Twitterからの多様な流言訂正情報獲得のための分類器の構築

高橋 弘志　　小町 守　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{hiroshi-t, komachi, matsu}@is.naist.jp

1　はじめに

2011年3月11日，日本の東北地方を中心に東日本大震災が発生した．震災の発生直後から，Twitter[1]をはじめとしたインターネットメディアはそれぞれの特徴を生かして情報伝達の手段として広く利用された．特に震災直後の携帯電話による通話や電子メールが使用しづらい状況において，Twitterは地域の生活情報や安否情報をリアルタイムに入手するために利用されており，災害時における情報共有手段としての可能性が示唆されている．

しかし有用な情報が共有されたのと同時に，Twitter上では震災に関連する裏付けをもたない情報，すなわち流言も大量に発生し，拡散した．

例えば，表1に挙げた流言はいずれも偽りの情報である．しかし知識のないユーザが一見しただけでは偽りであることを疑うことすらできない可能性が高い．そしてそのようなユーザは善意によって流言を拡散するだろう．そのため，下のような情報の不確かさについて言及したツイートを提供し，無知なユーザに流言の真偽を疑うきっかけを与える事ができればTwitter上での流言の拡散を防ぐことができると考えられる．

- サーバルームで死にそう発言はデマですよ。みなさんソースないのはむやみにRTしないように。

このような，流言の不確かさについて言及しているツイートを**流言訂正情報**と定義し，webからリアルタイムに獲得する研究が宮部ら[1]によって行われている．宮部らは，流言訂正情報に含まれる傾向が強い文字列を**流言マーカー**とし，流言マーカーを含むツイートを流言訂正情報か否かに分類する分類器を構築している．しかしこの手法では，設定する流言マーカーによって獲得される流言訂正情報のトピックに偏りが生じる事が考えられる．そのため本研究では，流言マーカーの見直しやツイート本文の情報をより多く利用することによってこの問題を改善し，より多様な流言訂正情報を獲得することを目指す．

表1: Twitter上で発生した流言の例

コスモ石油爆発で有害物質の雨が降る
関西電力が関東送電のために節電呼びかけ
ドワンゴ社員，サーバルームから救助要請
放射線対策にイソジンを飲むと良い

2　流言訂正情報の獲得手法

2.1　流言訂正情報獲得手法の概要

本研究では流言訂正情報分類器を構築し，これを用いて訂正情報を獲得する．分類器を構築する手順として，まず東日本大震災発生直後の日本語のツイートの中から，流言マーカー等の条件に従ったツイートを抽出する．次にこれらのツイートに対して流言訂正情報か否かのタグを人手で付け，コーパスを構築する．続いて，このコーパスを用いてツイートが流言訂正情報か否かを判別する訂正情報分類器を構築する．

2.2　流言マーカーと対象ツイートの制限

宮部らは「地震」を含む東日本大震災直後のツイートを収集し，流言マーカーを「デマ」として分類器を構築している（すなわち「地震」と「デマ」の両単語を含むツイートのみを扱った）．この手法は直感的に正しいが，ある潜在的な問題を抱えている．それは，「地震」と「デマ」を含むという制限により，獲得できる訂正情報のトピックが偏りうるという問題である．そこで本研究では，宮部らの手法からツイート収集の際の制限を取り払い，さらに流言マーカーとして「デマ」に「ガセ」「嘘」の2語を加えた3語を用いる．

[1] https://twitter.com

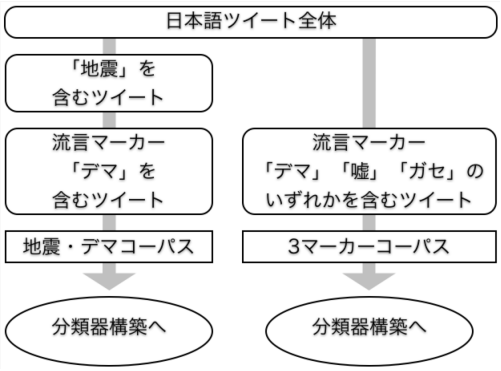

図 1: コーパス構築までの流れ

2.3 流言訂正情報タグ付きコーパスの構築

本研究では 2 種類の流言訂正情報タグ付きコーパスを構築する．一つは宮部らの手法を再現した「地震・デマコーパス」であり，もう一つが「デマ」「嘘」「ガセ」を流言マーカーとした「3 マーカーコーパス」である．これらコーパスの構築方法について述べる．まず，一定期間内に投稿された全日本語ツイートの中から，ツイート収集の際の制限や設定した流言マーカーに従い，それぞれのコーパスで扱うツイートを抽出する．続いてこれらツイートの一部に対して流言訂正情報か否かのタグを人手で付け，コーパスとする．図 1 にこの一連の流れを表す．また以下に，ツイートを流言訂正情報と判断する基準と「デマ」を流言マーカーとした場合にその基準に合致するツイートの例を示す．

ある情報の不正確さの記述が主題である

- 【デマ注意】日本では食料の空中投下が認められていない、という情報はデマです。

ある情報の不正確さの記述が含まれ，主題ではない

- あらゆる情報が錯綜してるなぁ。緊急なものは勿論、尾田先生の寄付もデマらしいし。

流言に関する情報をまとめたサイトを紹介している

- 太平洋沖地震、ネット上でのデマまとめ http://xxx.xxx

2.4 流言訂正情報分類器

2.4.1 ツイート本文素性

ツイートの情報の信頼性に関する研究は盛んに行われており，Castillo ら [2] はある話題に関するツイートの集合を信頼できるか否かに分類する分類器を決定木学習を用いて構築している．また Qazvinian ら [3] は噂を検出するベイズ分類器を構築している．どちらの研究でも素性として単語や品詞といった情報が利用されており，本研究ではこのようなツイート本文から得られる素性をツイート本文素性として取り扱う．ツイート本文素性として，表 2 に示したものを用いた．

表 2: ツイート本文素性

流言マーカーと周辺文脈[1]（window size ±1）		
人名の割合	組織名の割合	地名の割合
固有名詞の割合	疑問符の有無	感嘆符の有無

[1] window size を 1 から 5 まで変化させながら予備実験を行ったところ，これを 1 としたときに最も良い結果が得られた．また同一ツイート内に複数のマーカーが存在する場合，最初に現れたマーカーの周辺文脈のみを利用する．

表 3: Twitter 特有の素性

ハッシュタグの有無	リツイートの有無
引用による訂正の有無	URL の有無

2.4.2 Twitter 特有の素性

Twitter にはリツイートや URL，ハッシュタグなどといった，他のテキストには見られない様々な表現が存在する．これらの表現が持つ情報も積極的に利用されており，Qazvinian らは URL を，Castillo らはそれに加えてリツイートの深さをそれぞれ先述の研究に用いている．本研究では Twitter 特有の素性として表 3 に示したものを用いた．

2.4.3 分類器の学習アルゴリズム

本研究では，分類器の学習に SVM（Support vector machine）を用いる．今回は libsvm 3.14[2]を使用し（RBF カーネル）パラメータは付属の grid.py によって決定した．

3 訂正情報分類器の性能評価実験

3.1 実験の目的

実験では，構築した訂正情報分類器の分類性能を評価する．また，流言マーカーが獲得される訂正情報のトピックの多様性に与える影響を確認する．

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

3.2 使用するデータ

使用するデータは，東日本大震災ビッグデータワークショップ Project 311[3]において Twitter Japan 株式会社より提供された約 1 億 8000 万件のツイートである．これは，2011 年 3 月 11 日から一週間で投稿された全日本語のツイートに等しい．またこれらのツイートに対して，ツイートの投稿者のユーザ情報を Twitter API を用いて取得する．ただし，ツイートが投稿された 2011 年 3 月時点のユーザ情報を取得することはできないため，代替として 2012 年 11 月時点でのユーザ情報を取得，使用する．

3.3 前処理

使用するツイートに対して以下の前処理を施す．

改行文字の除去

リプライの除去 [4]

ユーザ情報の無いツイートの除去

bot アカウントの除去

重複するリツイートの除去

あるツイートが複数回リツイートされた場合，無作為にリツイートを一つ選択し，残りのリツイートを除去する．これは実験データ内に同じ文面のツイートが多く現れることを避けるためである．

形態素解析

以下の手順でツイートの形態素解析を行う．

1. ツイートから URL，ハッシュタグ，メンションを示す文字列を取り除く．
2. ツイートがリツイート，引用の場合，コメント部分とリツイート元部分に分割する．
3. 分割した文字列それぞれに対し MeCab 0.994（IPA 辞書 2.7.0）[5]を用いて形態素解析を行う．

以上の前処理の結果，地震・デマコーパスに対応するツイートを約 11,000 件，3 マーカーコーパスに対応するツイートを約 227,000 件得た．この中からそれぞれ無作為に 1,100 件のツイートを選択し，タグの付与を行いコーパスとしている．

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

[4] どのような情報の不確実さについて述べているのかがリプライ単体の訂正情報からは判断しづらいため，除去することとした．

[5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

3.4 ベースライン

ベースラインとしては，先行研究 [1] で提案された訂正情報分類器を基本にした分類器を用いる．以下に構築したベースラインの分類器の詳細を示す．

前処理

先に述べた前処理を施す．

流言マーカー

「デマ」を流言マーカーとする．

コーパス

地震・デマコーパスを使用する．

訂正情報分類器

素性には流言マーカーの周辺文脈を用いる．また学習アルゴリズムは提案法と同じものを用いる．

3.5 評価指標

分類器の分類性能の評価基準として，本研究では再現率と F 値を用いる．これは，流言訂正情報はある情報の不確実さについて述べたもので，ある情報の真偽を正確に断定できるものではないからである．

4 結果と考察

4.1 流言訂正情報分類器の分類性能評価

構築した分類器に対し 10 分割交差検定を行い性能を評価する．使用したコーパスと分類器の性能を示したものが表 4 である．まず地震・デマコーパスを用いてベースラインと提案法の分類性能を比較する．提案法は使用する素性を，全ての素性，全体からツイート本文素性を引いたもの，全体から Twitter 特有の素性を引いたもの，と変化させながら実験を行った．これからわかるように，ツイート本文素性も Twitter 特有の素性も分類性能の向上に貢献していた．より貢献が大きかったのは Twitter 素性である．これは，地震・デマコーパス中に大量に見られた外部ページへの URL を含む訂正情報の影響であると考えられる．最も多く獲得された訂正情報は，以下に例を挙げるような「地震デマのまとめサイトを紹介した」ツイートであり，地震・デマコーパス内の 530 件の訂正情報の内 239 件がこのようなツイートだった．

- 荻上式 BLOG「東北地方太平洋沖地震，ネット上でのデマまとめ」http://xxx.xxx

表 4: 使用コーパスと分類器の分類性能

	地震・デマ		3 マーカー	
	再現率	F 値	再現率	F 値
ベースライン	57.5	66.9	-	-
提案法（全素性）	79.4	80.7	65.9	69.6
提案法（ – 本文素性）	74.8	78.0	56.3	63.3
提案法（ –Twitter 素性）	73.7	76.2	62.7	65.6

続いて，3 マーカーコーパスを用いた分類器の分類性能を評価する．ここでも先程と同じく使用する素性を変化させながら実験を行った．表 4 から，地震・デマコーパスを使用した場合に比べて再現率，F 値共に悪化していることがわかる．実際に訂正情報を分類し誤った例を確認したところ，誤って分類されたツイートには以下の様な特徴がみられた．

訂正する対象の情報が曖昧（ False Positive ）

- ああ，デマですか … こういうのあるからやだわ … ＞ＲＴ

URL 等を持つ日常会話（ False Positive ）

- 読んでたら涙出そうだった…いや嘘です。出てました。http://xxx.xxx

URL 等のない訂正情報（ False Negative ）

- 築地の件はデマとの情報あり

これらの誤りを改善する方法としてはツイートの係り受け情報の利用や，固有名抽出と Twitter 上での単語のバースト検出とを組み合わせた新しい流言マーカーの設定などが考えられる．また，流言マーカーとして使用した「デマ」「嘘」「ガセ」の語を比較すると，圧倒的に「嘘」を含むツイートの誤り例が多かった．これは,「デマ」や「ガセ」と比較して,「嘘」は日常の様々な場面で使用される語であり，流言訂正情報に用いられる頻度が低いからだと考えられる．

4.2 流言マーカーと訂正情報のトピック

流言マーカーの変化が獲得できる訂正情報のトピックにどのような影響を与えるかを調査する．有志によって東日本大震災後に Twitter で発生した流言をまとめた web ページ[6]を参考に，それらを人手でトピックごとに分類した．また地震・デマコーパスと 3 マーカーコーパスを用いて構築した分類器をそれぞれ地震・デマ分類器と 3 マーカー分類器とする．これらの分類器を用いて 3 マーカーコーパスの訂正情報の分類を行ったところ，いくつかの流言トピックは 3 マーカー分類器でしか獲得できなかった．その流言トピックの例を表 5 に示す．ここで，地震・デマ分類器が獲得できなかったトピックのツイートは，追加した流言マーカー「デマ」「嘘」を含むといった特徴を持っていた．このことから，流言マーカーの設定が，獲得できる訂正情報のトピックに影響を与えていたことがわかる．

[6] http://twitter-dema.etc64.com

表 5: 3 マーカー分類器でのみ獲得されたトピック

築地に魚が余っているから買いに来てください
辻元清美議員が戦車投入に反対と発言
埼玉県の水道水を飲むと危険

5 おわりに

本研究では，Twitter 上の流言の拡散を防ぐ効果が期待される流言訂正情報を獲得するタスクに取り組んだ．分類器の性能評価実験の結果から，URL やリツイート等の Twitter 特有の情報は，流言訂正情報を獲得する際にも有用であった．また，構築に用いたツイートが二重に制限された地震・デマ分類器は，3 マーカー分類器と比べて得られる訂正情報のトピックが少なかった．今後の応用としては，Twitter 上に平時に発生した流言訂正情報への適応などが考えられる．

謝辞

本研究に使用した日本語ツイートのデータは，東日本大震災ビッグデータワークショップ Project 311 を通して Twitter Japan 株式会社様より提供していただいたものです．

参考文献

[1] 宮部真衣，梅島彩奈，灘本明代，荒牧英治
流言訂正情報に基づいた流言情報クラウドの提案，第 4 回楽天研究開発シンポジウム, pp. 1-4, 2011.

[2] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In WWW 2011, pp. 675-684, 2011.

[3] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In EMNLP, pp. 1589-1599, 2011.