IC2S2 '24
7/19/2024

Where on Earth Do Users Say They Are?: Geo-Entity Linking for Noisy User Input

Tessa Masis
they/them

Brendan O'Connor
he/him

University of Massachusetts Amherst, USA

Using social media users' locations

Location reference identification from tweets during emergencies: A deep learning approach

Abhinav Kumar*, Jyoti Prakash Singh

Department of Computer Science & Engineering, National Institute of Technology Patna, India

Real-Time Disease Surveillance Using Twitter Data: Demonstration on Flu and Cancer

Kathy Lee Ankit Agrawal Alok Choudhary
EECS Department
rthwestern University
Evanston, IL USA
houdhar}@eecs.northwestern.edu

Understanding U.S. regional linguistic variation with Twitter data analysis

Yuan Huang [a], Diansheng Guo [a,*], Alice Kasakoff [a], Jack Grieve [b]

[a] *Department of Geography, University of*
[b] *School of Languages and Social Sciences,*

Interregional and intraregional variability of intergroup attitudes predict online hostility

HANNES ROSENBUSCH*, ANTHONY M. EVANS and MARCEL ZEELENBERG

Department of Social Psychology, Tilburg University, Tilburg, The Netherlands

Noisy location references on social media

Գ♗Թ
@YerevanKnights

In the aftermath, we are because they were.

Կուկուռուզնիկ Joined Ja

Ana C is staying whimsical
@_ana_c

Toda orquestra é uma banda cover. Especialista em tendências do mercado etílico e de luxo, WSET 3. Better living through k-pop. (MULTISTAN, Ela/Dela)
Translate bio

SP-BH Joined July 2009

brendan o'connor
@brendan642

Faculty @UMassCS | Natural language processing and computational social science | brenocon.com | @brenocon.bsky.social

western mass (northampton) brenocon.com Joined June 2008

Geo-Entity Linking

Geo-Entity Linking

Use of the Edinburgh geoparser for georeferencing digitized historical collections

By Claire Grover[1,*], Richard Tobin[1], Kate Byrne[1], Matthew Woollard[2], James Reid[3], Stuart Dunn[4] and Julian Ball[5]

[1] *School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK*
[2] *UK Data Archive, University of Essex, Colchester CO4 3SQ, UK*
[3] *EDINA, 160 Causewayside, Edinburgh EH9 1PR, UK*
[4] *Centre for e-Research, King's College London, Strand, London WC2R 2LS, UK*
[5] *Hartley Library, University of Southampton, Southampton SO17 1BJ, UK*

CLIFF-CLAVIN: Determining Geographic Focus for News Articles

[Extended Abstract]

Catherine D'Ignazio
MIT Center for Civic Media
77 Massachusetts Avenue
Cambridge, MA 02139, USA
dignazio@mit.edu

Rahul Bhargava
MIT Center for Civic Media
77 Massachusetts Avenue
Cambridge, MA 02139, USA
rahulb@media.mit.edu

Ethan Zuckerman
MIT Center for Civic Media
77 Massachusetts Avenue
Cambridge, MA 02139, USA
ethanz@media.mit.edu

Luisa Beck
Independent
Emmi.Beck@gmail.com

GeoTxt: A Web API to Leverage Place References in Text

Morteza Karimzadeh[1,2], Wenyi Huang[3], Siddhartha Banerjee[1,3], Jan Oliver Wallgrün[1,2], Frank Hardisty[1,2], Scott Pezanowski[1,2], Prasenjit Mitra[1,3] and Alan M. MacEachren[1,2]

1) GeoVISTA Center,
302 Walker Building,
University Park, PA, 16802
+1-814-865-3433
{karimzadeh, wallgrun, hardisty}@psu.edu

2) Department of Geography,
The Pennsylvania State University,
302 Walker Building,
University Park, PA 16802
+1-814-865-3433
{scottpez,maceachren}@psu.edu

3) College of Information Sciences and Technology, 332 Info Science and Tech University Park,
PA 16802, USA
+1-814-865-3528
{wzh112,sub253,pmitra}@ist.psu.edu

Changes in Tweet Geolocation over Time: A Study with Carmen 2.0

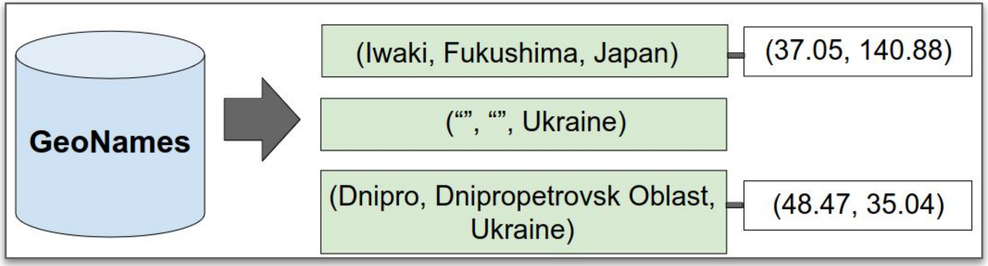
Jingyu Zhang and **Alexandra DeLucia** and **Mark Dredze**
Department of Computer Science
Johns Hopkins University
{jzhan237, aadelucia, mdredze}@jhu.edu

Data

Target location database: GeoNames

28,767 distinct locations

Cities are labeled with coordinates



Labeled geo-entity linking dataset: Twitter-Global

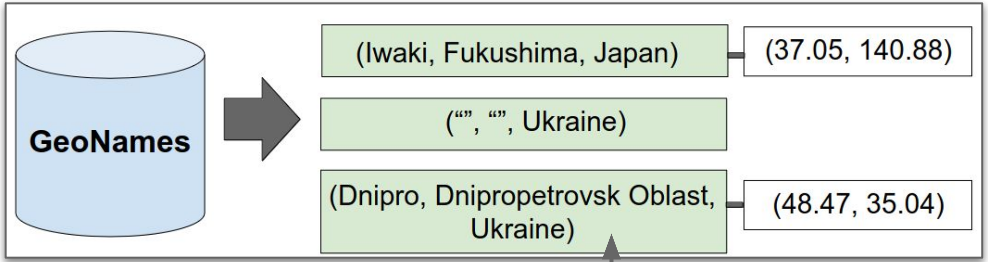
4.1M geocoordinate-tagged tweets

Data

Target location database: GeoNames

28,767 distinct locations

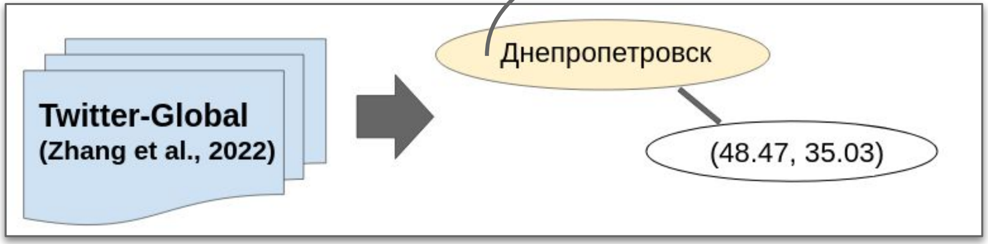
Cities are labeled with coordinates



Labeled geo-entity linking dataset: Twitter-Global

4.1M geocoordinate-tagged tweets

We link each poster's Location field to a ground truth location = the closest city in GeoNames database

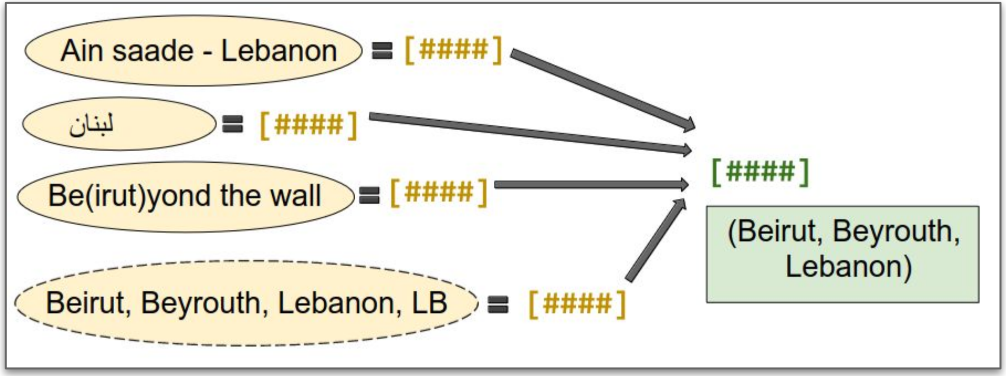


Proposed Method: UserGeo

1) **Training: For each target location in GeoNames, create a soft-alias location name representation by averaging SBERT embeddings of all linked Location fields in Twitter-Global**

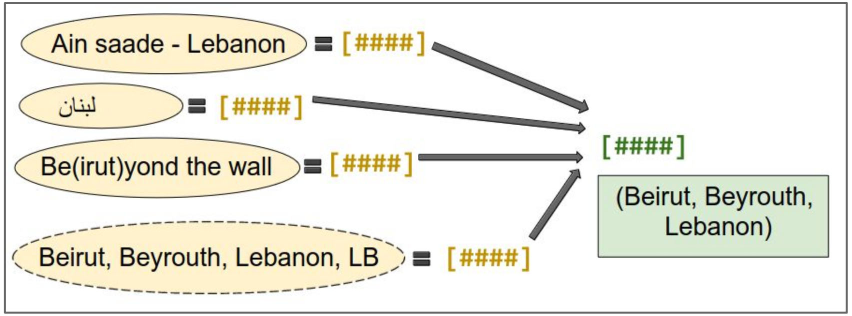
2) Predicting: For a new free text location mention, predict the location with the highest cosine similarity

3) If cosine similarity is below a certain threshold, make no guess i.e. `NULL`

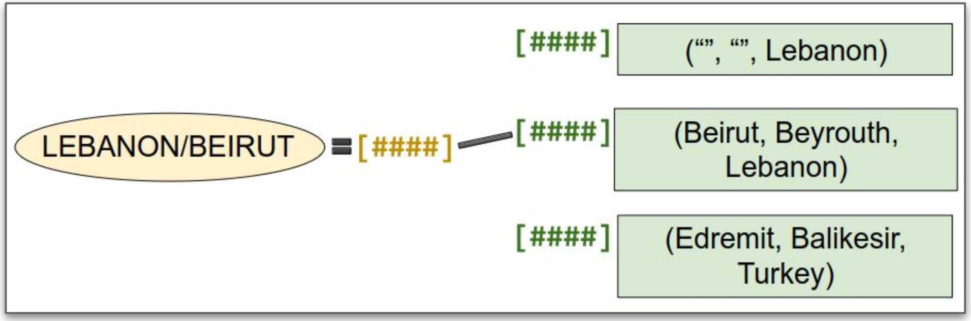


Proposed Method: UserGeo

1) Training: For each target location in GeoNames, create a soft-alias location name representation by averaging SBERT embeddings of all linked Location fields in Twitter-Global



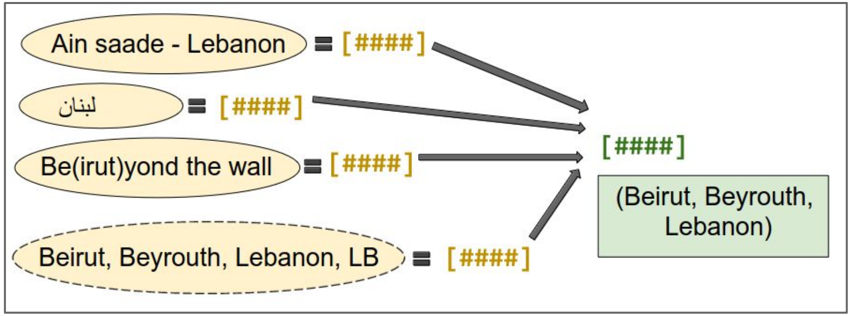
2) **Predicting: For a new free text location mention, predict the location with the highest cosine similarity**



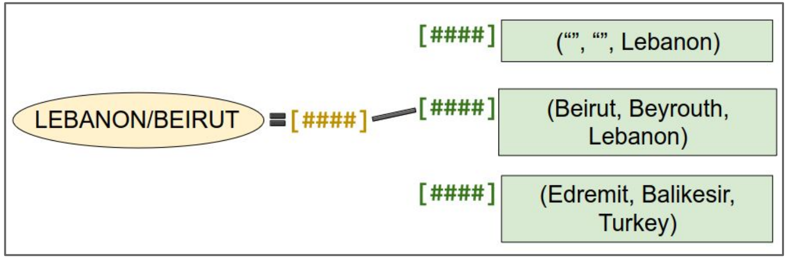
3) If cosine similarity is below a certain threshold, make no guess i.e. NULL

Proposed Method: UserGeo

1) Training: For each target location in GeoNames, create a soft-alias location name representation by averaging SBERT embeddings of all linked Location fields in Twitter-Global



2) Predicting: For a new free text location mention, predict the location with the highest cosine similarity



3) **If cosine similarity is below a certain threshold, make no guess i.e. NULL**

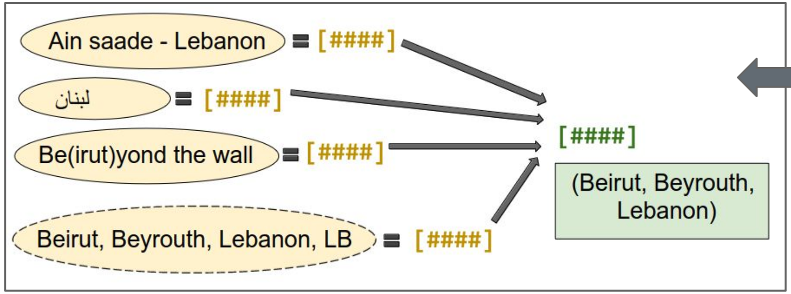
Results: Accuracy

Method	Country	Admin.	City

Results: Accuracy

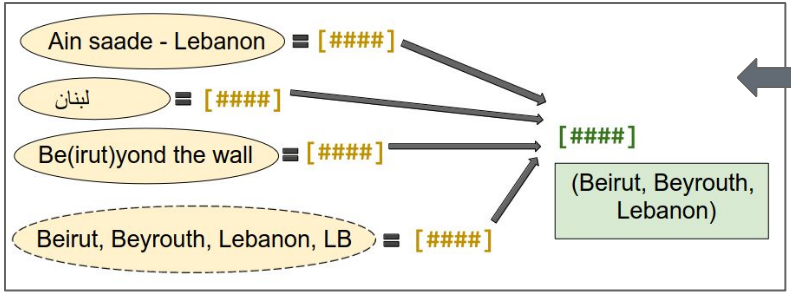
Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8

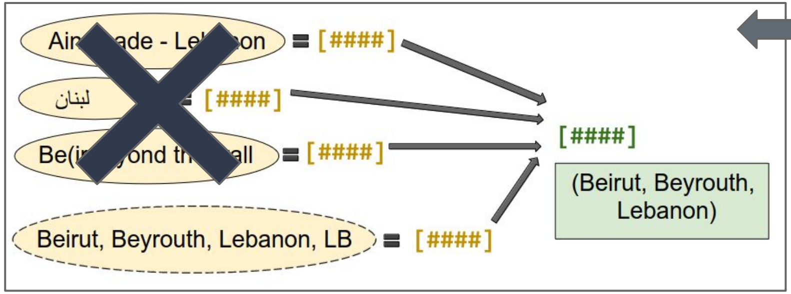
Results: Accuracy



Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
UserGeo	**67.8**	**44.2**	14.8

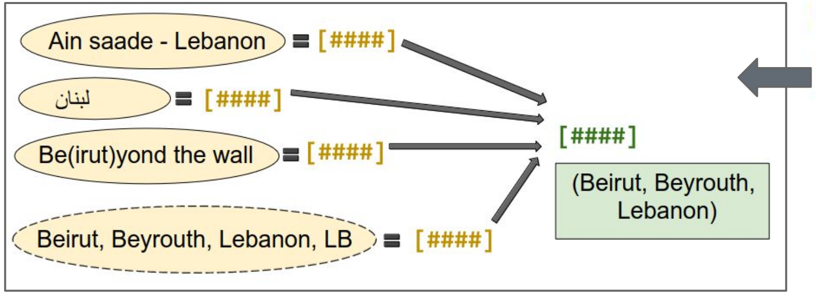
Results: Accuracy





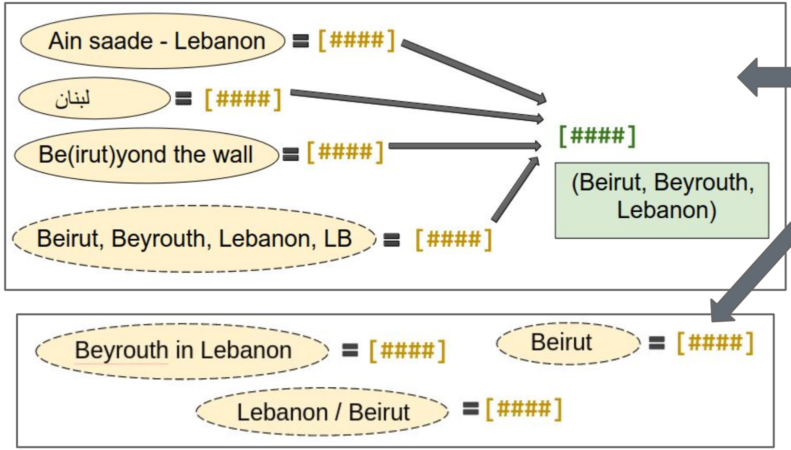
Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
UserGeo	**67.8**	**44.2**	14.8
NameGeo	59.8	37.7	14.3

Results: Accuracy



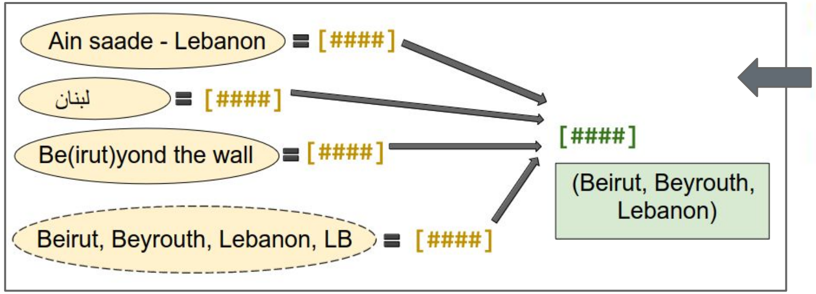
Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
UserGeo	**67.8**	**44.2**	14.8
NameGeo	59.8	37.7	14.3

Results: Accuracy



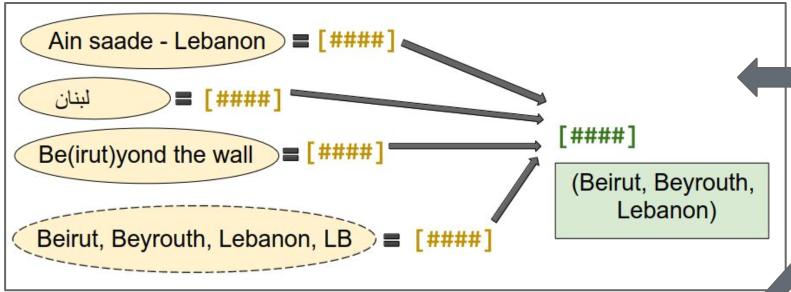
Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
UserGeo	**67.8**	**44.2**	14.8
+variants	66.0	43.7	15.3
NameGeo	59.8	37.7	14.3
+variants	62.0	40.9	**17.0**

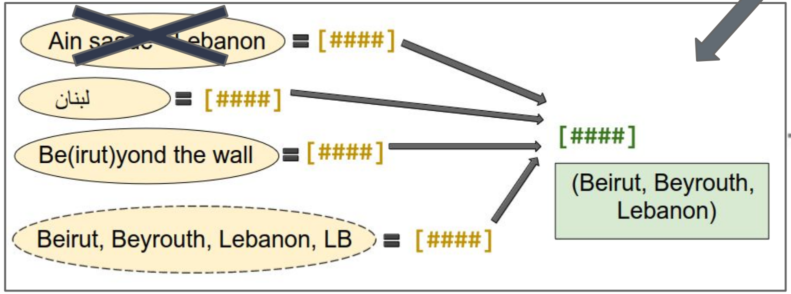
Results: Accuracy



Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
UserGeo	**67.8**	**44.2**	14.8
+variants	66.0	43.7	15.3
NameGeo	59.8	37.7	14.3
+variants	62.0	40.9	**17.0**

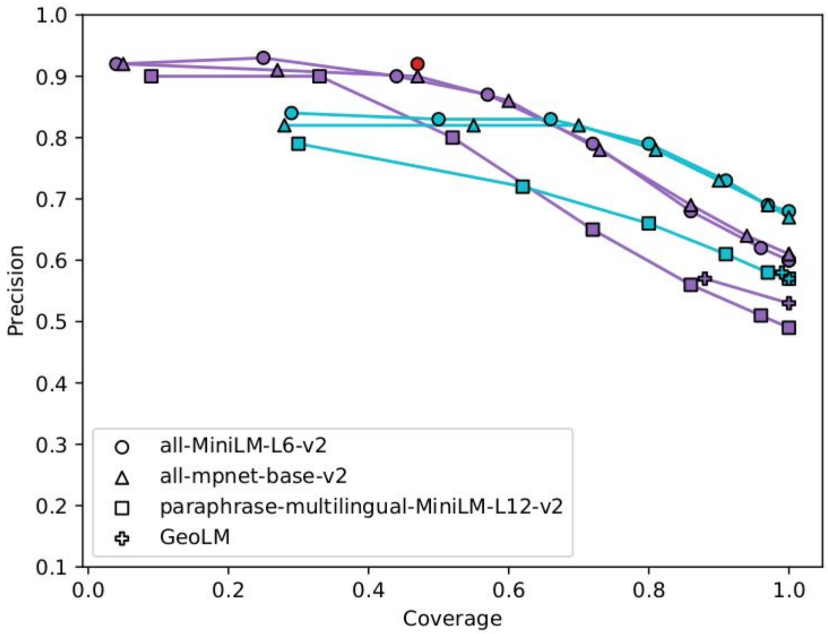
Results: Accuracy



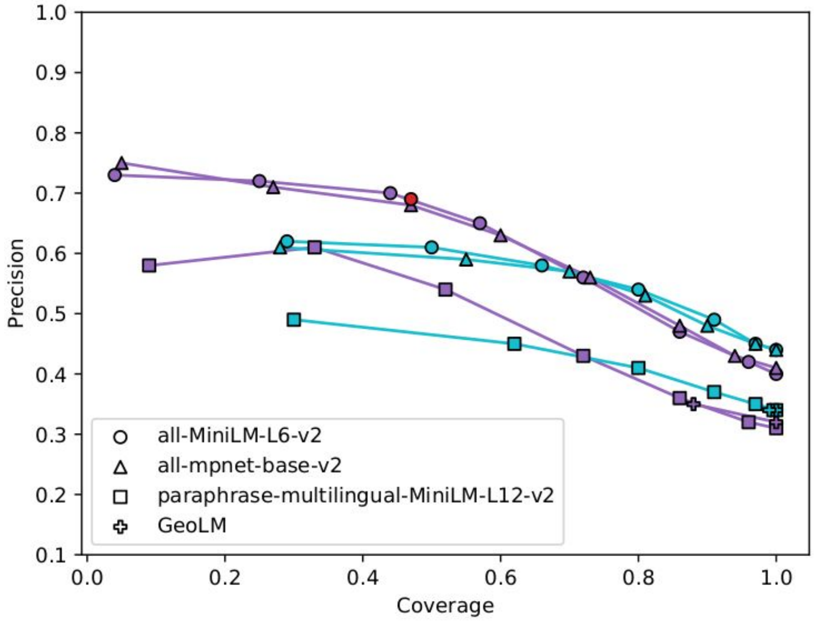


Method	Country	Admin.	City
Carmen 2.0	43.5	27.3	9.8
UserGeo	**67.8**	**44.2**	14.8
+variants	66.0	43.7	15.3
+pruning	63.5	41.4	13.2
+variants+pruning	65.2	43.4	13.9
NameGeo	59.8	37.7	14.3
+variants	62.0	40.9	**17.0**

Results: Precision–Coverage Curves



(a) Country



(b) Admin.

Results: Qualitative Analysis

Location field
TURKEY/SİNOP
福島県いわき市
Catskills
where the wild things are

Results: Qualitative Analysis

Location field	GPS location
TURKEY/SİNOP	Sinop, Sinop, Turkey
福島県いわき市	Iwaki, Fukushima, Japan
Catskills	Hyde Park, New York, United States
where the wild things are	La Vista, Nebraska, United States

Results: Qualitative Analysis

Location field	GPS location	Carmen 2.0 prediction
TURKEY/SİNOP	Sinop, Sinop, Turkey	`NULL`
福島県いわき市	Iwaki, Fukushima, Japan	`NULL`
Catskills	Hyde Park, New York, United States	`NULL`
where the wild things are	La Vista, Nebraska, United States	`NULL`

Results: Qualitative Analysis

Location field	GPS location	Carmen 2.0 prediction	NameGeo prediction
TURKEY/SİNOP	Sinop, Sinop, Turkey	`NULL`	“”, Sinop, Turkey
福島県いわき市	Iwaki, Fukushima, Japan	`NULL`	Zhongshu, Yunnan, China
Catskills	Hyde Park, New York, United States	`NULL`	Catalca, Istanbul, Turkey
where the wild things are	La Vista, Nebraska, United States	`NULL`	`NULL`

Results: Qualitative Analysis

Location field	GPS location	Carmen 2.0 prediction	NameGeo prediction	UserGeo prediction
TURKEY/SİNOP	Sinop, Sinop, Turkey	`NULL`	“”, Sinop, Turkey	Boyabat, Sinop, Turkey
福島県いわき市	Iwaki, Fukushima, Japan	`NULL`	Zhongshu, Yunnan, China	Iwaki, Fukushima, Japan
Catskills	Hyde Park, New York, United States	`NULL`	Catalca, Istanbul, Turkey	Hyde Park, New York, United States
where the wild things are	La Vista, Nebraska, United States	`NULL`	`NULL`	`NULL`

How many have a real location?

Location field	**GPS location**
TURKEY/SİNOP	Sinop, Sinop, Turkey
福島県いわき市	Iwaki, Fukushima, Japan
Catskills	Hyde Park, New York, United States
where the wild things are	La Vista, Nebraska, United States

How many have a real location?

Location field	GPS location
TURKEY/SİNOP	Sinop, Sinop, Turkey
福島県いわき市	Iwaki, Fukushima, Japan
Catskills	Hyde Park, New York, United States
where the wild things are	La Vista, Nebraska, United States

	Country	Admin.	City
Upper bound	72.5	58.3	49.2

Conducted a manual examination of the proportion of Location fields that reference an actual location; we use this as an upper bound of accuracy

How many have a real location?

Location field	GPS location
TURKEY/SİNOP	Sinop, Sinop, Turkey
福島県いわき市	Iwaki, Fukushima, Japan
Catskills	Hyde Park, New York, United States
where the wild things are	La Vista, Nebraska, United States

	Country	Admin.	City
Upper bound	72.5	58.3	49.2
NameGeo+variants	62.0	40.9	**17.0**
UserGeo	**67.8**	**44.2**	14.8

Conducted a manual examination of the proportion of Location fields that reference an actual location; we use this as an upper bound of accuracy

Summary

- Proposed methods for geo-entity linking noisy & multilingual social media data with selective prediction
- Of two best performing methods, UserGeo achieves SOTA performance at country and administrative levels while NameGeo+variants doesn't require training data
- Identified problems with geo-entity linking at the city level for social media data
- In future, plan to release our models and to extend to broader task of geoparsing unstructured text

Thank you!

This work was recently published at NLP+CSS Workshop at NAACL 2024!

Slides, abstract, and paper available at tmasis.github.io/

Tessa Masis

tmasis@cs.umass.edu

Brendan O'Connor

brenocon@cs.umass.edu

Results: Accuracy

Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
all-MiniLM-L6-v2			
NameGeo	59.8	37.7	14.3
UserGeo	**67.8**	**44.2**	14.8
all-mpnet-base-v2			
NameGeo	60.9	38.3	**14.9**
UserGeo	67.4	43.7	13.9
paraphrase-multilingual-MiniLM-L12-v2			
NameGeo	48.7	28.9	8.1
UserGeo	57.0	34.3	9.4
GEOLM			
NameGeo	52.5	30.5	12.1
UserGeo	57.4	33.9	10.7