4Q-9

訂正パターンに基づく誤情報の抽出と集約

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では，ソーシャルメディアは有益な情報源として活躍した．その一方で，「イソジンを飲むと放射線予防になる」に代表されるような誤情報の拡散が問題となった．このような誤情報の中には人々の安全に関わるもの多く，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．

誤情報の検出に関する研究は多くされている [1, 2, 3, 4]．Qazvinian ら [1] は，誤情報に関連するツイート群から，誤情報に言及しているツイートと，誤情報に言及していないツイートに分類し，さらに誤情報に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．梅島ら [2] は，訂正を明示する表現（「デマ」など）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．

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，東日本大震災後1週間の全ツイートから誤情報を自動的に収集する手法を提案する．提案手法により，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，さらに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 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」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短距離法を用いた．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．

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

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．$\text{hist}(l,t)$ 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，$\text{PMI}(t,w)$ 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，$\text{hist}(\text{len}_s, t)$ 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．

3 実験

評価実験では，東日本大震災時のツイートデータを用いて誤情報の抽出を行い，その精度と再現率を測定した．

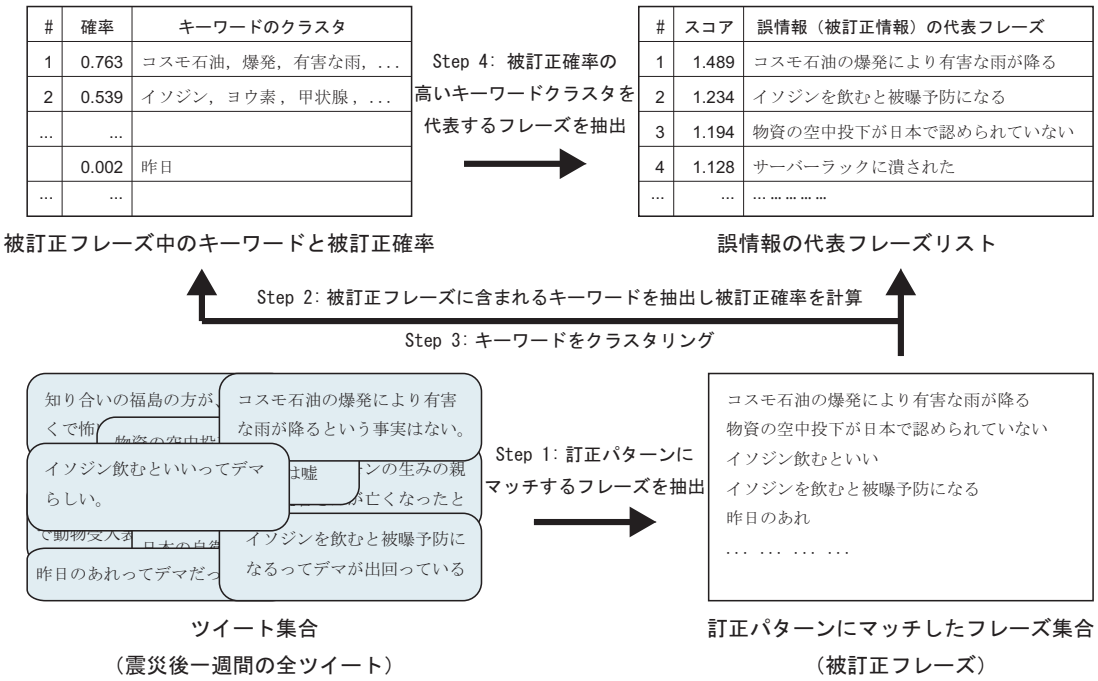

図 1: 提案手法の流れ

3.1 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[1] 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以下の 4 つのウェブサイト[2]に掲載されている事例のうち，Twitter データの投稿期間内に発信されたと判断できる 60 件の誤情報を用いた．提案手法で抽出された誤情報の正否は，同等の内容が 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．

3.2 実験結果

評価結果を表 1 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．

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

[2] 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: 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)

4 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] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proc. of EMNLP 2011*, pp. 1589–1599, 2011.

[2] 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.

[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, 2012.