Can Conversational Word Usage be Used to Predict Speaker Demographics?

Dan Gillick
University of California, Berkeley

Report：許志宇 黃予宏

Professor：陳嘉平

Abstract

This work surveys the potential for predicting demographic traits of individual speakers (gender, age, education level, ethnicity, and geographic region) using only word usage features derived from the output of a speech recognition system on conversational American English. Significant differences in word usage patterns among the different classes allow for reasonably high classification accuracy (60%-82%), even without extensive training data.

Index Terms: demographics, speech recognition, classification

Introduction

主要使用Mixer corpora 在 NIST 2008 Speaker Recognition ，此語料庫總共有1336位語者的資料 。

Trait	Summary Statistics
Gender	female (844); male (492)
Yr. of Birth	1922 - 1990; median=1974
Yrs. of Education	1 - 30; median=16
Native Language	U.S. English (765); 33 others
Occupation	Student (230); Homemaker (41) ...
Country Raised	U.S. (947); India (58); China (41) ...
State Raised	40 U.S. states represented
Ethnicity	White (457); Asian (388); Black (129) ...
Smoker	yes (142)
Height (cm)	134 - 198; median=168
Weight (kg)	36 - 160; median=68

Introduction

主要專注在基於 Automatic Speech Recognition(ASR)所輸出的n-gram的文字特徵進行分類工作，將每一種人口特徵(性別，年齡，教育程度，種族，地緣）在文字使用的差異，籍此進行群組分類，發現擁有另人訝異的驚確度。

Related Work

在聲學上：

利用聲學特性（如音調，抖動，閃爍，音量）與行爲特徵（如說話速度）可用於預測語者的年齡和性別，比如像 Muller 的研究，用類似的方法也可以用來做方言的辨別，來推測語者的地緣關係。

Related Work

在文字上：

Koppel的實驗預測了小說與非小說類出版物作者的性別

Schler的實驗預測了blog上以非正式口語對話語者的性別和年齡

（在13-17, 23-27, 33-42年齡群組中有接近80%的精確度）

Data

取用2008NIST Speaker Recognition Evaluation 資料集(Mixer Corpora的子集合)，包含538本地美國人的語音資料，利用SRI Decipher system進行ASR 的轉譯，並加入文字的錯誤率。

資料的分類如下

Gender		Age		Education		Ethnicity		Region	
Male	39%	20-29	31%	High school or less	18%	White/Caucasian	63%	Northeast	40%
Female	61%	30-39	27%	College	52%	Hispanic/Latino	6%	South	21%
		40-49	20%	More than college	30%	Black/African-American	19%	Midwest	16%
		50+	22%			Asian	12%	West	23%

Table 2: Classes derived from demographic data.

Data

在實驗之前，我們必須了解人口特徵之間相關性，像是年齡與學歷之間的相關度高，可能會有不同的作用。使用 Logistic regression models 以量化各類別間的關係。

Data

Traditional regression models as:

$$R^2_{McFadden} = 1 - \frac{\log \hat{L}(M_{full})}{\log \hat{L}(M_{intercept})}$$

M_{full} = Model with predictors

$M_{intercept}$ = Model without predictors

$\hat{L}$ = Estimated likelihood

Data

	Gender	Age	Educ	Eth	Region	ALL
Gender	–	0.006	0.006	0.008	0.013	0.036
Age	0.003	–	0.021	0.035	0.022	0.069
Educ	0.004	0.023	–	0.034	0.013	0.069
Eth	0.005	0.046	0.033	–	0.049	0.116
Region	0.007	0.023	0.005	0.032	–	0.061

值必須介於0~1，愈大代表人口特徵的相關愈高

Data

根據迴歸係數得知：

1.年齡和教育程度成負相關

年輕人受的教育程度較老年人高

2.亞洲人比較年輕且來自東方

非洲人比較年老且來自南方

3.亞洲人所受的教育程度較高

黑人及非裔的美國人和拉丁美洲人教育程度較低

Feature Selection

用一些n-grams做分類發現

使用bigram來分類比unigram有較好的表現。

資料中含有許多不含辨別資訊的bigram，所以根據Information Gain Formula選擇前2000個bigram。

Information Gain Formula

$$
\begin{aligned}
IG_f &= -\sum_c P(c)\log P(c) \\
&\quad +P(f)\sum_c P(c|f)\log P(c|f) \\
&\quad +P(\bar{f})\sum_c P(c|\bar{f})\log P(c|\bar{f}) \qquad (2)
\end{aligned}
$$

f: feature(bigram)

P(c): Entropy 資料亂度

c: class

Feature Selection

Class	Bigrams	
Female	my husband	oh my
Male	uh that's	uh uh
20-29	and like	cool [laugh]
30-39	it definitely	was living
40-49	excuse me	for president
50+	many years	kids that
H.S. or less	not into	yes uhhuh
College	I'm working	lot about
Grad. School	of other	yeah or
White/Caucasian	in touch	watch the
Hispanic/Latino	um more	my thing
Black/African-American	may have	them up
Asian	mm yeah	oh like
Northeast	gone to	was no
South	it's [laugh]	different in
Midwest	choose to	opposed to
West	cool and	about your

Table 4: Some sample high Information Gain bigrams are shown for each class.

Classification Experiments

用bootstrap method取得可靠的分類錯誤率估算

而分類器使用

Margin Infused Relaxed Algorithm(MIRA)

MIRA非常快速，可在一秒內跑完一個model

Experiment

	Baseline	MIRA	% Improvement
Gender	39%	18%	54%
Age	66%	35%	47%
Education	49%	33%	33%
Ethnicity	38%	28%	26%
Region	62%	40%	35%

Table 5: Baseline error rates (always predict the majority class), classifier error rates, and relative improvements are shown for each demographic trait.

Experiment

	Actual Gender	
	Male	**Female**
Male	85%	9%
Female	15%	91%

	Actual Age			
	20-29	**30-39**	**40-49**	**50+**
20-29	81%	25%	13%	10%
30-39	11%	52%	17%	9%
40-49	4%	11%	49%	9%
50+	4%	12%	21%	72%

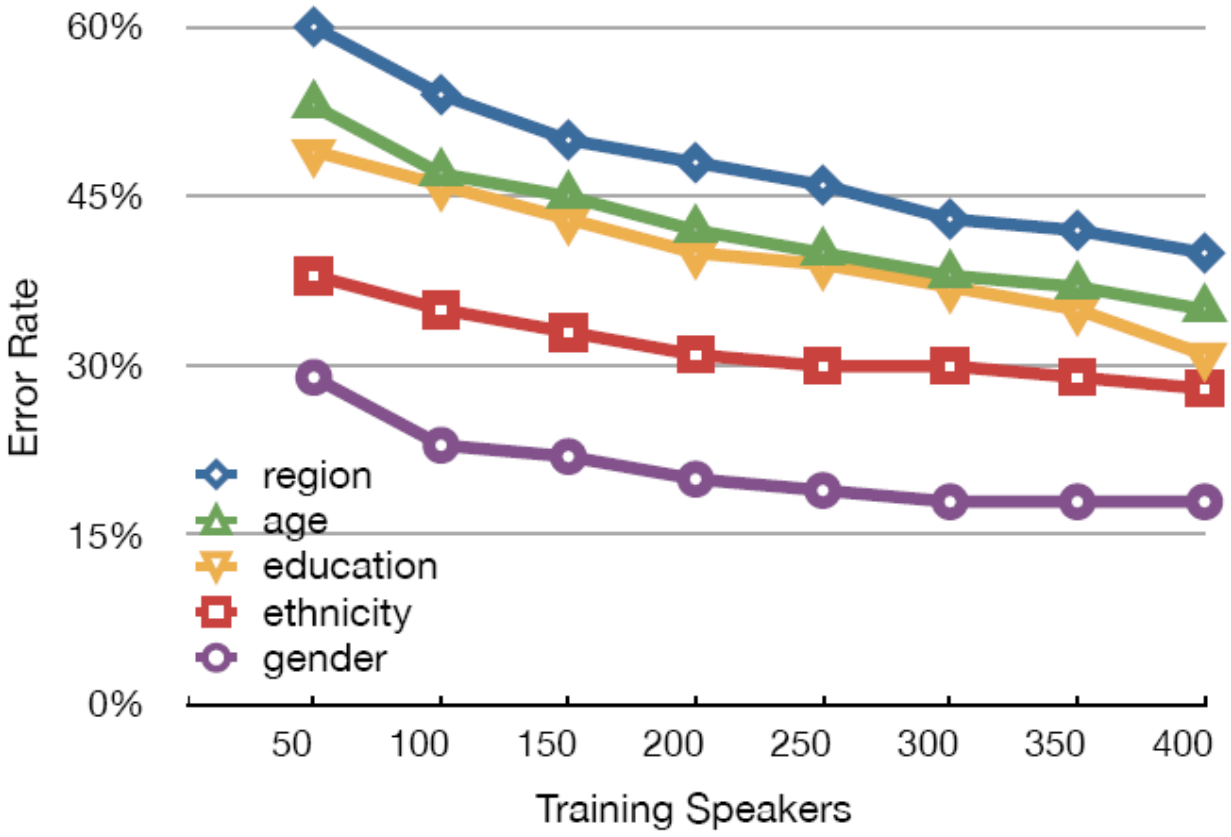
	Actual Education Level		
	H.S. or less	**College**	**Grad. School**
H.S. or less	56%	5%	6%
College	35%	79%	41%
Grad. School	9%	16%	53%

Experiment

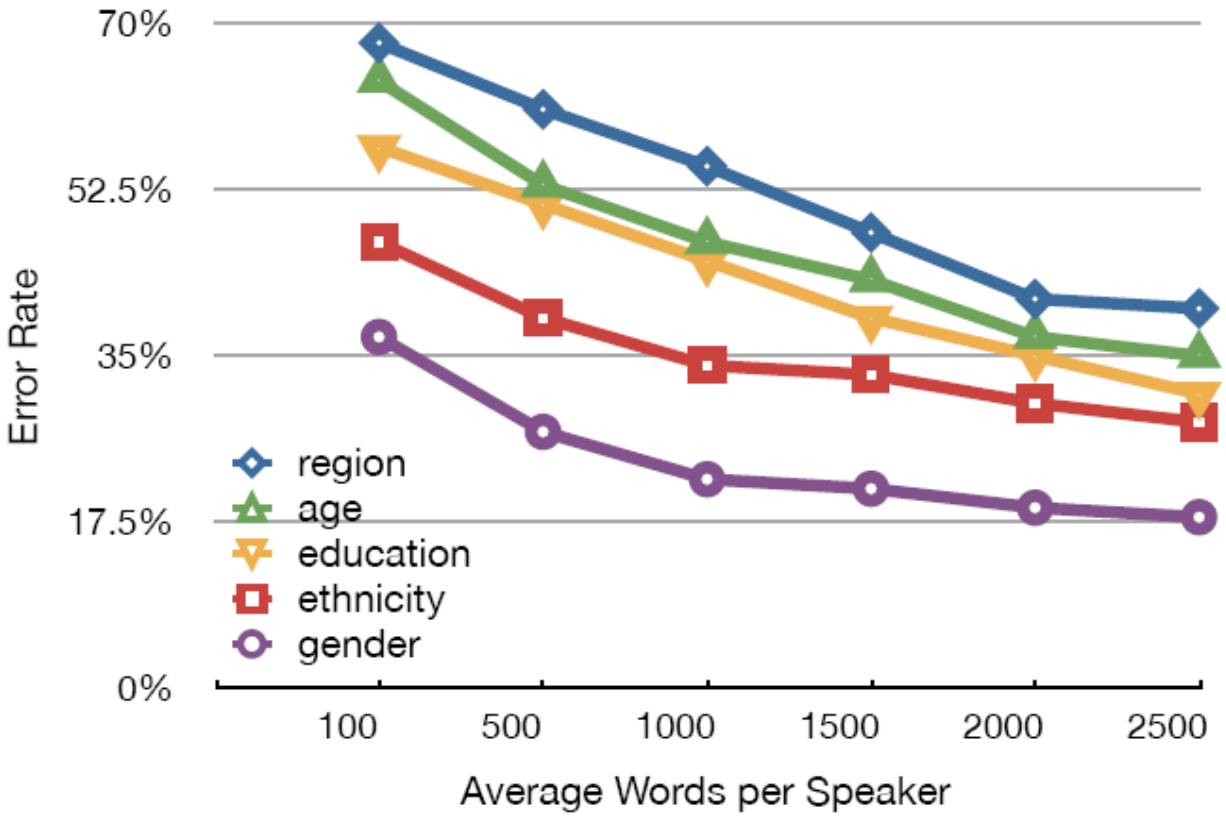
	Actual Ethnicity			
	White	**Hispanic**	**Black**	**Asian**
White	88%	47%	29%	42%
Hispanic	1%	14%	1%	1%
Black	5%	28%	61%	22%
Asian	6%	11%	9%	35%

	Actual Region			
	Northeast	**South**	**Midwest**	**West**
Northeast	70%	17%	26%	22%
South	10%	56%	20%	9%
Midwest	9%	14%	38%	7%
West	11%	14%	16%	62%

Experiment


60%
45%
30%
15%
0%
Error Rate
region
age
education
ethnicity
gender
50
100
150
200
250
300
350
400
Training Speakers

Experiment


70%
52.5%
35%
17.5%
0%
Error Rate
region
age
education
ethnicity
gender
100
500
1000
1500
2000
2500
Average Words per Speaker

Future Work

可以收集大量的MSN文字訊息進行訓練，從文字訊息中辨識出發送者的地緣關係

如：中南部人會在文字訊息中加入台語或方言