ISS-P-307

Twitterの発言における著者性別推定システムの検討

木村　颯斗†　　大山　実†

†東京電機大学 情報環境学部

1. はじめに

マイクロブログと呼ばれているTwitterにおける多数の発言を解析して，様々なアプリケーションに利用する事例が増えてきている．Twitterの解析において，発言者の性別が分かればより高度なアプリケーションへの利用が可能になる．そこで，本研究では発言者の性別識別に関する検討を行ったので報告する．

2. 手法

2.1 性別基準

先行研究[1]ではユーザの入力したプロフィール情報を性別の基準として用いている．この方式はユーザの入力した情報が正しいことを前提としている．本研究では性別が判明している協力者の発言の分析を行う．

2.2 特徴量定義

本研究では各Tweetを以下の①，②の二種類の手法を用いて解析し，その後，③の整形処理を行い性別推定の為の特徴量を抽出する．

①Yahooキーフレーズ抽出[2]を用いることによって，発言から複合名詞や係り受け上重要な文字列を抽出できるので，これをキーフレーズとする．

②MeCab[3]を用いて発言を形態素解析し，品詞ラベルを付与する．

③発言中に出現したキーフレーズとその周囲2形態素を抽出し，その組を発言の特徴とし，各々の頻度を数え上げる．数え上げの際には，各々の単語の品詞も同時に付与する．これは性別による発言の差が品詞の共起にも現れると推定したためである．

2.3 処理方式

男女の性別推定を行う際に線形分類器を利用する．本研究ではオンライン学習フレームワークとしてJubatus[4]を用い，アルゴリズムはAROWを利用する．

3. 実験

3.1 実験方法

実験で用いたユーザ数は男女それぞれ32人,33人である．学習データとしてランダムに選んだ男女各10名のユーザから各々500件，計10,000件の発言を用いた．また，分類には65名全てのユーザからランダムに発言を60,000件を抽出し，使用した．この中では既に学習された発言が含まれている可能性もあるが，含まれる比率は少ない．

3.2 実験結果

実験結果を表1に示す．それぞれ全体の識別率，先行研究の識別率，本研究の男女別の識別率を記載する．

表1．識別実験精度

項　目	精度(%)
全体識別率	64.4
先行研究	67.8
男性識別率	72.7
女性識別率	57.1

3.3 考察

女性の識別率が男性の識別率よりも低い原因は，本研究で用いた分類器は，その性質上先に学習させた事象の方を強く重み付けする．男女間で差が無い特徴を識別する際に,先に学習させたユーザ（この場合は男性側）の性別を正解と識別してしまったと考えられる．

4. むすび

キーフレーズと形態素を組み合わせた本方式では約64%の識別率となった．更なる精度の向上のためには，特徴量の再設計を検討する必要があると考えられる．性別の本人申告を信用している先行研究と比べて単純比較は出来ないが，ほぼ同一の識別率を得た．

また，日本人ユーザが主に利用している，顔文字やアスキーアートと呼ばれる記号を用いた表現等も使用して識別が行えないかの検討も今後行う．

---

[1] Discriminating gender on Twitter [John D. Burger, 2011]

[2] http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html [Yahoo!テキスト解析:キーフレーズ抽出,2014,2,11取得]

[3] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html [MeCab,2014,2,11取得]

[4] http://jubat.us/ja/ [Jubatus,2014,2,11取得]