Twitter Geolocation and Regional Classification via Sparse Coding

Miriam Cha, Youngjune Gwon, and H.T. Kung

HARVARD School of Engineering and Applied Sciences

Introduction

Twitter

- Most popular microblog service
- 500 million tweets per day
- Over 270 million active users

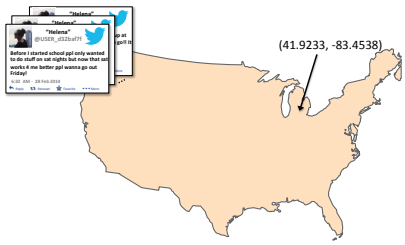
Motivation

- Twitter's location service enables users to add location information to their Tweets
- Location information of message is useful (*e.g.*, consumer marketing)
  - However, < 3% of messages are geo-tagged†
- Machine learning can geolocate based on message content
  - Previously, best results are from supervised model-based approaches
- We focus on unsupervised data-driven approach, namely sparse coding, to take advantage of abundance of unlabeled messages to geolocate

†C. Weidemann, Journal of Geoinformatics

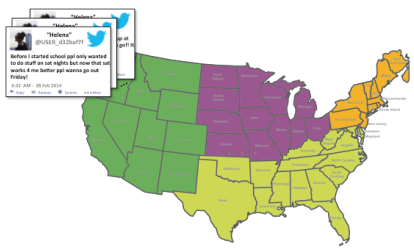
Objectives

Geolocation



- Estimate latitude and longitude with Twitter message content

Region and State Classification



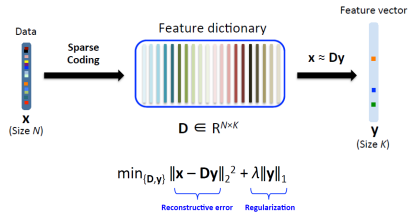
- Classify US state and region using message content

Develop Twitter geolocation and regional classification based on sparse coding and dictionary learning

Approach

Sparse Coding

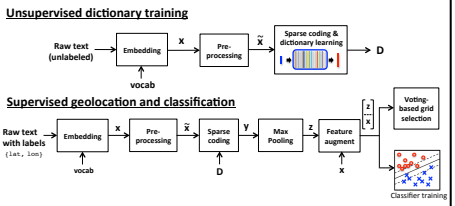
Represent **x** as a linear combination of basis vectors in **D**



$$\min_{(\mathbf{D},\mathbf{y})} \|\mathbf{x}-\mathbf{D}\mathbf{y}\|_2^2 + \lambda\|\mathbf{y}\|_1$$

Reconstructive error | Regularization

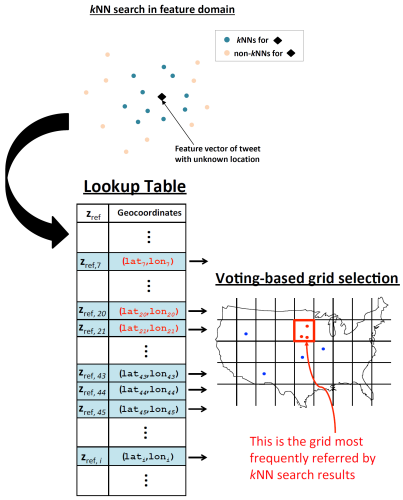
Geolocation and Region/State Classification



Voting-based Grid Selection in Geolocation

Similarity matching in feature space (sparse code domain)

- Find *k*-Nearest Neighbors (*k*NNs) in sparse codes
- Look up reference locations of *k*NNs to compute geocoordinate estimate



Evaluation

CMU GeoText Dataset

- Geo-tagged microblog corpus
  - http://www.ark.cs.cmu.edu/GeoText/
  - 377,616 Twitter messages from 9,475 users within continental US over one week
  - All user geolocations known (latitude, longitude)

Experimental Methodology

- Three types of embedding schemes (binary bag-of-words, word-counts, word-sequence)
- For a fair comparison, we follow the identical experimental methodology as Eisenstein *et al.* (2010)

Results

Geolocation errors. Values are mean (median) in km

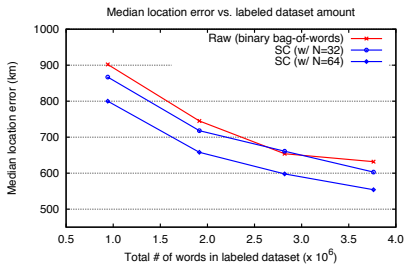
	Raw	SC	SC+PCA	SC+Raw	SC+voting	SC+all
Binary	1024 (632)	879 (722)	748 (596)	861 (713)	825 (529)	707 (489)
Word counts	1189 (1087)	1042 (887)	969 (802)	1022 (863)	998 (511)	926 (497)
Word sequence	–	767 (615)	706 (583)	671 (483)	715 (580)	**581 (425)**

Our method applied to word-sequence vectors achieves the best performance

Performance comparison summary

	Geolocation error		Classification accuracy	
	Mean	Median	Region	State
Our approach	**581**	**425**	**67%**	**41%**
Eisenstein *et al.*	845	501	58%	27%
W&B	967	479	–	–
Roller *et al.*	897	432	–	–

Achieves 9% gain for region classification and 14% gain for state classification



SC gives decreased error by ~100 km compared to Raw throughout all reported amounts of labeled data

As the amount of labeled training samples is limited in practice, such advantage of sparse coding is attractive

Conclusion and Future Work

Conclusion

- Twitter geolocation and regional classification can benefit from unsupervised, data-driven approaches such as sparse coding and dictionary learning
- Achieve competitive performance standing at 581 km mean and 425 km median location errors for GeoText dataset
- Our classification accuracy for 4-way US regions is a 9% improvement over the best topical model in prior literature, and the 14% for 48-way US states

Future Work

- Evaluation in both US and worldwide data
- Hybrid method that can leverage the strengths of data-driven and model-based approaches

Acknowledgement

This work is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1144152 and gifts from the Intel Corporation