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．

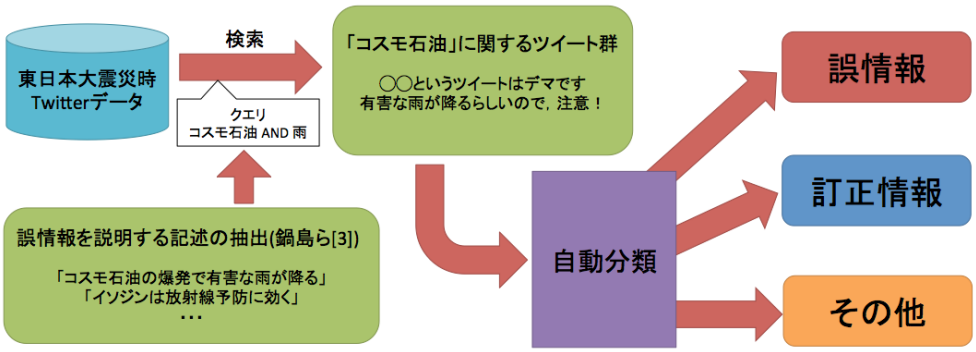

図 1: 誤情報モニタリングシステムの概要

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

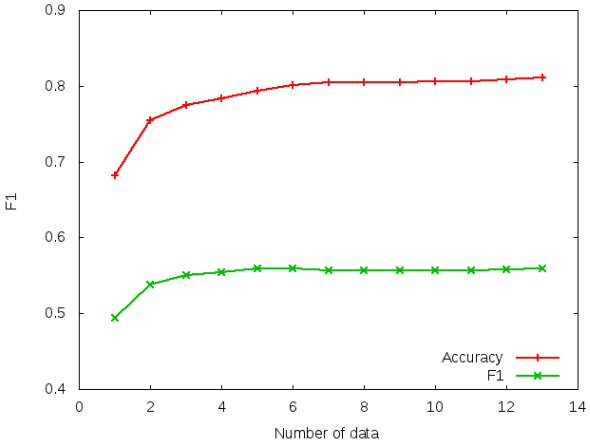

図 2: 学習曲線

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！

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.