4Q-5

返信・非公式リツイートで表明される論述関係の認識

大和田 裕亮 † 岡崎 直観 ‡†† 乾 健太郎 ‡ 石塚 満 †
† 東京大学大学院情報理工学系研究科 ‡ 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

1 はじめに

Twitter は情報発信の手軽さやリアルタイム性が魅力のメディアであり，東日本大震災でも大いに活用されたが，誤情報の拡散が混乱や 2 次的な損害にも繋がった．情報の信憑性をコンピュータが自動的に判断するのは技術面および実用面において非常に難しいため，我々はツイート空間の情報を整理し，各ツイートの裏を取るような情報を提示することで，人間が情報の信憑性を判断することの支援を目指している．図 1 に「イソジンを飲めば甲状腺がんを防げる」という内容のツイート（中心）に対する，周囲の反応の例を示した．このようにツイートに対する同意・反論・疑問などの論述関係を整理することで，様々な角度から，分かりやすい形で情報を検証することができる．

本研究では，ツイートのうち返信や非公式リツイートの関係にあるツイート（返信・引用ツイート）によって表明される投稿者の「同意」「反論」「疑問」などの態度を推定する分類器を教師有り学習で構築し，その推定性能を報告する．

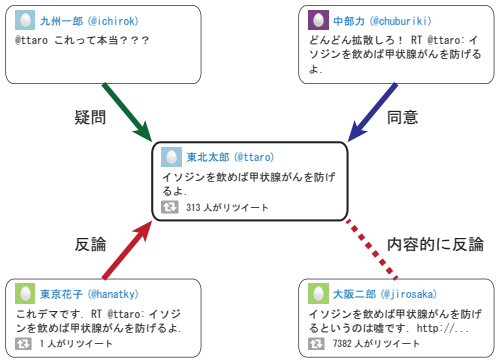

図 1: 返信・非公式リツイート，もしくは内容に基づくツイート間の論述関係．

2 手法

2.1 問題設定

本研究では，返信・引用ツイートを以下の 4 クラスに分類するタスクを考える．

同意 主張の支持や感情的な同調（感謝や崇拝も含む）など，返信・引用先のツイートに対して明確な同意の意図が感じられるもの．

反論 主張の否定や感情的な反発など，明確な反論の意図が感じられるもの．発言者に対し強く注意を促すようなものも含める．

疑問 返信・引用先に対して情報を要求しているが，明確な反論とは言えないもの．情報源を要求するようなものや，引用部に対する疑問の吐露なども含まれる．

その他 上記のどれにも分類できないもの．

「同意」「反論」による対立関係の抽出に加えて，「疑問」により人々の感じている問題点などを明らかにすることを目指している．なお，感情的な同調や反発のツイートの本文には有用な情報は少ないが，誰が誰に同調・反発したかというユーザー同士の関係は有用な情報であるため，これらも主張の支持・否定と同等に扱い分類を行なう．

2.2 返信・引用ツイートのアノテーション

本研究で行う実験で使用するデータは，ホットリンク社より提供された，2011 年の 3 月 9 日から 4 月 4 日までのツイートデータ（以下 hotto コーパス）である．hotto コーパスには，#tsunami や #jishin など震災に関連するハッシュタグまたはキーワードが含まれるツイートと，そのツイートを投稿したユーザーの情報などが収録されている．収集対象ユーザー数は約 100 万人，ツイート数は約 2 億 1 千万ツイートである．また，各ツイートには公式リツイート（以下 RT）や返信の関係が含まれており，これらはそのまま利用した．引用関係は含まれていなかったが，“`[RQ]T username:`” というパターンを含むツイートがあった場合，“`username`” の過去のツイートの中から近いものを探し引用元とした．

返信・引用ツイート収集のため，震災の際に出回ったデマ一覧[1] を参考に，「イソジン」「コスモ石油」など 20 個のクエリを選んだ．各クエリで hotto コーパスを検索し，RT が多い順に並べ，上位から順に返信・引用ツイートを（さらにそれらに対する返信・引用があれば続く限り）取得する．このようにして各クエリに対し返信・引用をそれぞれ 100 個ずつ重複を避けて取得し，計 4000 個の返信・引用ツイートを集めた．

4000 個のツイートに対し，三人のアノテーター（以下 A・B・C）の手により 4 クラスのラベル付けを行ったところ，3 人のラベルが一致したツイートは 2690 ツイートであった．ラベルの一致度を評価するため，ペアワイズに Cohen の一致係数を計算したところ，AB 間が 0.660，AC 間が 0.621，BC 間が 0.641 となり，十分な一致と見なせた．3 人のアノテーションが一致した 2690 ツイートを集めたものをデータセット X とする．データセット X には，同意が 1651 個，反論が 263 個，疑問が 282 個，その他が 494 個含まれる．

2.3 分類手法

返信・引用の態度を教師有り学習で分類する手法について述べる．学習に用いる素性は大まかに，ツイート内容・ツイート間・ユーザ間の 3 種類に分かれる．

ツイート内容の素性は，1-gram, 2-gram，URL 数，ハッシュタグ数，デマ表現，反論表現との一致度である．これらの素性では，返信ツイートであればツイート全文，引用ツイートであれば追加部分（ツイート全体から引用箇所以降を除いたもの）を対象とし，本文中のアカウント名や URL，ハッシュタグを抽出する．1-gram と 2-gram は，残りの部分から MeCab[2] を用いて抽出する．デマ表現は，「デマ」「ガセ」「誤報」「虚報」「削除」「訂正」の 6 つのうちどれかを含むという素性である．反論表現との一致度とは，あらかじめ作成した反論表現辞書を用い，反論に固有の表現をどの程度含んでいるかを表す．反論表現辞書は，データセット X の中から反論を表すと思われる表現（デマの否定とは限らない）を人手で抽出して作成する．反論表現には，「RT を取り消してください」などデマを否定する表現と，「あなたの発想が軽すぎる」などそれ以外の表現が含まれる．100 個の反論表現を含む辞書を作成し，反論表現との一致度を以下のように計算する．辞書中の各反論表現に対して，その 2-gram をツイート中にどの程度含むかを求め，それらの最大値を一致度とする．例えば，反論表現辞書が「それ/は/デマ/です」「これ/デマ/ね」の二つからなる場合を考える．「デマ/です」を含むツイートがあった場合，「それ/は/デマ/です」の 3 個の 2-gram のうち 1 個を含むので 1/3，「これ/デマ/ね」の 2 個の 2-gram は 1 個も含まないので 0 であり，この場合の一致度は 1/3 とする．

Statement Structure Recognition of Reply and Quoting Tweets on Twitter
Yusuke OWADA, Naoaki OKAZAKI, Kentaro INUI, Mitsuru ISHIZUKA

[1]「東北関東大震災に関するデマまとめ」のまとめ http://togetter.com/li/112556

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

表 1: 返信・引用ツイートの分類結果

	Precision	Recall	F 値	Accuracy
同意	0.855(1389/1624)	0.841(1389/1651)	0.848	
反論	0.488(120/246)	0.456(120/263)	0.472	
疑問	0.766(239/312)	0.848(239/282)	0.805	
その他	0.537(273/508)	0.553(273/494)	0.545	
全体				0.751(2021/2690)

返信先・引用元のツイートとの間にある素性では，ツイートのタイプ，空文，2-gram のコサイン類似度を使用する．ツイートのタイプとは返信か引用かであり，空文とは，引用でありながら自身のコメントを付加していないものに発火する素性である．2-gram のコサイン類似度とは，返信・引用ツイートと返信先・引用元ツイートの 2-gram のコサイン類似度である．

返信先・引用元のユーザとの間にある素性として，返信・引用の回数，返信・引用の方向性，RT の回数，RT の方向性，共通のツイートを RT した回数，共通の URL をツイートに含んだ回数，共通のツイートに返信・引用した回数を利用する．方向性とは，やり取りが双方向に存在するか，一方通行にしか存在しないかを表す．

3 評価

3.1 交差検定による性能評価

データセット X の 2690 ツイートを，前述した 4 クラスに識別する多クラス分類を行う．本実験では，分類器として最大エントロピーモデルを用いる．分類器の実装として，Classias[3]の pegasos.logistic（L2 正則化ロジスティック回帰）を使用した．2 クエリずつに分け 10 分割交差検定を行い，合計を取ったものを表 1 に示す．

「同意」については感謝を表す表現（ex. ありがとう）や 2-gram のコサイン類似度などが,「疑問」については疑問符や問いかけの表現（ex. ですか）などが特徴的であり，この 2 クラスの識別性能は高い．一方,「反論」と「その他」の分類性能は低くなっている.「その他」については，ツイートの意図が広範囲にわたるため共通の素性を得にくい.「反論」については反論表現との一致度が有力な手掛かりとなるが，反論を示唆する表現を辞書で全てカバーするのは難しく，n-gram などの単純な素性では一般性が確保出来ないため，識別性能が低かったと考えられる．他に反論に特徴的なものとしては，返信や引用が一方通行であるという素性や,「お前」などの表現が見られるが，反論を決定づけるような特徴ではないため，反論に関する学習が十分に行えなかったと考えられる．

3.2 各素性の有効性

本手法において特徴的なのは，反論に特有な表現を持つ事例を分類するための反論表現辞書と，そのような表現を持たない事例を分類するための構造的特徴を用いたことである．そこで，これらの有効性を評価するべく，以下の比較を行った．1-gram・2-gram・元ツイートとのコサイン類似度・URL 数，ハッシュタグ数のみを特徴として使用する場合をベースライン（Base）とする．ベースラインにデマ表現と反論表現辞書との一致度を加えた場合を反論表現あり（+Dictionary），構造的特徴（返信か引用か・共通の RT や URL 引用回数・相互の RT や返信引用回数・相互の RT や返信引用の方向性）を加えた場合を構造的特徴あり（+Structure），反論表現辞書と構造的特徴を加えた場合を全使用（+All）とする．これらの 4 つの場合において，それぞれ分類器を構築し，性能を比較したのが表 2 である．本研究では,「反論」の識別が重要であると考え，マクロ F 値,「反論」クラスの F 値，精度の 3 指標で性能を比較した．

この結果から，反論表現辞書は反論クラスの識別に対して大きく寄与することが分かる．このことから，反論表現辞書を充実させることで，分類器の性能をさらに改善することができると考えられる．一方，構造的特徴は反論クラスの識別性能を多少上げるのに貢献しているが，他のクラスについては性能を下げることもあり，今後の継続的な検討が必要である．例えば，返信や引用の回数・方向性については，分類対象のツイートに関連するようにトピックや時系列を限定することで，より正確にツイート間の関係を表す特徴となるかもしれない．

[3]Classias のホームページ: http://www.chokkan.org/software/classias/

表 2: 分類器ごとの性能

	MacroF 値	「反論」F 値	Accuracy
Base	0.634	0.368	0.748
+Structure	0.642	0.389	0.740
+Dictionary	0.664	0.457	0.759
+All	0.667	472	0.751

3.3 考察

本研究の目的であるツイート間の論述構造解析では,「反論」を高精度で識別することが重要であるが，現状では反論表現を含まないツイートの識別は難しい．例えば次のような引用ツイートがある．

- 私が言うのも変ですが水で冷やしている限りメルトダウンはしません。問題が水が送れるか否かです。RT @xxx 水で冷やして「炉心溶融・メルトダウン」を停止できるという原子炉の専門家はいない。

この例では，引用元（@xxx のツイート）が水で冷やしてもメルトダウンを防げないと言っているのに対し，引用者はその逆の主張をしているが，引用者の発言中に反論特有の表現は見当たらないため「反論」と識別するのは容易ではない．現在我々は，言論マップの矛盾認識の技術 [1] など，深い言語処理と組み合わせた相補的な分析を検討している．

4 関連研究

言論マップ [1] とは文間関係を「同意」「対立」などに分類することを目指している点が共通しているが，ツイートは単文とは限らない上に書式が不正確なものも多く，そのままの適用は容易ではない．本研究では，反論表現や構造的特徴が利用しやすい返信・引用に絞り，深い言語処理を必要としない手法を提示した．

その他の情報の分類や構造の可視化を行う研究として，ディベートサイトの参加者を同意見を持つグループに分ける研究 [2] などが挙げられる．彼らはまずユーザ間のポストを感情極性の正負で分類し，正のポストが多いユーザ同士は同意見，負のポストが多いユーザ同士は別意見と仮定してグループを分け，高い精度を得た．本研究は，より広い意図が存在する Twitter のデータを対象とし，信憑性判断の助けとするため疑問の識別も行う点が異なる．

5 おわりに

本研究では，ツイート間の論述構造を分析・可視化することで情報の裏を取ることを支援する構想を紹介し，その第一歩として返信・引用で表明される態度に着目し，教師あり学習による 4 クラスへの分類を行った．反論表現辞書や構造的特徴を用いることで性能を向上できることを示し，特に重要な論述関係である「反論」の F 値で 0.472，4 クラス全体の精度で 0.751 という性能を得た．今後は，反論表現辞書の充実や構造的特徴の精選，あるいは文間関係認識などにおける深い言語処理との組み合わせを検討し，性能向上を目指す予定である．さらに，一般のツイート間の関係分類への拡張や，話題の要約などへの応用も考えている．

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．

参考文献

[1] 水野淳太, 渡邉陽太郎, 村上浩司, 乾健太郎, 松本裕治. 文間関係認識に基づく賛成・反対意見の俯瞰. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3408–3422, 2011.

[2] A. Hassan, A. Abu-Jbara, and D. Radev. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proc. EMNLP-CoNLL2012*, 2012.