Binxuan Huang
binxuanh@cs.cmu.edu

Kathleen M. Carley
kathleen.carley@cs.cmu.edu

Carnegie Mellon University
www.casos.cs.cmu.edu

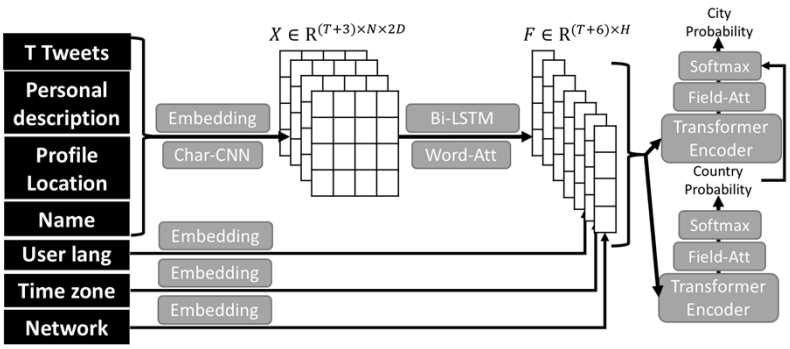
A Hierarchical Location Prediction Neural Network for Twitter User Geolocation

Introduction

Accurate estimation of user location is important for many online services. Previous neural network based methods largely ignore the hierarchical structure among locations. In this paper, we propose a hierarchical location prediction neural network for Twitter user geolocation. It first predicts the home country for a user, then uses the country result to guide the city-level prediction. Our model can accommodate various feature combinations and achieves state-of-the-art results over three commonly used benchmarks under different feature settings. It not only improves the prediction accuracy but also greatly reduces the mean error distance.

Method

- we first use word embedding and character-level CNN model to represent text as word vectors. Then we apply Bi-LSTM to learn semantic representations for these texts, which are further fed into word-level attention module to learn one representation vector for each text.
- Combining text representations, embeddings for user language, time zone, and network, we get T+6 representation vectors. Two transformer encoders are used to learn the correlation between these vectors and generates classification features for country and city. The country prediction results are further used to constrain the city-level prediction.



Experiments

- We validate our method on three widely used datasets, Twitter-US, Twitter-World, WNUT, with four different feature settings, text, text+metadata, text+network, text+metadata+network.
- Only using text feature from tweets, our model HLPNN-Text works the best among all these text-based location prediction systems and wins by a large margin. It not only improves prediction accuracy but also greatly reduces mean error distance.
- With text and metadata, HLPNN-Meta correctly predicts locations of 57.2% users in WNUT dataset, which is even better than these location prediction systems that use text, metadata, and network.
- Adding network feature further improves our model's performances. It achieves state-of-the-art results combining all features on these three datasets. On Twitter-US dataset, our model variant HLPNN-Net achieves a 4.6% increase in Acc@161 against previous state-of-the-art methods (Do et al., 2017) and (Rahimi et al., 2018). The prediction accuracy of HLPNN-Net on WNUT dataset is similar to (Miura et al., 2017), but with a noticeable lower mean error distance.

	Twitter-US			Twitter-World			WNUT			
	Acc@161↑	Median↓	Mean↓	Acc@161↑	Median↓	Mean↓	Accuracy↑	Acc@161↑	Median↓	Mean↓
Text										
Wing and Baldridge (2014)	49.2	170.5	703.6	32.7	490.0	1714.6	-	-	-	-
Rahimi et al. (2015b)*	50	159	686	32	530	1724	-	-	-	-
Miura et al. (2017)-TEXT	55.6	110.5	585.1	-	-	-	35.4	50.3	155.8	1592.6
Rahimi et al. (2017)	55	91	581	36	373	1417	-	-	-	-
HLPNN-Text	**57.1**	**89.92**	**516.6**	**40.1**	**299.1**	**1048.1**	**37.3**	**52.9**	**109.3**	**1289.4**
Text+Meta										
Miura et al. (2017)-META	**67.2**	**46.8**	**356.3**	-	-	-	54.7	70.2	0	825.8
HLPNN-Meta	61.1	64.3	454.8	**56.4**	**86.2**	**762.1**	**57.2**	**73.1**	**0**	**572.5**
Text+Net										
Rahimi et al. (2015a)*	60	78	529	53	111	1403	-	-	-	-
Rahimi et al. (2017)	61	77	515	53	104	1280	-	-	-	-
Miura et al. (2017)-UNET	61.5	65	481.5	-	-	-	**38.1**	**53.3**	**99.9**	1498.6
Do et al. (2017)	66.2	45	433	53.3	118	1044	-	-	-	-
Rahimi et al. (2018)-MLP-TXT+NET	66	56	420	58	**53**	1030	-	-	-	-
Rahimi et al. (2018)-GCN	62	71	485	54	108	1130	-	-	-	-
HLPNN-Net	**70.8**	**31.6**	**361.5**	**58.9**	59.9	**827.6**	37.8	**53.3**	105.26	**1297.7**
Text+Meta+Net										
Miura et al. (2016)	-	-	-	-	-	-	47.6	-	16.1	1122.3
Jayasinghe et al. (2016)	-	-	-	-	-	-	52.6	-	21.7	1928.8
Miura et al. (2017)	70.1	41.9	335.7	-	-	-	56.4	71.9	**0**	780.5
HLPNN	**72.7**	**28.2**	**323.1**	**68.4**	**6.20**	**610.0**	**57.6**	**73.4**	**0**	**538.8**

Table 3: Comparisons between our method and baselines. We report results under four different feature settings: Text, Text+Metadata, Text+Network, Text+Metadata+Network. "-" signifies that no results were published for the given dataset, "*" denotes that results are cited from (Rahimi et al., 2017). Note that Miura et al. (2017) only used 279K users added with metadata in their experiments of Twitter-US.

Conclusion

In this paper, we propose a hierarchical location prediction neural network, which combines text, metadata, network information for user location prediction. Our model can accommodate various feature combinations. Extensive experiments have been conducted to validate the effectiveness of our model under four different feature settings across three commonly used benchmarks.

This work was supported in part by the Office of Navel Research/UALR through, Grant ONR N00014182106 and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the Office of Navel Research/UALR or the U.S. government.

isr institute for SOFTWARE RESEARCH