Twitter における不快なつぶやきの検出

庄司 茜

ソーシャルネットワークサービスの代表格である，Twitter では短い文章のやりとりなどがなされており，他のユーザをフォローすることで，そのユーザの投稿を常に閲覧できる状態になる．従って，あえて特定の宛先に送らず，単なる投稿として質問文を投げかけた結果，多くのユーザの目に留まり，結果的に様々な回答が得られる場合がある．これは，広範囲に情報を発信するだけでなく，広範囲から情報が得られる有用なコミュニケーション手段と考えられるが，反面，ふとしたことで軋轢を生みやすいことも事実である．このようなトラブルを防ぐため，本研究では，不快なつぶやきのみをあらかじめ検出することで，Twitter 上でのコミュニケーションをより円滑にする基盤となる手法を提案した．提案にあたり，不快のつぶやきの傾向について，調査・分類を行った．調査に基づき，非難，自虐・自慢，わいせつの 3 グループに分類した．本研究では，3 種の内，「非難」，「自虐・自慢」が明確な手がかり語が多く得られやすいと考え，これらを「不快なつぶやき」として扱った．予備調査の妥当性を検証するため，調べたカテゴリ定義に基づき，2 名の被験者を用いて，双方が提供する「不快なつぶやき」の判定についてどれだけ一致するかを検証した．結果，不快さを感じる確率に関しては，それぞれの収集方法である程度の差が認められた．提案手法を実装するために，ユーザがタイムラインに反映しかねると判断したつぶやきとして，3 種類のサービスからつぶやきを学習データとして収集した．分類には，SVM ライブラリの一種である LIBSVM を用い，つぶやきデータの形態素解析には MeCab を用いた．実験では，SVM による 2 値分類において，正例データに Web 上で自動収集した愚痴や不満のつぶやき，負例データにパブリックタイムラインから自動収集した日本語つぶやきデータを暫定的に採用した．素性には，学習データのつぶやき中に現れる全ての品詞 (1)，名詞 (2)，動詞 (3)，形容詞 (4)，名詞・動詞・形容詞の組み合わせ (5) をそれぞれ用い，5 分割の交差検定を行い正解率を評価した．正解率の評価尺度としては，分類の正確さ（classification accuracy）を採用した．正例データには，関連サービス 3 種から 3 分の 1 ずつランダムに得たつぶやきデータ 1,000 件，負例データには，パブリックタイムラインから得た日本語のつぶやきデータのうちランダムに得た 1,000 件，合わせて 2,000 件を用いた．結果から，名詞・動詞・形容詞の内容語を素性とすることで，パブリックタイムラインから獲得したつぶやきと，3 種類のつぶやき収集サービスから獲得したつぶやきが，7 割程度の精度で分類できることが明らかとなった．この愚痴や毒舌のサービスを元に訓練データを収集した場合と, 先行研究に用いられた誹謗中傷を元に収集した場合とを，比較したところ，正確さと再現率は低下したが，精度はわずかに向上した．今後の方針としては，アノテーションマニュアルの精緻化や，クラウドソーシングサービスを用いたアノテーションにより，人手によって判定したデータを用いた評価を行う予定である．また素性についても，特徴素選択により素性を絞り込む効果などを検証していく．

（指導教員　関 洋平）