Language Influences on Tweeter Geolocation

Ahmed Mourad, Falk Scholer, Mark Sanderson
{ahmed.mourad, falk.scholer, mark.sanderson}@rmit.edu.au

ECIR 2017

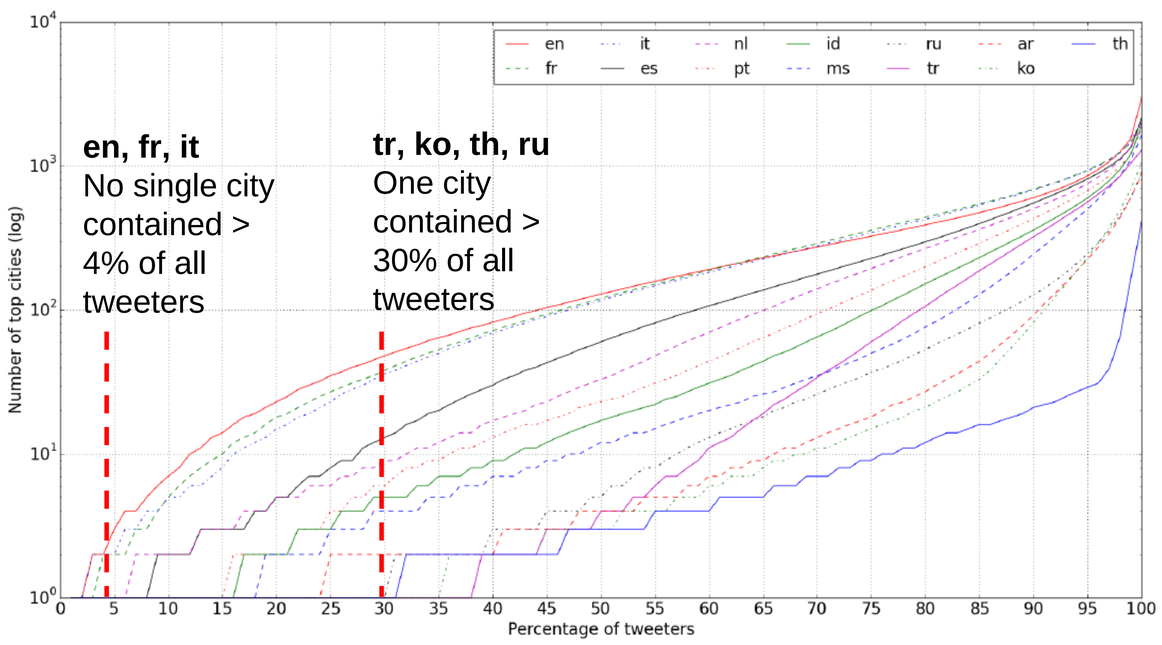
Hypothesis

Users tweeting in languages with restricted geographical coverage (e.g. Turkish, Arabic, Russian and Indonesian) are easier to geolocate compared to English. [1]

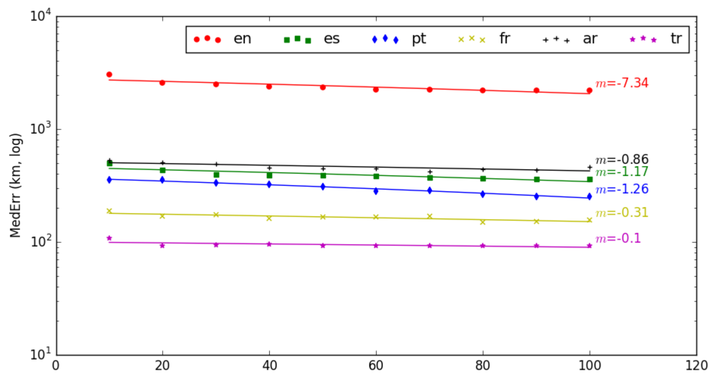
Every language is GLOBAL

		en	es	it	pt	id	nl	fr	ms	ko	ru	ar	th	tr
#cities (k)	WORLD	2.9	2.2	2.1	1.8	1.9	2	2	1.6	1.1	0.9	0.9	0.4	1.3
	TwArchive	3.2	2.3	2.2	1.9	2	2	2.2	1.7	1.7	1	1.6	0.7	1.6

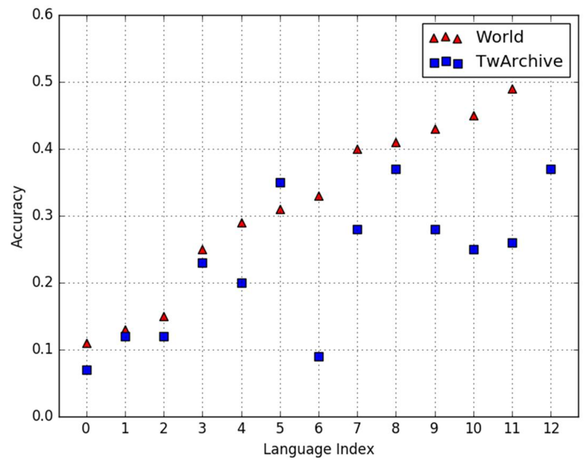
Some languages are MORE imbalanced than others



Median error is NOT an appropriate measure for small datatsets



Relative accuracy for a language changes across data collections



Correlation with features

	Feature	Acc				Acc@161				MedErr			
		WORLD		TwArchive		WORLD		TwArchive		WORLD		TwArchive	
		r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2
1	Entropy	**-0.87**	**0.76**	-0.69	0.47	-0.62	0.38	-0.29	0.08	0.52	0.27	0.43	0.19
	#Cities	-0.76	0.57	-0.4	0.16	-0.57	0.32	-0.26	0.07	0.54	0.3	0.57	0.32
	Entropy.test	-0.83	0.69	-0.7	0.49	**-0.85**	**0.73**	**-0.79**	**0.62**	0.82	0.68	0.89	0.79
	#Cities.test	-0.55	0.3	-0.51	0.26	-0.67	0.45	-0.55	0.3	0.81	0.66	0.93	0.87
	#LIW words	0.4	0.16	0.37	0.14								
2	Avg dist					0.12	0.01	0.51	0.26	-0.33	0.11	-0.3	0.09
	Nbr avg dist					-0.46	0.21	-0.22	0.05	0.55	0.31	0.53	0.28

1. Distribution of tweeters (entropy) has a greater impact on the accuracy of geolocation prediction than other features
2. Acc@161 is not an appropriate measure

Accuracy skewed towards big cities

Prec.	en	es	pt	fr	ar	tr	id	it	nl	ru	ms	th	ko
$MC\ P_\mu$	0.02	0.12	0.23	0.10	0.39	0.54	0.27	0.09	0.16	0.34	0.25	0.32	0.45
$MNB\ P_\mu$	0.11	0.29	0.31	0.13	0.49	0.54	0.40	0.15	0.25	0.33	0.41	0.43	0.45
$MC\ P_M$	0.000	0.000	0.001	0.000	0.004	0.007	0.002	0.000	0.003	0.003	0.002	0.008	0.006
$MNB\ P_M$	0.047	0.027	0.036	0.033	0.059	0.027	0.079	0.018	0.077	0.006	0.086	0.267	0.046

How languages compare to each other using different averaging techniques?

Micro (μ) averaging is biased towards big cities, while Macro (M) averaging does not

Across Averaging Techniques		Precision				Recall			
		WORLD		TwArchive		WORLD		TwArchive	
		W	M	W	M	W	M	W	M
	μ	0.41	-0.08	0.38	0.08	1.00	0.08	1.00	0.15
	M	0.00		0.08		0.05		0.15	

Geographical coverage has LESS impact than Expected

Across Data Collections	Precision			Recall		
	μ	W	M	μ	W	M
	0.46	0.13	0.00	0.46	0.49	0.03

Conclusion

- Distribution of tweeters over cities is strongly correlated to accuracy
- Scale of a test set was found to have little influence on accuracy
- Reporting both micro and macro averaging, or using a weighted average, provides valuable additional insight
- Other geolocation techniques will be considered
- Language influence needs to be investigated further

References

[1] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. Journal of Articial Intelligence Research, pages 451—500, 2014.

Acknowledgements

This work was made possible by NPRP grant# NPRP 6-1377-1-257 from the Qatar National Research Fund (a member of Qatar Foundation).

RMIT UNIVERSITY