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

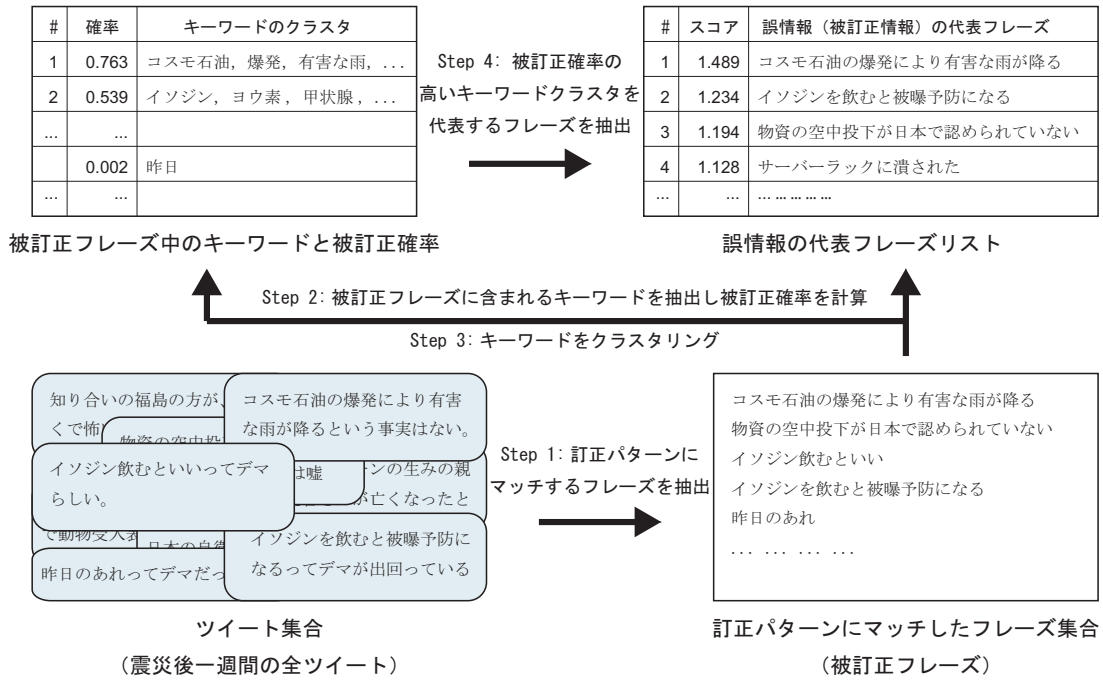

図 1: 誤情報抽出の流れ

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) キーワードの抽出による誤り

「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事例である．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，被訂正フレーズ集合には含まれるが，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

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

(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．

(h) ランキング外

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.