ツイートの文章に使われている句読点に基づく属性推定

江口　大賀 †　　菊池　浩明 †
明治大学総合数理学部 †

1　はじめに

SNS 上で投稿される文章は，身の回りの出来事や感じた事を口語体で投稿される事が多い．そこで使われる言葉使いや単語は，ユーザーの年代や性別によって変わる事が予想される．例えば長浜らは，ツイートから得られた単語の χ 二乗値を用いるアルゴリズムを用いて，ユーザーの性別の推定を行った．その結果，男女間で比較すると，男子は「僕，俺」などの名詞を多用し，女子は「*)，ω」などの記号を多用する傾向を発見した．さらに，収集した文章の単語を「名詞，動詞，形容詞，副詞，連体詞，助詞，接頭詞，助動詞，接続詞，感動詞，記号」の 12 個に分類し，出現割合を比較も行った．その結果，品詞の出現割合では，男女間で大きな偏りがなかった [1]．

そこで，従来では自然言語処理の段階でストップワードに指定されて削除されがちである句読点等に注目する．本研究では，Twitter に投稿された文章から，ユーザーの年齢と性別の属性推定を行う事を目的とする．本実験のシステム構成図を，図 1 に示す．

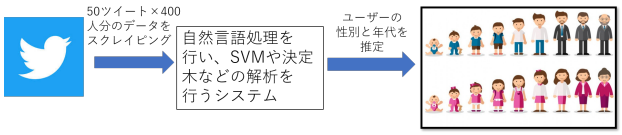

図 1　システム構成図

2　提案手法

2.1　データ収集

本研究のデータ収集では，プロフィールから Twitter のユーザを検索する「ツイプロ」[2] というサービスを用いる．このサービスは各アカウントのプロフィールから，年齢・性別・地域・職業・趣味などを自動分類している．「ツイプロ」を用いて，プロフィールから年代と性別が分かるユーザのツイートを収集する．

その後，収集したツイートを自然言語処理によって品詞分解する．合計 400 人分のツイートを収集したが，収集したユーザの属性と人数を表 1 に示す．1 人当たり 50 個のツイートを収集したので，総計 20,000 個のツイートを収集した．

表 1　収集したユーザの属性の統計値

性別\年代	10 代	20 代	30 代	40 代	合計
男性	50	50	50	50	200
女性	50	50	50	50	200
合計	100	100	100	100	400

2.2　各手法について

本実験では，2 つの手法による検証を行う．各手法の仮説と用いたデータ数を表 2 に示す．各手法の説明変数と決定変数を表 3 に示す．

表 2　手法の仮説とデータ数一覧

手法	立てた仮説	学習データ	評価データ
1	各属性で最も差が出る単語は、句読点である	200	200
2	若い年代は，句読点を使わない	100	100

表 3　手法の説明変数と決定変数

手法	説明変数	決定変数
1	400 人のユーザーの内の誰かが 3 回以上使用した 1,2052 単語の出現回数	10 代男，20 代男，30 代男，40 代男，10 代女，20 代女，30 代女，40 代女の 8 種類
2	「。」と「、」の 2 単語の出現回数	ある年代とそれ以外の年代の 2 種類

3　実験方法

様々な機械学習を用いて年代や性別の学習を行い，各属性を推定する．手法 1 ではランダムフォレスト，手法 2 ではサポートベクタマシンを用いた．

3.1　実験環境

本実験では，全て python 上で実行した．ツイートの収集では，urllib と pyquery を用いた．得られたツイートの自然言語処理には janome を用いた．属性推定における決定木とサポートベクタマシンには，sklearn を用いた．

Attribute estimation based on punctuation usage used in Tweets
†Taiga Eguchi and Hiroaki Kikuchi, School of Interdisciplinary Mathematical Science, Meiji University.

3.2 実験結果：特徴語の抽出

手法 1 から抽出された特徴語の上位 7 個と、その語の重要度を表 4 に示す．RandomForstClassifier を用いて，重要度を算出した．重要度とは，各々の説明変数の値が，目的変数を算出するのにどれ位重要かを示す物である．句読点の 2 単語が，重要度の上位であった．

表 4　特徴語と重要度

単語	重要度 [%]
を	0.93
。	0.87
、	0.72
私	0.70
まし	0.57
です	0.51

3.3 実験結果：句読点による世代ごとの分類

各世代の句読点の出現回数の平均を表 5 に示し，標準偏差を表 6 に示した．この結果から，世代が上がるにつれて，句読点の出現回数が多くなる事が分かった．

表 5　各世代の句読点の出現回数の統計量

句読点\年代	10	20	30	40
。	9.4	30.2	45.1	53.9
、	15.9	27.3	40.9	47.4

表 6　各世代の句読点の出現回数の標準偏差

句読点\年代	10	20	30	40
。	14.8	43.6	42.1	43.6
、	19.5	26.2	31.5	37.9

句読点の出現回数から，ある世代かそれ以外かの推定の結果を表 7 に示す．この表での再現率は，100 個の評価データの属性の予想の正解したデータ数の割合とする．10 代の再現率が，顕著に高い事が分かった．最も再現率が高かった 10 代の推定の結果を，散布図で図 2 に示す．

表 7　ある年代 100 人とそれ以外の年代 100 人の句読点による分類

	10 代	20 代	30 代	40 代
再現率 [%]	76.2	54.5	61.8	62.9

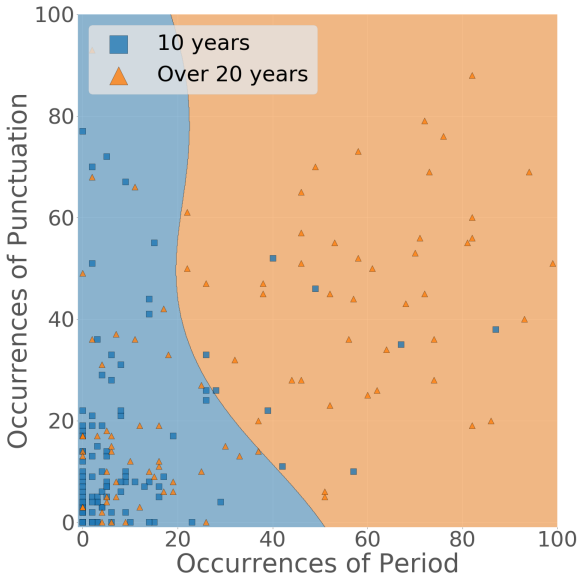

図 2　10 代 100 人と 20,30,40 代の 100 人の句読点による分類の散布図

3.4 考察

「ツイプロ」からユーザの収集をしている時感じた事だが，20 代以上の年代のユーザは，議員や役員などの役職が多かった．そのため，10 代のユーザより言葉使いが丁寧になると予想される．一方で 10 代のユーザーは，学生が殆どであり，20 代以上のユーザーより，単語のみのツイート等が多い傾向があると考えられる．以上のような理由から，10 代とそれ以外の世代で，句読点の使う頻度に差が出ると考察される．

4 おわりに

本研究では，従来の実験ではストップワードに指定されがちである句読点に注目し，性別や年代等の属性推定を行った．その結果，10 代とそれ以上の世代の推定には，句読点の出現頻度を比べる事が有効である事が分かった．今後は，文章中で「、」を使う直前の単語などにも注目して，年代や性別の推定も行いたい．

参考文献

[1] 長浜祐貴，遠藤聡志，當間愛晃，赤嶺有平，山田考治，”Twitter の投稿文章による人物像の推定”，JSiSE 学生研究発表会，2013.

[2] s21g Inc，”ツイプロ”，(https://twpro.jp/，2019 年 12 月参照)

[3] 岩朝史展，松本和幸，吉田稔，北研二，”Twitter ユーザの属性別感情推定の検討”，言語処理学会 第 22 回年次大会 発表論文集，pp.389-392， 2016.