は大きく4つに分けられる．以降では，各ステップについて説明を行う．

ステップ1 被訂正フレーズの抽出: ステップ1では，ツイート本文から被訂正フレーズを見つけ出す．被訂正フレーズとは，「イソジンは被曝を防げるというのはデマだ」の下線部のように，「デマ」や「間違い」といった訂正表現で打ち消されている箇所のことである．被訂正フレーズと訂正表現は，「という」や「のような」といった連体助詞型機能表現で繋がれており，被訂正フレーズに続く表現を「訂正パターン」と呼ぶ．人手で作成した368個の訂正パターンのいずれかにマッチするツイート本文に対して，文頭から訂正パターンの直前までを被訂正フレーズとして抽出する．本ステップをツイート全体に適用し，抽出した被訂正フレーズの集合を D とする．

ステップ2 キーワードの抽出: 前節で抽出された被訂正フレーズには，「昨日のあれはデマだ」の「昨日のあれ」

[1] https://twitter.com/#!/jishin_dema

[2] 例えば「ツイート上には様々なデマが流れているので注意を！」というツイートには「デマ」という表現を含んでいるが，特定の情報を訂正しているわけではない

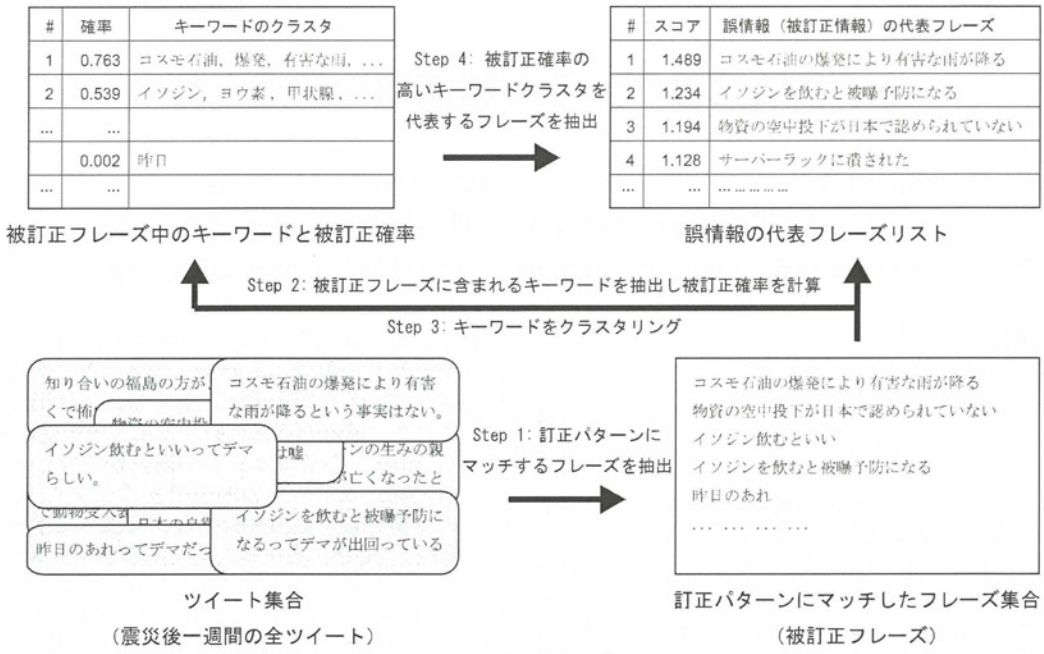

図 1: 提案手法の流れ

のように，具体的な情報に言及していないフレーズも含まれている．これらは誤情報としては不十分であるため，取り除く必要がある．そこで，被訂正フレーズ中の単語が訂正パターンとよく共起しているかどうかを調べる．具体的には，ある語 w がツイートで言及されるとき，その語が被訂正フレーズ集合 D に含まれる条件付き確率，

$$P(w \in D|w) = \frac{w\text{が訂正パターンと共起するツイート数}}{w\text{を含むツイート数}} \quad (1)$$

を算出し，確率が高い上位 500 単語を誤情報のキーワードとして選択する．

ステップ 3 キーワードのクラスタリング: 被訂正フレーズには，「コスモ石油の火災により有害物質を含む雨が降る」と「コスモ石油の爆発は有害だ」のように，同一の情報に言及しているが，表現や情報量の異なるフレーズが含まれている．誤情報を重複なく抽出するために，これらをまとめる必要がある．そこで，ステップ 2 で抽出されたキーワードをクラスタリングする．キーワード間の距離（類似度）として，キーワードと文内で共起する内容語（名詞，動詞，形容詞）を特徴量とした文脈ベクトルのコサイン距離を用いた．文脈ベクトルの特徴量には，キーワードと各単語との共起度合いを測定する尺度である自己相互情報量を用いた．クラスタリング手法として最長距離法を用いた．各クラスタにおいて，ステップ 2 の条件付き確率が高いものを代表キーワードとする．

ステップ 4 代表フレーズの選択: 前ステップで得られた各クラスタに対し，そのクラスタ中のキーワードを含む被訂正フレーズの中で代表的なものを選択し，誤情報として出力する．誤情報を過不足なく説明できる被訂正フレーズを選択するため，以下の式でスコアを計算する．

$$\mathrm{score}(s,t) = \mathrm{hist}(\mathrm{len}_s, t) \times \sum_{w \in C_s} \mathrm{PMI}(t,w) \quad (2)$$

ここで，s は被訂正フレーズ，t は誤情報クラスタを代表するキーワード，C_s は s 中の内容語の集合，len_s は被訂正フレーズ s の単語数を示す．$\mathrm{hist}(l,t)$ は，最重要キーワード t を含み，かつ単語数が l である文の出現頻度，$\mathrm{PMI}(t,w)$ は t と単語 w の自己相互情報量を示す．式 (2) は，キーワードとよく共起する内容語を多く含み，かつ標準的な長さの被訂正フレーズに対して，スコアが高くなるように設計されている．すなわち，$\mathrm{hist}(\mathrm{len}_s, t)$ は，最重要キーワードを含むフレーズの中で典型的な長さのフレーズに高いスコアを与え，極端に短いフレーズ・長いフレーズに対して低いスコアを与える補正式である．

4 実験 1：訂正パターンの評価

提案手法は，訂正パターンで表明されない誤情報を獲得することができない．そこで本節では，人手で整備した訂正パターンの性能を評価する．

4.1 実験設定

誤情報の抽出元となるコーパスには，東日本大震災ビックデータワークショップ[3] で Twitter Japan から提供された 2011 年 3 月 11 日 9 時から 2011 年 3 月 18 日 9 時までの 179,286,297 ツイートを利用した．評価実験の正解データとして，誤情報を人手でまとめた 4 つのウェブサイト[4]に掲載されている事例のうち，震災後 1 週間以内に発信されたと判断できる 60 件の誤情報を用いた．

訂正パターンは，適合率と再現率で評価した．収集した被訂正フレーズ集合約 2 万件からランダムに 150 件サンプリングし，その中で発信者が訂正パターンで情報を否定・訂正していると判断できる割合を適合率とした．再現率は，収集した被訂正フレーズ集合約 2 万件によって正解データの誤情報 60 件をカバーできた割合とした．

[3] https://sites.google.com/site/prj311/

[4] 収集したサイトは以下の通り
http://www.kotono8.com/2011/04/08dema.html
http://d.hatena.ne.jp/seijotcp/20110312/p1
http://hara19.jp/archives/4905
http://matome.naver.jp/odai/2130024145949727601

表 1: 訂正パターンの適合率と再現率

適合率	再現率
0.79 (118/150)	0.83(50/60)

表 2: 抽出された被訂正フレーズの内訳

被訂正フレーズの種類	件数
(あ) 情報を訂正していると判断できる被訂正フレーズのうち，内容が十分なもの	76
(い) 情報を訂正していると判断できる被訂正フレーズのうち，内容が不十分なもの	42
(う) 誤抽出のうち，パターンが曖昧な事例	24
(え) 誤抽出のうち，著者の態度が不明な事例	8
合計	150

4.2 結果と分析

表 1 に訂正パターンの適合率と再現率を示す．約 8 割の適合率，再現率で誤情報を抽出することができた．表 2 に抽出された被訂正フレーズの内訳を示す．

(あ) と (い) は表 1 の評価で正解と判断した事例である．そのうち，(い) は「昨日のあれはデマだ」の「昨日のあれ」のように，具体的な情報に言及していないフレーズや，「イソジンの件ってデマだったのか。」の「イソジンの件」のように説明が不足している事例である．ステップ 2 の条件付き確率によるランキグや，ステップ 4 の代表フレーズの選定を行うことで，(い) のような訂正フレーズを取り除くことができると考えられる．

(う) と (え) はどちらも誤って抽出された事例である．そのうち，(う) は「こういう災害の時ってデマがよく流れる」のように，訂正パターンの用法の違いにより訂正されていないフレーズを抽出した事例である．(え) は「募金するとモテるってデマを流せばいい」のように，訂正パターンに続く表現により，著者の訂正に対する態度が曖昧になっている事例である．

また，抽出出来なった誤情報 10 件を調査したところ，表 3 にある 3 つに分類することができた．

(お) は今回整備した訂正パターンでは網羅できなかった事例である．例として「天皇が 24 時間御祈祷に入ってるってのはソースがない」の下線部の訂正パターンは，今回整備した訂正パターンには含まれていなかったが，今後パターンを拡充することで抽出できる．

(か) は本研究が対象とする訂正パターンの型によらず，誤情報を訂正した例である．例として，「日本に韓国が借金の申し出。しかも管は快諾」という誤情報に対して以下のような訂正ツイートが存在した．

> これデマなんじゃ?ソースないし。RT @xxx　RT こんな非常事態の日本に韓国が借金の申し出。しかも管は快諾！

この例のように，元のツイートにコメントする形で，情報を訂正するツイートがいくつか見られた．

(き) の誤情報は今回の実験で用いたツイート内に存在するが，それに対する訂正ツイートが存在しない事例である．本手法は，誤情報には何らかの訂正ツイートが存在することを前提としているため，抽出は困難であるが，その数は少ない．

5 実験 2：誤情報の集約の評価

本節では，3 節のステップ 2 から 4 を評価する．前節で抽出された被訂正フレーズ集合から，(い) に含まれる具体的な情報に言及していない被訂正フレーズが取り除けたか，誤情報を過不足なく説明する被訂正フレーズを抽出できたか，という観点で評価をする．

表 3: 抽出できなかった誤情報の内訳

原因	件数
(お) 新しい訂正パターンが存在	3
(か) 訂正ツイート内に手がかりあり	4
(き) 訂正ツイートなし	3
合計	10

表 4: 抽出された誤情報の精度・再現率

N	精度 (4 サイト)	精度 (人手判断)	再現率
25	0.44(11/25)	0.64(16/25)	0.18(11/60)
50	0.34(17/50)	0.58(29/50)	0.28(17/60)
75	0.33(25/75)	0.56(42/75)	0.42(25/60)
100	0.30(30/100)	0.52(52/100)	0.50(30/60)

5.1 実験設定

提案手法で抽出された誤情報の正否は，同等の内容がまとめサイトに掲載されている 60 件の正解データに含まれるかどうかを一件ずつ人手でチェックを行うことで判定した．また，これらの 4 つのまとめサイトに収録されていないが，誤情報であると判断できるものもある．そこで提案手法が抽出した情報が正解データに含まれなかった場合は，人手で調査を行い，実際には誤情報だったのか判断した．本研究の目的は，誤情報を網羅的に抽出することであるので，抽出した誤情報のうち，同じ内容と判断できるものが複数ある場合，正解は 1 つとした．評価方法について，提案手法はスコアの高い順に N 件まで出力可能であるため，N を変化させたときの精度，再現率を計測した．

5.2 実験結果と分析

評価結果を表 4 に示す．N が 100 のとき，提案手法が抽出した情報のうち，正解データにも存在する情報は 3 割である．さらに，今回の正解データには含まれないが，誤情報と判断できる事例が約 2 割あり，提案手法は約 5 割の適合率で誤情報を抽出できた．不正解だった事例のうち，約半数は同じ誤情報を別のフレーズで表現したもの (重複) が占めるため，提案手法が抽出する誤情報の約 7 割は正解と見なすことができる．

抽出された誤情報の上位 100 件のうち，不正解であった 48 件の誤判定の原因を調べたところ，6 種類の原因に分類できた．表 5 に理由と件数を示す．

(a) から (d) は，明らかに抽出誤りと判断できる事例である．(e) と (f) は，人間でも誤情報であるか判断が難しい事例である．以下でそれぞれの詳細と，改善案を述べる．

(a) トピック抽出による誤り
「なんちゃら」，「どさくさ」，「○○」といった，誤情報を説明する中心的なキーワードとしては不適切な単語を抽出してしまったことが原因である．対策としては，ステップ 2 で，ひらがなのみで構成される単語や，記号の含有率が高い単語などを，キーワードとして抽出しないことが考えられる．

(b) クラスタリングによる誤り
抽出された誤情報上位１００件のうち，同じ内容と判断できる誤情報が重複している事例である．例を以下に示す．括弧の中は，選定に利用したトピック単語である．

> 市原市のコスモ石油千葉製油所 LPG タンクの

表 5: 精度に対する誤り分析

原因	件数	割合 (%)
(a) トピック抽出による誤り	12	25.0
(b) クラスタリングによる誤り	20	41.7
(c) 内容が不明確な情報	5	10.4
(d) 正しい情報	1	2.1
(e) 未来予測	5	10.4
(f) 真偽不明	5	10.4
合計	48	100.0

爆発により，千葉県，近隣圏に在住の方に有害な雨などと一緒に飛散する (コスモ石油千葉製油所)

千葉県の石油コンビナート爆発で，空気中に人体に悪影響な物質が空気中に舞い雨が降ると酸性雨になる (石油コンビナート爆発)

これはステップ 3 でクラスタリングを行ったとき，同じクラスタに分類できなかったため，重複して表れた．トピック単語のクラスタリングには，被訂正フレーズの中で共起する単語を素性としているが，素性にキーワードそのものの表層の情報を加えることで，誤りを減らすことができると考えられる．

(c) 内容が不正確な情報

抽出された誤情報の内容が，誤情報を説明するのに内容が不足していると思われる事例である．以下に例を示す．

餓死者や凍死者が出た．

正解データの中には「いわき市で餓死者や凍死者が出た」というものが存在するが，それと比べると具体性に欠けているため，不正解とした．これらの被訂正フレーズを含むツイートの数が少ないため，閾値を設けて取り除く必要がある．

(d) 正しい情報

誤情報として抽出されたが，事実を確認したところ，誤情報ではなかった事例である．以下に例を示す．

東京タワーの先端が曲がった

これは突拍子のない話だったため，誤情報と思った人が多かったと考えられる．しかし事例数は 100 件中 1 件と少ないので，他に比べあまり問題ではないと考える．

(e) 未来予測未来に起こる事象について述べたもの抽出した事例である．以下に例を示す．

福島で核爆発が起こる

(f) 真偽不明

いくつかのウェブサイトを検索して調査したところ，誤情報かどうかを判別できなかった事例である．以下に例を示す．

サントリーが自販機無料開放

次に，正解データにある誤情報 60 件のうち，候補中には含まれるが，抽出されなかった誤情報 20 件についても同様に原因を調査したところ，2 つに分類できることが分かった．2 つの原因の件数と割合を表 6 に示す

(g) クラスタリングによる誤り

訂正パターンにより候補の抽出はできたが，クラスタリングにより，誤って他の誤情報に含まれた事例である．しかし，全体に比べ事例数が少ないため，それほど問題ではないと思われる．

表 6: 再現率対する誤り分析

原因	件数	割合 (%)
(g) クラスタリングによる誤り	2	10.0
(h) ランキング外	18	90.0
合計	20	100.0

(h) ランキング外これは訂正パターンにより候補を抽出できたが，条件付き確率が低かったため，キーワードとして抽出できなかった事例である．例えば，「東京電力を装った男が表れた」という誤情報では，「東京電力」というキーワードは誤情報以外の話題でも頻出したため，条件付き確率が低くなった．対策としては，キーワード単独をスコアリングするのではなく，被訂正フレーズそのものをスコアリングするような手法が必要である．

6 おわりに

本研究では，誤情報を訂正する表現に着目し，誤情報を自動的に収集する手法を提案した．実験では，誤情報を人手でまとめたウェブサイトから取り出した誤情報のリストを正解データと見なして評価した．抽出された情報の中には，まとめサイトに掲載されていない誤情報も存在し，提案手法は誤情報の自動収集に有用であることが分かった．今後は，訂正パターンの拡充や被訂正フレーズのスコアリングの改良を進め，誤情報抽出の性能を向上させるとともに，リアルタイムでの誤情報獲得に取り組む予定である．

謝辞

本研究は，文部科学省科研費 (23240018)，文部科学省科研費 (23700159)，および JST 戦略的創造研究推進事業さきがけの一環として行われた．データを提供して頂いた Twitter Japan 株式会社に感謝いたします．

参考文献

[1] 野村総合研究所. プレスリリース：震災に伴うメディア接触動向に関する調査. http://www.nri.co.jp/news/2011/110329.html, 2011.

[2] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 1589–1599, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[3] 藤川智英, 鍛治伸裕, 吉永直樹, 喜連川優. マイクロブログ上の流言に対するユーザの態度の分類. 言語処理学会第 18 回年次大会, 2012.

[4] 鳥海不二夫, 篠田孝祐, 兼山元太. ソーシャルメディアを用いたデマ判定システムの判定精度評価. デジタルプラクティス, Vol. 3, No. 3, pp. 201–208, jul 2012.

[5] 梅島彩奈, 宮部真衣, 荒牧英治, 灘本明代. 災害時 twitter におけるデマとデマ訂正 rt の傾向. 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2011, No. 4, pp. 1–6, jul 2011.

[6] 梅島彩奈, 宮部真衣, 灘本明代, 荒牧英治. マイクロブログにおける流言マーカー自動抽出のための特徴分析. 言語処理学会第 18 回年次大会, 2012.

[7] 宮部真衣, 梅島彩奈, 灘本明代, 荒牧英治. 流言情報クラウド: 人間の発信した訂正情報の抽出による流言収集. 言語処理学会第 18 回年次大会, 2012.

Twitter上での誤情報と訂正情報の自動分類

渡邉研斗† 鍋島啓太† 岡崎直観†‡ 乾健太郎†

東北大学† 科学技術振興機構 さきがけ‡
{kento.w, nabeshima, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

東日本大震災時にTwitterは安否確認や情報交換に大いに役立った．しかし「コスモ石油の爆発で有害な雨が降る」等の誤情報がTwitter上で拡散し，社会に混乱をもたらした[1, 2]．一方，「○○というツイートはデマです」のような，誤情報を訂正・阻止する訂正ツイートも多数見られた．

我々は，Twitter上の情報をリアルタイムでモニタリングし，誤情報と思われるツイート群とその訂正ツイート群を一緒に提示することで，情報の信憑性の判断を支援するシステムを構築している．このようなシステムを実現するには，Twitter上で拡散している誤情報に関連するツイートを収集し，各ツイートが誤情報に言及しているか，誤情報を訂正しているか，判別・整理する必要がある．本研究では，鍋島ら[3]の手法を用いて，震災時に拡散した誤情報を説明する記述（例えば「コスモ石油の爆発で有害な雨が降る」や「イソジンは放射線予防に効く」）の検出と，関連キーワードの抽出が実現すると仮定する．例えば，東日本大震災時では「コスモ石油」「イソジン」などのキーワードが誤情報に多く含まれる．

これらのキーワードを含むツイートは，誤情報の拡散もしくは訂正を行っている可能性が高い．そこで，本研究ではキーワードで収集されたツイートを誤情報の支持・拡散ツイート，誤情報の反論・訂正ツイートに分類するためのコーパスを整備する．そのコーパスを用い，教師あり学習を用いて自動分類手法を提案する．評価実験では提案手法の性能を報告し，今後の課題を整理する．

2 提案手法

2.1 コーパスの準備

本研究では，東日本大震災ビッグデータワークショップでTwitter Japanより配布された震災直後1週間分の全ツイートを対象に，鍋島ら[3]の手法で獲得した14件の誤情報を説明する記述を用いた．各誤情報を説明する記述（例えば「コスモ石油の爆発で有害な雨が降る」）に対し，適切な検索クエリ（例えば「コスモ石油 AND 雨」）を選び，誤情報を拡散するツイート，訂正するツイートの両方を区別せずに収集した．なお，影響力の大きいツイートを重点的に調べるため被リツイート数の多いツイートを優先的に採用した[1]．それらのツイートに対し，誤情報（誤情報を拡散・支持する情報），訂正情報（誤情報を訂正・阻止する情報），その他（誤情報に言及していない情報）のいずれかのラベルを手作業で付与した．

手作業での分類はコストが大きいので，本研究ではクラスタリングを用いて，効率的にアノテートした．似た表現を用いたツイート群は，同一の主張である場合が多いので，まずツイート群を類似した文字列でクラスタリングした（この時点で「誤情報」・「訂正情報」・「その他」クラスタが多数生成される）．次に各クラスタ内に別の主張が混ざっていないかをチェックした（例えば「誤情報」クラスタ内に「訂正情報」のツイートが混ざっていたらクラスタを分割する）．最後に，各クラスタを「誤情報」・「訂正情報」・「その他」の3クラスタにマージした．全部で5195件のツイートを対象とし，2462件の誤情報ツイート，2376件の訂正情報ツイート，357件のその他のツイートを同定した(表1)．

2.2 分類器の構築

訂正情報には「○○という情報はデマです」のように「デマ」や「風説」のような訂正表現が含まれている可能性が高い．我々は事前研究[4]において訂正表現の有無でツイートを自動分類した．しかし，この方法では「誤情報」と「訂正情報」の分類にしか対応しておらず，誤情報とは無関係な「その他」のツイートを分類することができない．そこで，本研究では2.1節で構築したコーパスを訓練事例として，最大エントロ

[1]実際には，被リツイート数がx件以上のツイートのみを採用した．誤情報によって関連するツイート数が異なるため，閾値xは誤情報毎に調整した．

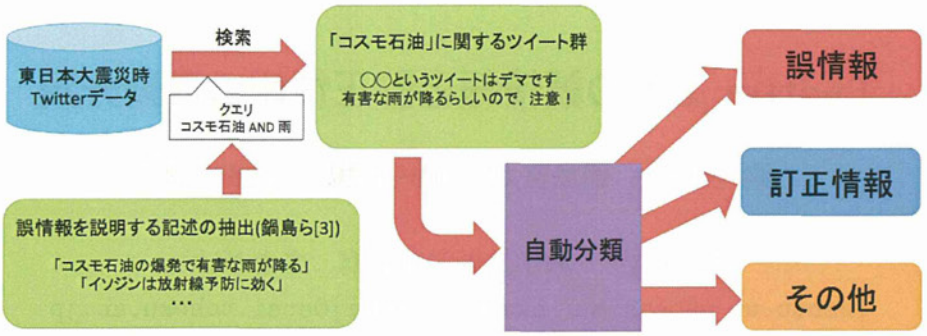

図 1: 誤情報モニタリングシステムの概要

表 1: 構築したコーパスに含まれる誤情報（トピック）と誤情報・訂正情報・その他の内訳

誤情報（トピック）	ツイート数	誤情報	訂正情報	その他
サーバーラックが倒れて動けない	1155	742	401	12
コスモ石油の爆発で有害な雨が降る	979	382	499	98
放射線対策にイソジン（うがい薬）が利く	925	162	700	63
阪神大震災では三時間後に一番大きい揺れが来る	610	506	84	20
ONEPEACE の作者尾田栄一郎が 15 億円寄付	311	170	134	7
東大が合格者の入学取り消し	249	140	81	28
天皇陛下が京都御所へ避難	171	25	129	17
支援物資の空中投下が認められていない	165	38	58	69
トルコが 100 億円支援	164	100	47	17
フジテレビの募金は日本ユニセフに行く	153	82	64	7
阪神大震災でレイプが多発した	152	69	82	1
福島第一原発が核爆発の恐れ	74	16	45	13
辻本補佐官が米軍の救助活動に抗議	46	28	16	2
ポケモンクリエーターの田尻智が死去	41	2	36	3
合計	5195	2462	2376	357

ピー法を用いて 3 クラス分類モデルを学習した．

本研究では以下の素性を設計した．

- 訂正表現の有無 (T):
  本文中に「デマ」や「風説」のような訂正表現が含まれていれば，訂正情報である可能性が大きい．本研究では，震災時のツイートから 121 個の訂正表現を手作業で収集したものを使用する．この素性は渡邉ら [4] のルールに対応するものである．
- Bag of words (B):
  拡散したい情報がある場合，ユーザは情報をそのままコピー&ペーストする可能性が高い．よって拡散される情報内には，特定の単語（「拡散希望」「コピペ」等）が用いられる傾向にある．
- URL の有無 (U):
  訂正情報の中にはしばしば誤情報であるという根拠を提示するために URL を記載している場合がある．よって URL がツイート本文中にあれば訂正情報の可能性が高い考えられる．
- 拡散 (RT @) の有無 (R):
  「RT @」が文字列が含まれている場合，ツイートを拡散させようとしているので，誤情報か訂正情報である可能性が高い．
- 訂正表現周辺の単語 (TW):
  単に訂正表現の有無のみでは，「デマではありません」などの訂正表現を否定しているツイートのように，実際は誤情報であるツイートを訂正情報にしてしまう可能性がある．よって訂正表現の周辺単語を調べることにより，それらのツイートを正しく分類できることが期待できる．本研究では訂正表現の前後 5 単語を素性として加える．
- 訂正表現から誤情報キーワードまでの距離 (D):
  ある誤情報を訂正したい時は「(誤情報キーワード)についてはデマです」のように，定型的な言い回しが多い．よって誤情報に関するキーワードから訂正表現までの距離 (文字数) が小さければ，訂正情報である可能性が高い．ここで，誤情報に関するキーワードは，2.1 節でコーパスを作成した際に用いた検索クエリ (例えば「イソジン」と「うがい薬」) とする．
- 誤情報とツイートの類似度 (SU，SB):

表 2: 提案手法の性能と素性セットによる性能の違い

素性		スコア			
		Accuracy	Precision	Recall	F1
ベースライン	T	0.7578	0.5337	0.5413	0.5204
全 8 素性		0.6562	0.5540	0.5333	0.5266
7 素性	-B	0.8125	0.5437	0.5816	0.5606
6 素性 (B を除く)	-SB	0.8181	0.5480	0.5855	0.5644
	-SU	0.8120	0.5485	0.5808	0.5601
	-D	0.8169	0.5462	0.5848	0.5637
	-TW	0.8088	0.5437	0.5787	0.5578
	-T	0.8094	0.5415	0.5793	0.5585
	-R	0.8092	0.5466	0.5789	0.5582
	-U	0.7870	0.5245	0.5634	0.5431

誤情報を説明する記述とツイート本文の類似度を素性にすることで，誤情報を支持するツイート認識をできると考えられる．本研究では，誤情報を説明する記述とツイート本文の単語ユニグラムと単語バイグラムのコサイン距離をもとに類似度を算出し，素性として用いた．(それぞれ SU，SB)

3 実験

3.1 実験設定

提案手法を評価するため，2.1 節で作成したコーパスに含まれる 14 件の誤情報（表 1）ごとに，学習データを 14 グループに分割し，交差検定を行う．つまり，「コスモ石油の爆発で有害な雨が降る」などのトピックを評価データとして，「イソジンは放射線予防になる」などのその他のトピックを学習データとして評価する．なお，評価では Accuracy をマイクロ平均で算出し，Precision，Recall，F1 をマクロ平均で算出する．

誤情報と訂正情報を自動分類する最も単純な方法は，ツイート本文中に訂正表現が存在するかどうかで分類する方法である．よって，実験でのベースラインは，素性「訂正表現の有無（T）」のみを使用した分類器の精度とする．

3.2 実験結果

表 2 に提案手法の精度，適合率，再現率，F1 スコアを示した．提案手法の全ての素性を用いた時（全 8 素性）の精度は 0.6562，マクロ F1 スコアは 0.5266 であった．訂正表現のみを素性に用いた場合（ベースライン）の精度は 0.7578 で，全素性を用いた提案手法の性能の方が悪くなってしまった．この現象を調べたところ，Bag of words 素性が性能低下の原因となっていることが判り，これを除いた提案手法（7 素性）の精度は 0.8125，マクロ F1 スコアは 0.5606 であった．Bag of words 素性を用いた時に性能が低下するのは，誤情報のトピックと関連が深い単語を分類器が丸暗記してしまうためだと考えられる．

表 3: 分類器のモデル（6 素性 -SB）

重み	素性	ラベル
3.88	誤情報とツイートの類似度 (SU)	誤情報
1.29	訂正表現あり (T=True)	訂正情報
0.83	訂正表現周辺の単語 (TW[0]=デマ)	訂正情報
0.78	訂正表現なし (T=False)	誤情報
0.74	訂正表現周辺の単語 (TW[2]=ない)	その他
0.73	訂正表現周辺の単語 (TW[4]=周辺)	その他
0.72	URL あり (U=True)	訂正情報
0.68	拡散あり (R=True)	誤情報
-0.68	URL あり (U=True)	誤情報
-0.87	訂正表現なし (T=False)	訂正情報
-0.92	拡散なし (R=False)	誤情報
-1.03	訂正表現あり (T=True)	誤情報
-1.04	訂正表現周辺の単語 (TW[1]=じゃ)	訂正情報
-1.20	訂正表現周辺の単語 (TW[2]=ない)	訂正情報
-1.73	誤情報とツイートの類似度 (SU)	その他
-2.16	誤情報とツイートの類似度 (SU)	訂正情報

さらに，7 素性の設定から，残りの素性を削除している場合の結果を 6 素性として表 2 に載せた．削除することにより性能が低下した素性は，分類に貢献したことになり，逆に上昇した素性は過学習を引き起こすなどして，分類の邪魔になっていったと考えられる．特に貢献していた素性は「URL の有無 (U)」で，貢献していなかった素性は「単語バイグラムを用いた，誤情報とツイートの類似度 (SB)」であった．

表 3 に，学習により高い重みが与えられたトップ 8 の素性と，低い重みが与えられたトップ 8 の素性を示した．素性セットとしては，表 2 の評価で最も高い精度を示した 6 素性 (-SB) を用いた．最も高い重みが与えられた素性は．誤情報とツイートの類似度 (SU) から誤情報を予測する素性で，重みが 3.88 であった．得られたモデルからは，直観的に理解できるような重みが与えられていることが判る．例えば，「○○はデマじゃない」といった，ツイートに対しては，訂正表現周辺の単語 (TW) の素性によって，訂正情報ではないと判断されやすくなる．また，URL が本文中に存在すると，訂正情報であり，誤情報ではないと判断されやすくなる．

本研究では，実験設定でも述べたように，「コスモ石油〜」のような，ある誤情報で機械学習を行い，「イソジン〜」などの別の誤情報の分類の性能を測定した．誤情報の種類をさらに増やすことで，分類器の性能が向上するかどうか見積もるために，学習曲線を求めた．具体的には，14 種類の誤情報から，ランダムに 1 個をテストデータとして選び，残りの 13 個を訓練データとした．次に 13 個の誤情報から成る訓練データから，1 個ずつ誤情報をランダムに選ぶことで学習データの量を調整し，学習曲線をプロットした．ランダムにデータを選ぶという以上の試行を，学習曲線の形が安定する

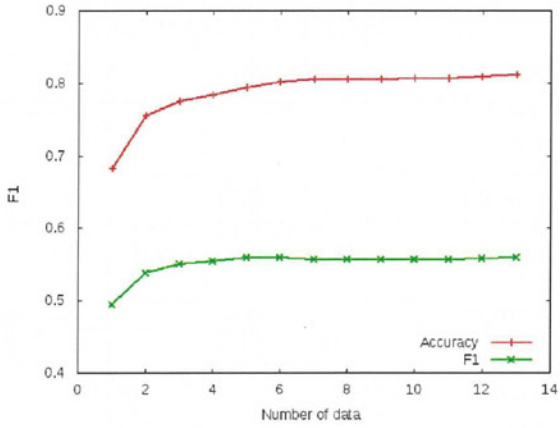

図 2: 学習曲線

まで（280 回）繰り返した．以上の方法で，表 2 で最も高い精度を示した 6 素性 (-SB) を用いて，学習曲線をプロットした．

図 2 によると，精度の上昇傾向は緩やかに続いており，誤情報の種類を増やすことによって，精度のさらなる向上が期待できる．しかし，劇的な精度向上を達成するには，提案手法の問題点を明らかにする必要がある．

3.3 考察

実験では，分類器の性能について様々な分析を行い，ベースライン手法である訂正表現の有無による分類よりも，精度の高い分類ができることを確認した．中でも「URL の有無」の素性は有効に働き，下のように訂正表現では分類しにくいツイートを正しく分類できた．

うがい薬「飲まないで」と専門家　買い求め客が急増 http://...

また，訂正表現周辺の単語を素性にすることで，「デマじゃない」のような訂正表現を否定するツイートを正しく認識できることを期待していた．

万が一原発から放射能が漏れ出した際，被爆しない為にイソジンを１５ cc飲んでおいて下さい！原液です！ガセネタではありません．お医者さんからの情報です．これは RT ではないので信じてください！

しかし，コーパス内でこのような表現を用いたツイートが少ないため，学習がうまく行えなかった．但し，図 2 の学習曲線から判るように，訓練データの規模が大きくなると精度の向上が見られるため，学習データの量を増やすことで，有効な素性になると期待できる．

さらに，何の手がかりもないが，誤情報を訂正するツイートも存在する．

厚生労働省です不特定多数の方に送信されている，コスモ石油千葉製油所における火災関連のメールについては，厚生労働省からの発表情報ではありませんのでご留意願います

このツイートでは，「デマ」「嘘」などの訂正表現や，URL や RT は一切使われておらず，また誤情報の内容（「コスモ石油の火災により有害物質の雨が降る」）も説明していないが，内容から誤情報を訂正するツイートであると判断できる．このようなツイートを訂正ツイートと認識するためには，深い処理（例えば「火災関連のメール」を「火災により有害物質の雨が降るというチェーンメール」と解釈する）や，ツイートやユーザ間の関係（例えば，厚生労働省はこの誤情報に関連して別のツイートで訂正表現を用いて打ち消しを行った，等の手がかり）を用いる必要がある．

4 おわりに

本研究ではキーワードで収集されたツイートを誤情報の支持・拡散ツイート，誤情報の反論・訂正ツイートに分類するためのコーパスを構築した．そのコーパスを用い，教師あり学習を用いて自動分類手法を提案した．その結果，訂正表現だけを用いた分類よりも，良い性能を示す分類器を作成できた．

今後の課題は分類器のより細かな分析を通じて，さらなる分類精度を計ることである．また，別の災害や平常時など様々な環境下での検証も必要である．他にもシステムのリアルタイム化するにあたり，誤情報から適切な検索クエリの自動生成や，「誤情報」や「訂正情報」以外の「懐疑情報」や「検証情報」など，より細かい情報の分類に取り組む予定である．

謝辞

本研究は，文部科学省科研費（23700159），および JST 戦略的創造研究推進事業さきがけ，および総務省・情報通信ネットワークの耐災害性強化のための研究開発事業の一環として行われた．貴重なデータを提供して頂いた Twitter Japan 株式会社に感謝いたします．

参考文献

[1] 萩上チキ. 検証 東日本大震災の流言・デマ. 光文社, 2011.

[2] 情報支援プロボノ・プラットフォーム（iSPP）. 3.11 被災地の証言ー東日本大震災　情報行動調査で検証するデジタル大国・日本の盲点ー. インプレスジャパン, 2012.

[3] 鍋島啓太, 水野淳太, 岡崎直観, 乾健太郎. マイクロブログからの誤情報の発見と集約. 言語処理学会 第 19 回全国大会 発表論文集, 2013.

[4] 渡邉研斗, 鍋島啓太, 水野淳太, 岡崎直観, 乾健太郎. Twitter における誤情報の拡散収束過程の可視化. 情報処理学会 第 75 回全国大会 発表論文集, 2013.

2012年度　卒業論文

ツイートの自動分類による
誤情報の拡散収束過程の分析

2013年3月18日

情報知能システム総合学科
(学籍番号: A9TB2251)

渡邉 研斗

東北大学工学部

概 要

自然災害時では，被災地の状況確認や信憑性の判断などに時間がかかり，ニュースや自治体からの情報の伝達が遅くなるため，情報拡散に優れている Twitter は非常に役立つツールとして注目されている．しかし，事実と反する情報 (誤情報) が出現し，社会に混乱をもたらす原因になる可能性がある．誤情報が出現した場合でも，それを訂正する情報 (訂正情報) の出現により誤情報の拡散が収まることもあり，ユーザは注意して情報を収集する必要がある．誤情報が拡散し続ける事態を防ぐために誤情報を自動的に提示する様々な研究がなされた．しかし，いずれも訂正情報に含まれる「デマ」などの表現（訂正表現）を元に誤情報の自動判定をするので，誤情報の提示が後手に回り，迅速な対応が必要なときの実用性に疑問が残る．

本稿では誤情報が拡散してから収束するまでの過程（誤情報の拡散収束過程）を分析することによって，誤情報の提示の実用性を調査し，加えて誤情報の拡散を促進させている要因の解明に迫る．更に，誤情報の拡散収束過程の提示を自動化するために，機械学習を用いて誤情報と訂正情報を自動的に分類する手法を提案する．拡散収束過程の分析の際に作成したデータを教師データとし，様々な素性をもとに最大エントロピー法を用いた機械学習を行なう．

実験によって，訂正表現を用いた手法との比較を行い，自動分類に有効な素性を示す．さらに結果から得られた失敗例から，自動分類の今後の問題点についても述べる．

目 次

第1章　序論

近年，災害時においての情報交換の手段として**SNS(ソーシャル・ネットワーキング・サービス)**が注目されている．SNSとは社会ネットワークをインターネット上で構築するサービスの事であり，誰でも即座に情報を収集・発信でき，規模の大きな情報交換ツールとして有用である．大規模な自然災害時には，被災地の状況把握や信憑性の判断に時間がかかるため，ニュースや自治体から最新の情報収集することが困難になる可能性がある．このような状況では，情報は人々の行動の判断材料になり，情報発信・収集ができるSNSは非常に重要な役割を果たす．

2011年3月11日に発生した東日本大震災においても**Twitter(ツイッター)**と呼ばれるSNSで安否確認・救助要請・危険警告などの情報収集・発信をする人々が多かった．Twitterとは140字以内の**ツイート**と呼ばれる短い文を投稿できるSNSであり，世界で1.4億人のユーザが利用している(2012年4月時点)．Twitterは140字という制限のある文字情報なので，細い通信路でも情報交換がしやすい．また，他のユーザのツイートを自動的に取得する**フォロー**という機能があり，より多くのフォローをされているユーザの発言は非常に強い影響力を持つ．また，Twitter独特の機能として**リツイート**と言うものがある．これはフォローしているユーザのツイートを有用な情報だと感じ，他のユーザにも教えたい時に，そのツイートをそのまま投稿する機能である．このリツイートは情報の拡散が非常にしやすく，ユーザを伝って情報を広範囲に発信できる．実際に東日本大震災では，Twitterで救助要請をし，知事がツイートを見つけたことにより救助された例もある．しかし，同時に「コスモ石油の爆発で有害な雨が降る」等の事実とは反する情報がTwitter上で拡散し，社会に混乱をもたらした[9, 10]．これら事実とは反する情報である**誤情報**を人々が信じてしまうと，誤った判断・行動をしてしまう可能性がある．一方,「有害な雨が降るというツイートはデマです」のような，誤情報を訂正する立場にある情報も多数見られた．これら誤情報を訂正する立場にある**訂正情報**を人々が得ることで，誤情報の拡散が収まった．この誤情報が発生・拡散し，訂正情報により誤情報が収束する，もしくは収束するまでの過程を**誤情報の拡散収束過程**と呼ぶ．

東日本大震災に拡散した誤情報の拡散収束過程の例で「被災した入試合格者が入学手続きできないため，東大が合格を取り消した」というものが存在した．この誤情報に関して，東大は即座に公式HPで「合格者本人の意志を確認せずに入学の取消すようなことはない」という訂正情報を発表したため，2時間程度で誤情報が収束した．一方で「支援物資の空中投下は法律で認められていない」という誤情報も拡散され，こちらも訂正情報が即座に出現したが，そのまま長時間に渡り拡散し続けた．

このような社会混乱を引き起こしかねない誤情報の拡散を防ぐ方法が必要になってくる．宮部ら[1, 2]は誤情報を発信したユーザとフォローユーザとの関係の分析や，平常時と震災時の誤情報の拡散収束過程を比較し，自動的に収集した誤情報をユーザに提示するシステムを提案した．梅島ら[3, 4]は誤情報と訂正情報の傾向をツイート文の特徴から数種類に分類し，拡散防止に迫った．白井ら[6]は，感染症疾患モデルを拡張した誤情報拡散収束モデルを構築し，誤情報を発言したユー

ザへの提言を行うことで，誤情報を収束させる方法を提案した．また，鳥海ら [7] は誤情報を判定する手法を提案し，誤情報らしさを確かめるシステムを構築した．これらの研究の共通点に

- 誤情報がある程度拡散して，訂正情報が出現しているという仮定を持つ
- 大量の誤情報の中からユーザは社会混乱に陥りそうな誤情報を探す必要があるという仮定を持つ

という点が挙げられる．1 項目については技術的に誤情報の検出は難しく，訂正情報の検出のほうが比較的に容易という背景がある．つまり関連研究の手法は，誤情報を検出した時には誤情報がある程度拡散している可能性があり，誤情報の拡散を防ぐ手法としては実用性に疑問が残る．よって，訂正情報を検出した後に，誤情報を提示する実用性の検討をしなければならない．2 項目に関しては，関連研究の提示した方法では，大量にある誤情報の中から「イソジンを飲むと放射線予防になる」のような，社会混乱に陥りやすいものをユーザが自ら探す必要がある．よって，どの誤情報がより社会混乱を引き起こしそうかを自動的に判断し，ユーザに提示する必要がある．

関連研究の問題点を踏まえ，

1. **自然災害時において，訂正情報が出現した後に誤情報を提示しても実用性があるかどうかを調査する**
2. **社会混乱になりそうな拡散収束過程の要因について分析する**

ことを本研究の目的とする．本研究では誤情報の拡散収束過程の分析を行い，分析の結果，訂正情報の検出による実用性があることを確認した．また，社会混乱を起こしそうな誤情報について，ツイートの量や時間に着目し新たな知見を得ることができた．

最終的には，誤情報の拡散収束過程を政府や自治体に提示・分析させることで，誤情報の重要性を与え社会混乱を防ぐ支援をするシステムを構築したい．このシステムを実現するには，Twitter 上で拡散している誤情報に関連するツイートを収集し，各ツイートが誤情報なのか訂正情報なのかを自動的に判別・整理する必要がある．本研究では，鍋島ら [5] の手法を用いて，震災時に拡散した誤情報を説明する記述 (例えば「コスモ石油の爆発で有害な雨が降る」や「イソジンは放射線予防に効く」) の検出ができると仮定する．検出できた内容について言及しているツイートは，誤情報の拡散もしくは訂正を行っている可能性が高い．そこで，本研究ではさらに，以下を目的とする．

3. **キーワードで収集されたツイートを誤情報の支持・拡散ツイート，誤情報の反論・訂正ツイートに分類するためのコーパスを整備**
4. **そのコーパスを用い，教師あり学習を用いた自動分類手法を提案する**

評価実験では提案手法の性能を確かめ，訂正表現のみを用いた手法と比べて精度が向上することを示す．

本論文では 6 章で構成されている．本章に続く 2 章では自然災害時における誤情報の拡散を防ぐための関連研究について述べる．3 章では誤情報の拡散収束過程を提示する上で必要になるデータや，技術的問題点，実用面での問題点について述べる．4 章では東日本大震災時の誤情報の拡散収束過程について分析し，訂正情報の量や，収束までの時間などに着目し，本手法の実用性につ

いて述べる．5章では誤情報と訂正情報に自動分類する手法を提案し，実験設定と結果，及び考察について述べる．6章では本研究で明らかになった点，及び今後の課題を述べ，本研究のまとめを行う．

第2章　関連研究

この章では，自然災害時に誤情報の拡散を防ぐための関連研究について述べる．まず，誤情報の拡散について分析を行った関連研究を挙げ，それらを踏まえた上で誤情報の拡散を防ぐ方法と，そのために必要な技術的問題について取り組んだ研究について述べる．

2.1　誤情報の特徴分析

誤情報の拡散を防ぐためには，誤情報の拡散収束過程についての分析をすることで，何らかの特徴を捉える必要がある．宮部ら [1] は災害時と非災害時の誤情報について網羅的に分析して，様々な知見を得ている．まず宮部らは誤情報の原因となったツイートが，どのように他のユーザに拡散していくかを辿り，口伝えでの誤情報拡散よりも Twitter での誤情報拡散の方が，誤情報の内容の変容が起こりにくい可能性があることを明らかにした．また，誤情報を拡散したユーザのツイートを観察することで，誤情報を発信したユーザが後に自分の発信した情報に対して訂正を行わない傾向があると明らかにした．拡散収束過程を観察し，訂正情報が出現しても直ちに誤情報が収束することはないという知見も得ている．

梅島ら [4] は災害時の Twitter の誤情報と訂正情報のリツイートの傾向について仮定を立て，それらについての調査を行った．まず，梅島らは多くの人にリツイートされるツイートを「実用的な特徴か」「私情を含む情報か」「実話か」「経験情報を含むか」の 4 つの特徴をもとにして，「情報発信計」,「経験談」,「私見」,「小話」,「ジョーク」の 5 つに分類できると仮定した．特に梅島らは，5 つに分類されたリツイートの内，災害時において実用的かつ，私情を含まない「情報発信系」のリツイートについて，細かい分析をしており以下の 4 つの特徴が有ると明らかにした．

- **拡散**:他のユーザに RT を希望する旨を記載することで，ツイートの拡散を助長するもの．ここでは,「拡散希望」「RT してください」等と記載されていることが多い．
- **URL**：そのツイートに関連する情報が得られる web ページの URL を記載しているもの．
- **ハッシュタグ**：そのツイートに関連するハッシュタグを記載しているもの．ハッシュタグとは，ツイートでキーワードや話題を明示的に表現するために用いられる#で始まる文字列のことである．#に続く文字列は半角英数字であり，ハッシュタグの前後にはスペースを必要とする．
- **詳細情報**：そのツイートに関する詳細情報を記載しているもの．詳細情報とは，情報の詳細を知るために必要な情報のこと．住所や電話番号等．

このような分析することで梅島らは URL を含むリツイートは情報源がはっきりしているため誤情報である可能性が低いという仮定を立て，URL を含むツイートの観察を行った．その結果，訂正

情報のほうがURLを多く含むことがわかり，誤情報にURLが含まれている場合は，ほとんどは確証の無い，個人のページであるということを明らかにした．また，誤情報と非誤情報を比較することで，誤情報には「行動を促す」「ネガティブな」「不安を煽る」内容が多いという仮定についての検証をし，この事実を明らかにした．同時に「行動を促す」「ネガティブな」「不安を煽る」内容のツイートはリツイートされやすいということも明らかにした．

以上の関連研究では，誤情報を拡散しているユーザには悪意はなく，誤情報と知らずに善意で拡散している場合がほとんどであるとことも明らかにしている．よって，誤情報の拡散を防ぐためには，誤情報を発信したユーザや拡散しようとしているユーザに対して，訂正情報を提示する手法が有効であるとも述べている．しかし，自然災害という危機的状況の中，善意で情報を拡散しようとしているユーザは危機意識が強く正義感にかられている場合が多く，頑なに拡散しつづけるので，訂正情報を提示しても拡散防止の効果はあまり得られないと考えられる．そこで本研究では，誤情報の拡散収束過程から誤情報を収束させた要因について調査し，情報を拡散しているユーザに訂正情報を提示する以外の手法について考える．

2.2 誤情報の拡散を防ぐための手法

誤情報の拡散を防ぐためのシステムに関する研究も存在する．宮部ら [2] は事前研究 [1] を踏まえ，情報が誤っていることをユーザに提示できれば，誤った情報の拡散を防ぐことができる可能性があると考え，自動的に誤情報を収集・提供する仕組みとして，流言情報クラウドを提案した．流言情報クラウドでは，あらかじめ訂正情報からリアルタイムに誤情報を蓄積し，人手を介すことなく情報提供可能であるが，人手による精査も可能とすることにより，提供する情報の信頼性を向上できるようにしてある．また，蓄積した情報をユーザに提供することにより，誤情報の拡散を防ぐシステムである．

鳥海ら [7] らは Twitter 上に投稿された情報が誤情報の可能性が高いかどうか，他のユーザによって誤情報であると判断され訂正されているか過去のツイートを確認することによって判断するシステムを提案・構築した．この誤情報判断システムでは，判断対象となる文が入力されると文中に含まれるキーワードを抽出し，同一キーワードを含むツイートを Twitter から取得する．このツイートの中に「○○はデマです」と誤情報に対して訂正していると判断できる単語を含むかどうかを確認し，その結果を元に判定対象にスコアを付与する．スコアに応じて「誤情報の可能性が非常に高い」,「可能性が高い」,「可能性がある」,「誤情報ではないかもしれない」,「誤情報ではない」の 5 段階評価を与える．鳥海らは誤情報判断スコアを以下のように定義した．

$$Score = \frac{l(n\text{のうち，誤情報に対して訂正しているツイート数})}{n(\text{誤情報について言及しているツイート数})} \tag{2.1}$$

この手法はシンプルであるが誤情報判断の精度が 8 割を超え，十分に実用に耐えうる結果が得られている．

しかしこれらの手法は，訂正情報の検出をしないと運用できない仕組みになっている．本研究では訂正情報を検出した後に誤情報を提示しても実用性が得られるのかを調査し，誤情報の出現から訂正情報の出現までの時間を計測することで，実用性が得られるか検証する．また，関連研究の手法はユーザが興味あるものに対してのみに有効な方法である．しかし，自然災害時では大量に有る誤情報の中から，拡散したらより危険なものを提示することが重要になる．例えば「漫