ツイートの長さと句読点に基づく年齢・性別の推定

明治大学　総合数理学部
先端メディアサイエンス学科
菊池研究室 4年　江口大賀

研究背景

・10代女性のツイート例

あきばっち！
紅茶タピオカミルクティー
ホットでいちばん甘いの！！

・40代男性のツイート例

25,26の両日やるみたいですよ。
党のホームページに掲載されてます。
僕はどこに行くか未定です。

SNSの文章で使われる
言葉使いや単語は,
ユーザの年代や性別によって変わる事
が予想される.

先行研究

[長浜 2013]

Twitterで使われている単語の $\chi 2$ 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

[長浜祐貴，”Twitter の投稿文章による人物像の推定”，
教育システム情報学会学生研究発表会，2013．]

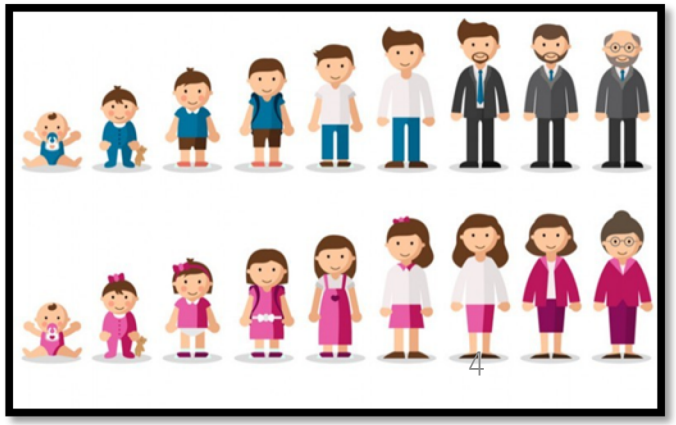
研究目的

句読点等に注目し，Twitter に投稿された文章から
ユーザ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



2019年度の研究の懸念点

若い年代の
句読点の使用回数が少ないのは,
投稿する文章量が少ないから??

10代男性のツイート：

ありがとう、イナズマイレブンSD。

40代男性のツイート：

私事ですが、先日、成田山新勝寺に家族で行き、紅葉まつりにお邪魔しました。

本研究の新規性

文章の長さも考慮し，
句読点を使用する割合に
注目した解析を行う．

提案手法：データ収集

「ツイプロ」とTwitterAPI等を用いて400人分のツイートを収集する．
１人当たり50個，合計20,000個のツイートを収集する．
2019年度のデータと合わせて，合計40,000個のツイートを収集する．

	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

提案手法：3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人のユーザの誰かが3回以上使用した単語(2020年度は1,1147個)の出現回数	400人のユーザの誰かが3回以上使用した単語の使用件数	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単語の使用件数	ある年代とそれ以外の年代の2種類	SVM
3	若い年代は，文章が短い		「。」と「、」の2単語の生起確率		

提案手法：句読点の生起確率とは

ユーザのツイートの文章の中に，
句読点がどれ位の割合で含まれるかを表した数値

句読点の生起確率

$$= \frac{\text{ユーザの句読点の出現回数}}{\text{ユーザのツイートの文字数の合計}}$$

例：読点(「、」のこと)の生起確率を求める

10代男性のツイート：

ありがとう、イナズマイレブンSD。

40代男性のツイート：

私事ですが、先日、成田山新勝寺に家族で行き、紅葉まつりにお邪魔しました。

10代男性の「、」の出現確率=1/17=0.058

40代男性の「、」の出現確率=3/36=0.083

実験方法

本実験では，全て python 上で実行する．

	用いたライブラリ
ツイートの収集	urllib，pyquery，TwitterAPI
収集したツイートの 自然言語処理	janome
属性推定で用いる 決定木(手法1)	sklearn
属性推定で用いる サポートベクタマシン (手法2・3)	

手法1の実験結果：
特徴語と重要度の上位5個

順位	単語	重要度[%]
1	。	0.84
2	私	0.74
3	、	0.62
4	を	0.61
5	まし	0.46

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
重要かを示す.

手法2の実験結果：句読点の使用件数の世代差

統計値 / 年代	平均		標準偏差	
	。	、	。	、
10代	7.1	15.5	11.8	15.9
20代	27.5	27.6	43.1	29.4
30代	43.3	39.3	39.5	32.5
40代	52.0	51.2	40.6	37.9

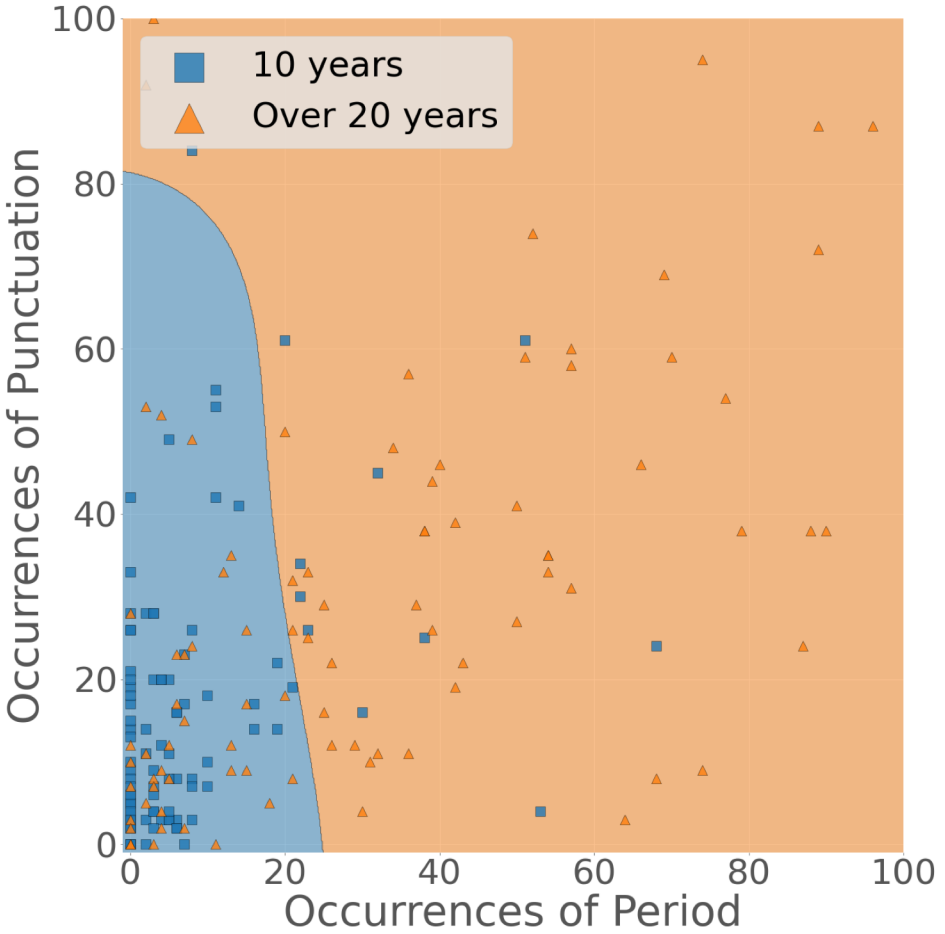
10代の句読点の使用件数が
少ない事が分かった

手法2の実験結果：年代の推定

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 [%]	70.6	56.0	62.3	67.4

10 代とそれ以外の句読点による散布図



**10代か20代以上かの
推定の再現率が高かった.**

2019年度の実験結果との比較

手法1:特徴語と重要度

	2019年度		2020年度	
順位	単語	重要度[%]	単語	重要度[%]
1	を	0.93	。	0.84
2	。	0.87	私	0.74
3	、	0.72	、	0.62
4	私	0.7	を	0.61
5	まし	0.57	まし	0.46

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
2019年度の再現率[%]	76.2	54.5	61.8	62.9
2019年度の再現率[%]	70.6	56.0	62.3	67.4

手法1の結果より,
句点(「。」のこと), 読点(「。」のこと)ともに
重要度の値が2019年度より下がった.

手法2の10代の推定の再現率が, 2019年度の結果よりも
5.6%低下した.

手法3の実験結果：
各属性の文字数(1ツイート当たり)と句読点の生起確率 [%] の統計値

	1ツイートの文字数		句読点の生起確率			
統計値 / 年代	平均	標準偏差	平均		標準偏差	
			。	、	。	、
10代	36.0	19.8	0.36	0.89	0.49	0.68
20代	46.8	22.6	1.06	1.08	1.20	0.84
30代	57.9	25.8	1.41	1.29	1.06	0.80
40代	63.4	28.0	1.54	1.51	1.06	0.86

年代が上がるにつれて，文章量が多くなる.

10代の句点(「。」)と読点(「、」)の生起確率は
顕著に低い.

手法3の実験結果：年代の推定

ある年代 100 人とそれ以外の
年代 100 人の句読点の出現確率による再現率

	10代	20代	30代	40代
再現率 [%]	73.0	55.9	59.2	66.0

10代か20代以上かの推定の再現率が
73.0%であり，最も高い結果となった.

これは手法2の推定よりも，2.6%高い値である.

結論

手法3の結果より,
年代が上がるにつれて,
文章量と句読点を使用する割合が
高くなる.

また, 10 代か 20 代以上かの推定においては,
句読点の使用割合に注目する事が
有効である事が分かった.