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

Q)このツイートをした
ユーザの年代と性別は??

そうなんです。まだまだなんですけど、
ようやく少〜しだけ変化してきました！
来年2月に間に合わなかったら、その後の
大会まで無理せずじっくり作っていきます

男子??　女子??
10代？？
20代？？
30代？？
40代？？

正解は
40代女子

研究目的

SNSの文章で使われる
言葉使いや単語から,
ユーザーの年代や性別を推定する事.

先行研究

[長浜祐貴，”Twitter の投稿文章による人物像の推定”，JSiSE 学生研究発表会，2013．]

Twitterで使われている単語の χ 乗値を，
男女間で比較した

男子：「僕」，「俺」などを多用
女子：「*)」，「ω」などを多用
という傾向が分かった

本研究の新規性

従来ではストップワードに指定されて削除されることが多い句読点に注目した

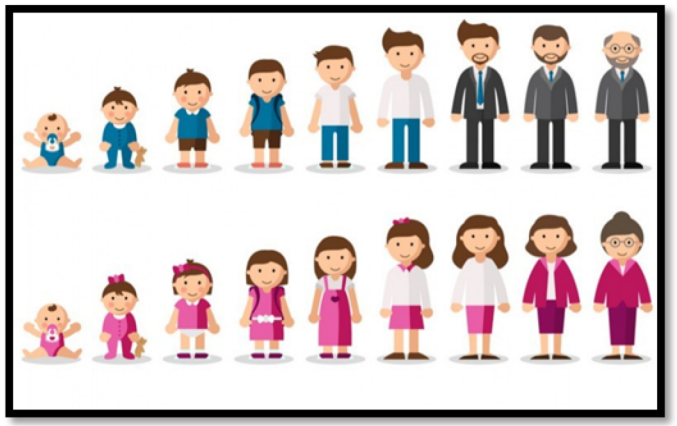
研究概要

句読点等に注目し，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．

20,000個の
データを
スクレイピング

自然言語処理を
行い、SVMや決定
木などの解析を
行うシステム

ユーザの
性別と年代を
推定



データ収集

「ツイプロ」を用いて400人分のツイートを収集した.
1 人当たり50個, 総計20,000個のツイートを収集した.

	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つの推定手法

手法	仮説	データ数	説明変数	目的変数	アルゴリズム
1	各属性で最も差が出るのは、句読点だ	400	どこかのユーザーで3回以上出現した12058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	ランダムフォレスト
2	若い世代は，句読点を使わない	200	「。」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	SVM

手法1：算出された特徴語と重要度の上位5個

特徴語	重要度
を	0.93
。	0.87
、	0.72
私	0.70
まし	0.57

各属性で最も差が出る単語は句読点だという事が分かった

重要度
＝説明変数の値がどれくらい目的変数を算出するのに重要か

手法2：句読点の利用頻度の世代差

統計値 / 年代	。		、	
	平均	標準偏差	平均	標準偏差
10代	9.4	14.8	15.9	19.5
20代	30.2	43.6	27.3	26.2
30代	45.1	42.1	40.9	31.5
40代	53.9	43.6	47.4	37.9

10代の句読点の出現回数が
少ない事が分かった

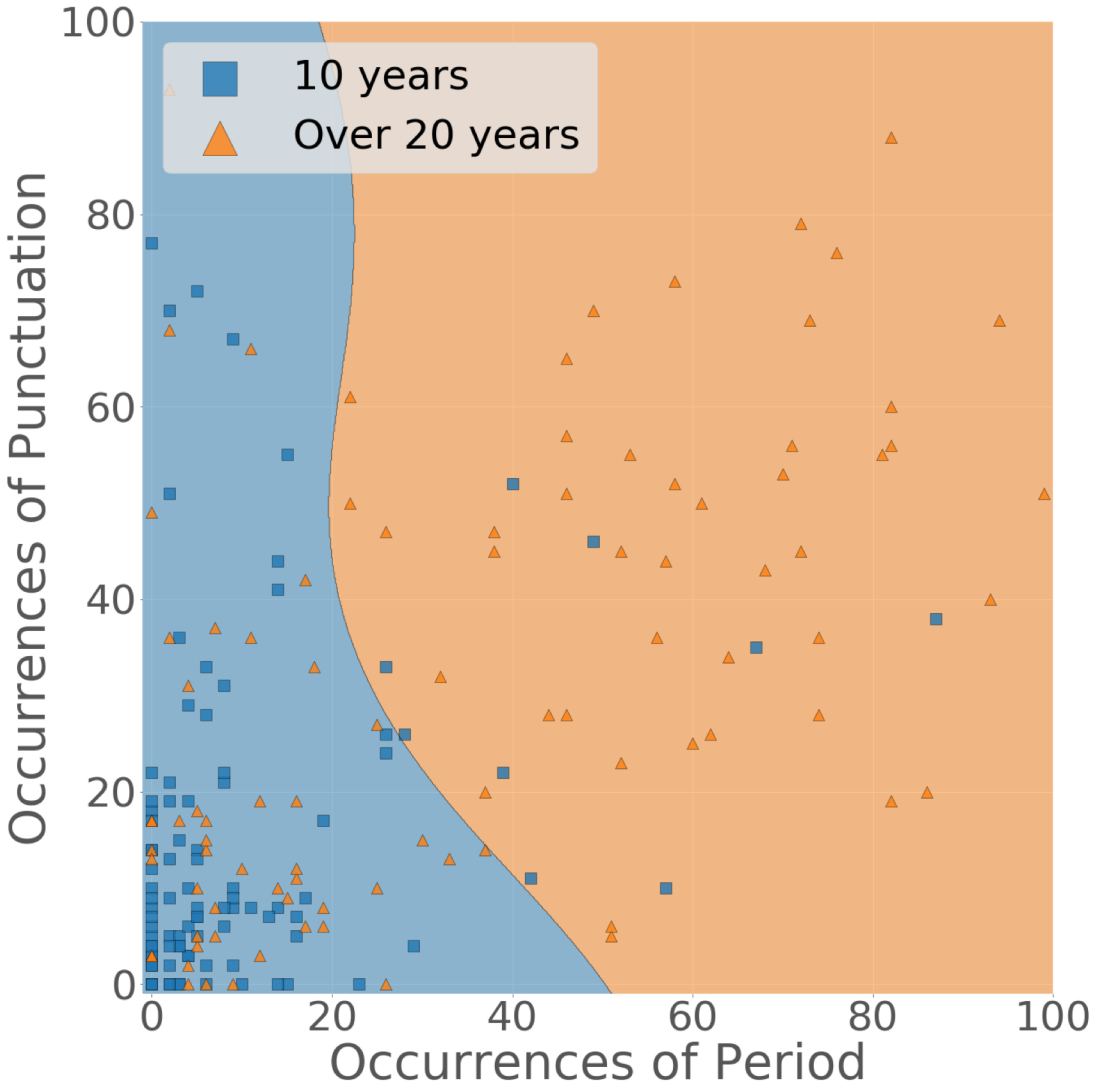
推定結果：手法2

ある年代 100 人とそれ以外の
年代 100 人の句読点による再現率

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

10代のみ、
高い精度で分類できた

10 代とそれ以外の句読点による散布図



人間と機械の年代推定再現率[%]の比較

大学生5人に、合計200人分のユーザの年代の推定を
行ってもらった.

	機械	人間
10代	76.2	48.0
20代	54.5	46.0
30代	61.8	24.0
40代	62.9	42.0

結論

10代とそれ以降の世代では，句読点の使用頻度に大きく差が生じた.
句読点は，他の単語と比較して，世代と性別を識別する重要な単語である.
今後の課題としては，読点の直前の単語にも注目したい.