Microblog Geolocation using Language Variation Deep Learning

David Zucker
Department of Computer Science
Stanford University
Stanford, CA 94305
zuckerd@stanford.edu

Introduction

This experiment investigates the feasibility of geographically locating Twitter users based solely on tweet content through the identification of geographic regional language and dialect patterns. Currently, fewer than 3% of current tweets are configured to include geographical information (Liu, 2015) and this experiment provides an approach to augment existing user location efforts. This experiment considers whether accurate geolocation is possible through the use of recurrent neural networks with without the use of external information including user specified 'hometown' data, named entity recognition of location names, gazetteer lists, or Twitter social graph data.

The ability to locate Twitter users based on tweet content has many practical applications including improving our understanding of language and dialect differences across geographies, measuring changes in language patterns in online media, detecting anomalies in tweet speech patterns to identify users not speaking the dialect of their current location (e.g., travelers, or non-native speakers), and segment users by geography for marketing or emergency response purposes.

Background and Related Work

Three major studies have been conducted in the past several years focusing on geolocation of Twitter users based on geographic language and dialect patterns. Many previous experiments relied heavily on statistical NLP techniques and analysis to drive geolocation predictions. Results for previous studies are included in the figure below.

	Models	Region (%)	State (%)
Eisenstein (2010)	Geo topic	58.0	24.0
	Unigram	53.0	19.0
	LDA	39.0	4.0
	Regression	41.0	4.0
	kNN	37.0	2.0
Liu (2015)	SDA-1	61.1	34.8
	Baseline Naïve Bayes	54.8	30.1
	Baseline SVM	56.4	27.5
Cha (2015)	Sparse Vector	67.0	41.0

In 2010, Eisenstein, et. al, conducted the first major study attempting to predictively locate Twitter users. Their multi-pronged experiments included topical modelling, k-nearest neighbors, and several statistical methods including LDA and regression. Eisenstein successfully located users at the 4-way regional and 48-way state level with 58% and 24% accuracy, respectively.

In 2015, Liu and Inkpen expanded upon Eisenstein's previous work through the use of a 3 layer stacked denoising autoencoder feed forward neural network. Their experiment consisted of combined classification and regression efforts that predicted latitude and longitude coordinates and binned them into classes. Using the Eisenstein corpus, Liu and Inkpen improved upon Eisenstein's accuracy results by 3% and 10% on the 4-way regional and 48-way state level classification tasks, respectively.

In 2015, Cha, Gwon and Kung expanded upon Eisenstein's previous efforts through the use of unsupervised sparse vector training and supervised classification to predict user location based on *k*-nearest neighbor tweets with a cosine similarity measure. Using the Eisenstein corpus, Cha

improved upon Eisenstein's accuracy results by 9% and 14% on the 4-way regional and 48-way state level classification tasks, respectively.

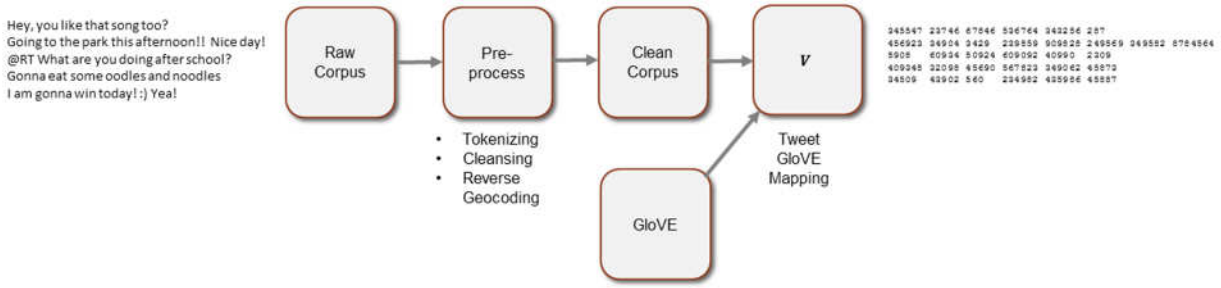
Based on research, I was unable to identify any significant previous studies leveraging recurrent neural network (RNN) approaches to predict Twitter user geolocation based solely on tweet content. This seems surprising considering the prevalence of social media data and the recent increased level of research in the deep learning field. Given the size and quality of the Eisenstein corpus, I intend use this experiment to expand upon the previous work stated above and attempt to improve geolocation accuracy.

Approach

The general approach of this experiment is to build a recurrent neural network (RNN) consisting of one or more LSTM layers to classify tweets into geographical 4-way regions as defined by the United States Census Bureau region and division mapping.[1] This experiment will leverage pretrained GloVe[2] word vectors and the Carnegie Mellon (CMU) Eisenstein Twitter corpus[3] (Eisenstein corpus) to train a neural network capable of predicting a given user's location based on a previously unseen tweet.

The neural network model is trained and evaluated on the Eisenstein corpus collected by Eisenstein as part of his experiments in 2010. The corpus contains 377,616 tweets from 9,475 unique users as collected during a 7 day period in March 2010. All tweets in the corpus include geolocation information as part of the raw dataset and correspond to physical locations within to the continental United States.

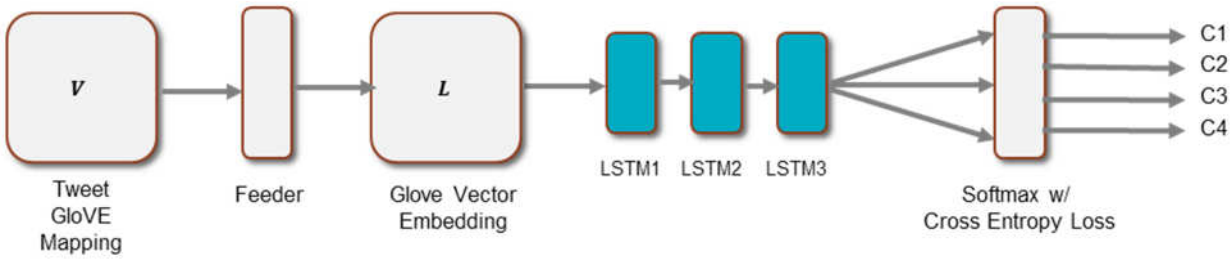
Prior to conducing the experiment, a preprocessing pipeline is constructed to prepare and standardize the tweet data. Preprocessing steps include tokenization using the included corpus tokenizer, cleansing to remove excessive punctuation and retweet tags, and reverse geocoding[4] to convert latitude / longitude data points into standardized street addresses. Reverse geocoding has the added advantage of providing the user state location which can be mapped to the regional categorical variable. Tweets containing only punctuation, retweet tags or other usernames are removed. Since latitude / longitude are recorded at the tweet level, a given user can have multiple geolocations. To account for this phenomenon, all users are assigned the geolocation coordinates of their first tweet.



Embeddings, L_1 through L_4 are constructed to capture vectorized representations of the tweet data. This is accomplished through the construction of a unique tweet vocabulary and a lookup against the precalculated GloVe twitter word vectors collected from 27B tweets (reference). GLoVe vectors of length 25, 50, 100, 200 are leveraged in this experiment. Once fully processed, the corpus contains approximately 368,000 tweets and a 142,182 words vocabulary.

[1] https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
[2] http://nlp.stanford.edu/projects/glove/
[3] http://www.cs.cmu.edu/~ark/GeoText/
[4] Reverse geocoded using mapping services provided by http://maplarge.com

The neural network architecture includes an input layer to feed vectorized tweet data into the network, an embedding layer containing embedding objects L_1 through L_4, a multi-component hidden layer consisting of between one and three RNN LSTM layers, and an output layer to map predictions to the regional classifier.



Each of the LSTM layers utilize the following equations where i, f, o correspond to input, forget, and output, respectively.

$$i = \sigma(x_t T^{(i)} + s_{t-1} W^{(i)})$$
$$f = \sigma(x_t U^{(f)} + s_{t-1} W^{(f)})$$
$$o = \sigma(x_t U^{(o)} + s_{t-1} W^{(o)})$$
$$g = tanh(x_t U^{(g)} + s_{t-1} W^{(g)})$$
$$c_t = c_{t-1} \circ f + g \circ i$$
$$s_t = \tanh(c_t) \circ o$$

The following softmax and cross entropy equations are used in the final output layer to generate categorical region predictions.

$$softmax(\boldsymbol{x})_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_i y_i log(\frac{e^{\theta_i}}{\sum_j e^{\theta_j}})$$

The neural network model is then trained on the Eisenstein corpus with each tweet considered a training example $(x^{(i)}, y^{(i)})$ where $x^{(i)}$ is the tweet and $y^{(i)}$ is the tweet regional classifier. The success of the model is then evaluated by comparing the predicted regional classifier location $(\hat{y}_i)$ against the ground truth 4-way regional classifier (y_i).

To evaluate the success of the deep learning models, the cross entropy loss (negative log likelihood) will be used in a softmax output layer to generate the input tweet predicted geographic location $(\hat{y}_i)$. This predicted location will be compared against the actual ground truth tweet geographic location (y_i). This process will be repeated across all geographical categories included in the experiment.

Experiment

Results

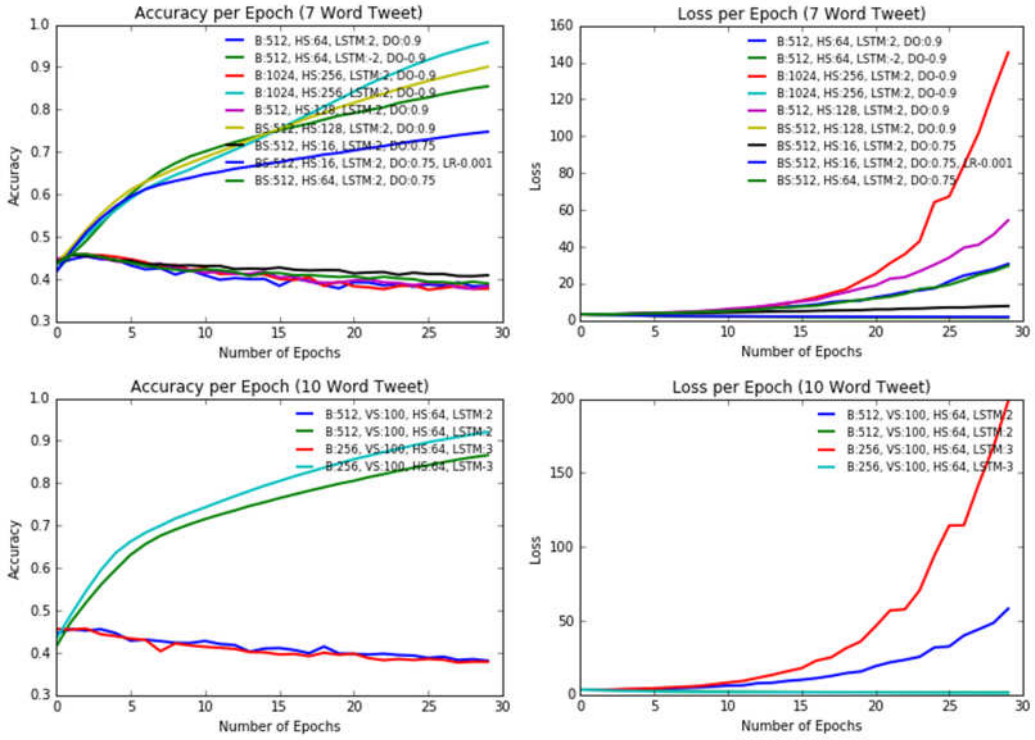
To begin the experiment, the Eisenstein corpus is randomized then divided into train, test, and dev subsets according to a 60%, 20%, 20% data split. To conduct the experiments we explored the data and neural network structure across several dimensions to determine the model and dataset attributes that result in the best outcome. The figure below summarizes the experimental variations.

Category	Values
Dataset	
Tweet Size	7 & 10 word
GloVe Vector Size	25, 50, 100, 200
Model Structure	
LSTM Layer Count	2 & 3
Hidden Layer Size	64, 128, 256
Hyper-parameter	
Dropout	0.75 - 1.0
Learning Rate	0.001 - 0.1
L2 Regularization	0 - 0.01
Batch Size	64, 128, 256, 512, 1024

To measure the experiment outcomes, we define accuracy as the proportion of tweets assigned to the correct 4-way regional classifier out of the total number of tweet assignments attempted.

The experiment results indicate that the 10 word tweets are more predictive than the 7 word tweets, although not substantially. GloVe word vectors of size 100 are just as predictive as the size 200 word vectors but are computationally less expensive and preferable for experimentation. Adjusting the hidden layer size between 64, 128 and 256 did improve accuracy with 10 word tweets.

Findings



Our biggest finding takeaway is that the accuracy on the test data set does not rise above 47% regardless of data, model, or hyper parameter adjustments. We found baseline RNN approaches with a single LSTM layer provide significant predictive power on the Eisenstein corpus and are sufficient to exceed the 41% 4-way regional classification accuracy implemented in Eisenstein, et al. Although multi-layer LSTM models improve predictive power, they do not exceed the 67% 4-way regional accuracy levels achieved by Cha, et. al.

One possible explanation for these findings is the very high dimensional tweet vocabulary, yet limited tweet word count (implied by the character limit). This seems to limit the full predictive power of the LSTMs. Another possible explanation is the data processing that standardizes the punctuation and unknown words to the extent that the variability is lost. Although, the RNN with LSTMs is able to achieve high accuracy results on the training set, these results are likely due to overfitting and do not generalize well to the test set.

Human Classification Performance

Correctly classifying tweets into 4-way regional categories is difficult. Using a random sample of 100 tweets, two humans were able to achieve an accuracy of 31%, slightly above random classification. This may have been due to the small size or regional skewness in the dataset. Interestingly, tweets with increased slang and poor language tended to be from the Northeast region.

Correctly classified tweets	Incorrectly classified tweets
The first episode of everybody hates chris is dead on	Not feelin today at all
No Suffolk county	Off the early morning though lol
The hoodrats with no food stamps	This month sucks already
I miss my tenth grade year	Thank you have fun snowboarding

Corpus Issues

Tweets within the Eisenstein corpus are distributed unevenly across regions with almost 80% of the dataset representing the Northeast and South regions –likely reducing 4-way regional classification accuracy.

Region	Count	Percent
Midwest Region	43,170	11.71%
Northeast Region	139,132	37.73%
South Region	140,291	38.05%
West Region	46,157	12.52%
	368,750	100.00%

Conclusion

This experiment expands on several existing approaches leveraging the Eisenstein dataset to geolocate Twitter users based solely on tweet content. Of the three major previous studies, only Liu, et al. leverage neural network approaches. Although this experiment was not successful in exceeding the accuracy levels achieved in the other studies, we attempted a new approach with RNN and LSTM layers that can hopefully be expanded upon in the future.

Future directions

Expand to additional datasets – This experiment can be expanded to include larger datasets including geocoded data gathered through the Twitter API, or the Roller Twitter dataset mentioned in Liu 2015. An interesting expansion of this experiment would be classification of foreign language tweets.

Additional model structure and parameter tuning – Further development of the neural network including enhancement of the LSTM layers to incorporate peepholes and improved initialization schemes may be promising. Additional parameter tuning will likely improve model performance.

References

Cha, Guang, Kung, Geolocation with Subsampled Microblog Social Media

Cha, Guang, Kung, Twitter Geolocation and Regional Classification via Sparse Coding

Collobert, et al., Natural Language Processing (almost) from Scratch

Backstrom, Sun, Marlow, Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity

Bergsma, Dredze, Van Durme, Wilson, Yarowsky, Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter

Cheng, Caverlee, Kyumin, You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users

Eisenstein, O'Connor, Smith, Xing, A Latent Variable Model for Geographic Lexical Variation

Han, Cook, Baldwin, Text-Based Twitter User Geolocation Prediction

Hochreiter, et al., Long Short Term MemoryLONG SHORT-TERM MEMORY

Jozefowicz, An Empirical Exploration of Recurrent Network Architectures

Liu, Inkpen, Estimating User Location in Social Media with Stacked Denoising Auto-encoders

Mahmud, Nichols, Drews, Home Location Identification of Twitter Users

Nand, Perera, Sreekumar, Lingmin, A Multi-Strategy Approach for Location Mining in Tweets

Sixto, Pena, Klein, Lopez-de-Ipina, Enable tweet-geolocation and don't drive ERTs crazy!

Improving situational awareness using Twitter