O-001

Twitter の発言における著者性別推定システムの検討

A study on Gender Estimation system in Twitter

木村　颯斗†1　　大山　実†1
Hayato Kimura　　Minoru Ohyama

1. はじめに

マイクロブログと呼ばれている Twitter[1]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い，その性能を評価したので報告する．

2. 性別基準

2.1　Twitter について

Twitter は同社が提供するウェブサービスであり，ユーザーが端末を通して発言を行いコミュニケーションを図るツールである．この発言は 1Tweet(発言)当り 140 文字に制限されている．Tweet は図 1 のように発言中に”@id”が含まれている場合，その id(xxxx)を持つユーザーへの会話とし，相互に会話が出来る．また”#”はハッシュタグと呼ばれ，その話題についての発言であることを示す記法である．Twitter には自分のプロフィールを記載することができるが，性別の入力欄は存在しない．またプロフィールの一部として自身の WebSite 等の URL を記載することも出来る．

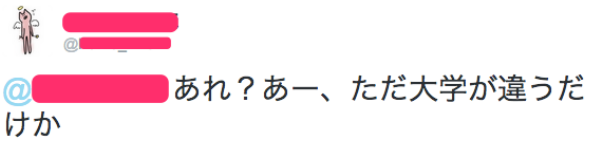

図1　実際の発言の例

2.2　性別の基準

John らによる先行研究[2]ではユーザの入力したプロフィール情報 を解析し，その中に性別が特定できそうな文字列が含まれていた場合や，別サービス(例：FaceBook)上や Blog への URL が記載されていた場合，そちらの情報を解析しその性別を確定している．本研究ではより確実な性別の基準として，既に性別が既知の協力者のデータを用いた．

†1 Tokyo Denki University

3. 性別推定手法

3.1　処理

本研究では 1 つの発言を「Yahoo キーフレーズ解析」と「MeCab による形態素解析」の 2 つの経路を用い解析し，それらを組み合わせる複合処理を行い，特徴量とした．実際には図 2 のようなフローで処理を行う．

3.2　Yahoo キーフレーズ解析

Yahoo キーフレーズ解析[3]とは Yahoo!Japan 社が提供するテキスト解析システムである．このシステムの API に文章を送信することで，その文章で重要な語句と，その語句の重要度を示すスコアを抽出することができる．これら重要度の高い語句を以下キーフレーズと呼称する．

3.3　MeCab

MeCab[4]とは形態素解析エンジンであり，今回は辞書として日本語の揺れにもある程度強いとされる UniDic を用いている．各発言を形態素解析し，形態素と品詞それぞれ取得する．

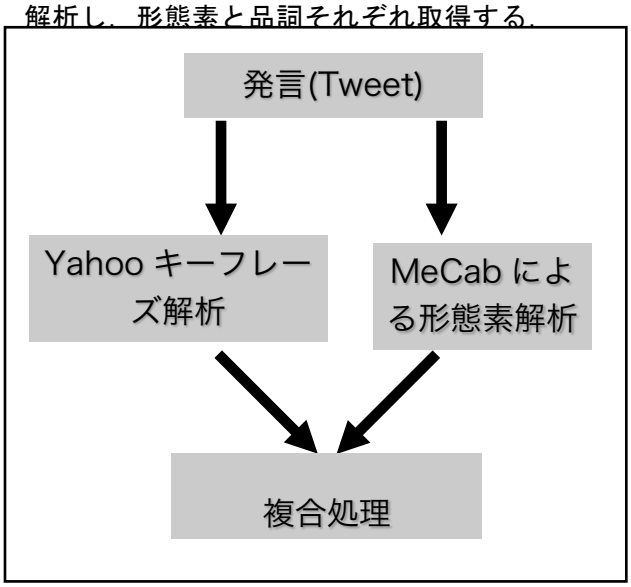

図 2:処理フロー

3.4　複合処理

この処理では最もスコアの高いキーフレーズと，その周辺にある形態素とを抽出する処理を

行う．キーフレーズが複数形態素にまたがっていた場合は，形態素を排除し，キーフレーズと該当形態素の周辺を抽出する．また，この際に形態素に付随する品詞も同様に特徴量として利用する．実際の例として「私達は新潟にこの電車で行く」というという発言をこの処理方式で処理した場合，形態素配列として[私,達,は,新潟,に,この,電車,で,行く]となる．同一の文字列を Yahoo キーフレーズ解析を用い解析を行うと，「新潟」と「電車」がキーフレーズとして抽出され，それぞれのスコアが 100,70 として解析される．この場合はスコアが最も高い新潟の周辺の形態素，[達,は,新潟,に,この]と各々に対応した品詞「接尾辞,助詞,キーフレーズ,助詞,連体詞」が特徴量となる．

3.5 分類器

各発言の男女の性別推定を行う際には，線形分類器を利用する．本研究では線形分類器のフレームワークとして Jubatus[5]を用い,アルゴリズムは AROW[6]を利用する．

4. 実験

4.1 実験に用いたデータ

協力者として合計 65 名（男性 32 名，女性 33 名）のユーザーの性別情報を得ることが出来た．また,それぞれのユーザーから最大で 3,600 件，合計 234,000 件の発言を取得した．

4.2 分類器による実験

学習用の教師データは、協力者全体から男女それぞれランダムに 10 名ずつ，計 20 名のユーザーから各 500 件，合計 10,000 件の発言を用いた．3 章で述べた手法により処理を行い，学習に用いた．また，分類には全ての協力者 65 名からランダムに発言を 60,000 件収集し処理を行い，分類を行った．この中には既に 3.5 節で述べた学習に利用された発言が含まれている可能性もあるが，十分に低い頻度だと考えられる．

4.3 人間による分類

分類器の精度を評価するため,実際の Twitter ユーザーに協力者の発言の性別推定を行ってもらった．35 名(男性 15 名，女性 20 名)の被験者に既に収集した 23 万件の発言からランダムに抽出した 100 件の発言を性別推定してもらった．35 名の被験者は殆どがデータ収集の協力者 65 名とは異なる人物である．

4.4 結果

4.2 節の分類器による実験，4.3 節の人間による実験での結果をそれぞれ表 1 に示す．識別率が本研究の精度を示し，男性の発言に対しての識別率が男性識別率，女性の発言に対しての識別率が女性識別率である．また，人間による識別率が 4.3 節の人間による分類の精度である．分類器による識別率（64.4%）は人間による識別率（62.1%）と比較して僅かに高い．また，分類器での男性に対する識別率（72.6%）の方が分類器での女性に対する識別率(57.1%)よりも高い．

表1 性別推定の精度

	精度
分類器による識別率	64.4
分類器での男性識別率	72.6
分類器での女性識別率	57.1
人間による識別率	62.1
先行研究識別率	67.8

４．５ 考察

人間による分類の精度と機械による分類の精度が同程度という結果により，1 発言当りの解析の精度はこれ以上大幅に改善することは難しいと考えられる．また，男女間で識別率に差があるのは，男性に多い極端に短い発言を本処理方式で解析を行うと，特徴量が 0 となり，発言を分類した際に男性と識別してしまう為であると考えられる．

5. おわりに

現在では発言単位についての性別推定であるが，実際のアプリケーションではユーザー単位での推定を行うことで，利用出来る場面が増えると考えられる．今後は精度の向上はもとより，ユーザー単位の性別推定や，ある話題についての男女の比率などを推定することを検討する．

[1]https://twitter.com/ [2014/5/16 日取得]
[2]Discriminating Gender on Twitter [John D. Burger,2011,http://www.mitre.org/publications/technical-papers/discriminating-gender-on-twitter]
[3]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,2014,5,16 取得]
[4]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html [MeCab,2014,5,16 取得]
[5]http://jubat.us/ja/ [Jubatus,2014,5,11 取得]
[6]Adaptive regularization of weight vectors. [Crammer,Koby,Kulesza,Alex,and Dredze,Mark.In NIPS, pp.345-352,2009b.]