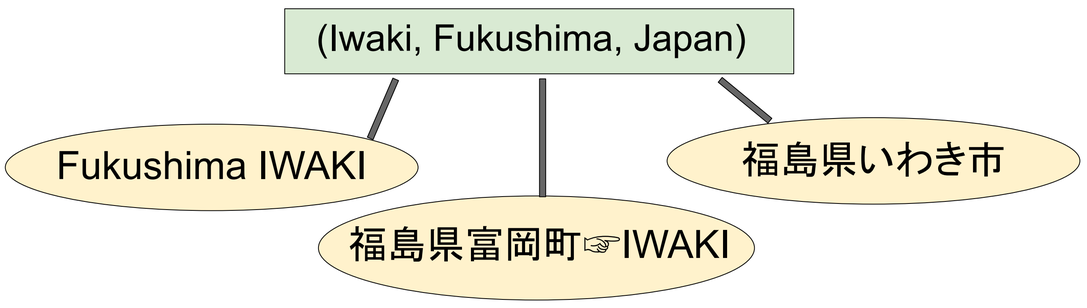
Where on Earth Do Users Say They Are?: Geo-Entity Linking for Noisy Multilingual User Input

Tessa Masis (tmasis.github.io/) and Brendan O'Connor

University of Massachusetts Amherst

Geo-Entity Linking

Locations can be referred to on social media with varying writing styles, informal names, and in different languages or writing scripts

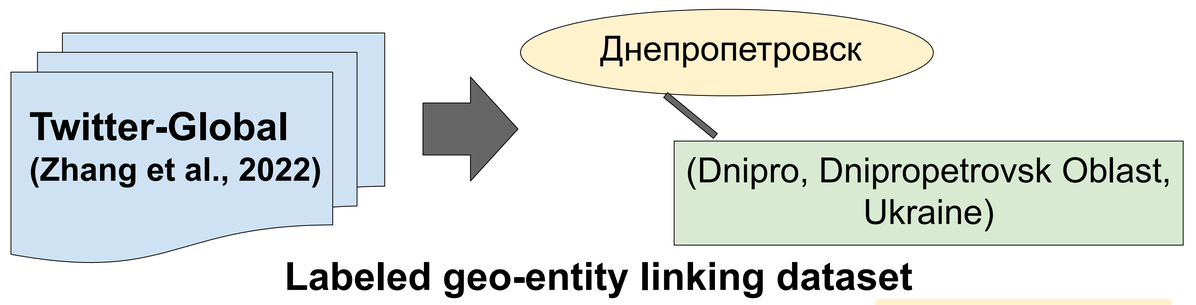


<u>Goal:</u> given a noisy free text Location field mention, link it to the correct real-world geographic entity

Data

Target location database

28,767 distinct locations; cities labeled with coordinates

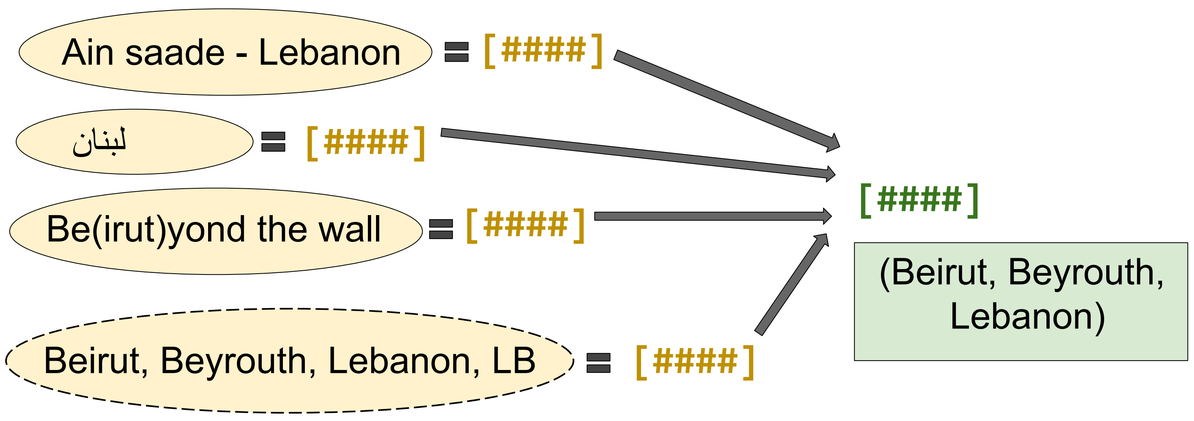


Labeled geo-entity linking dataset

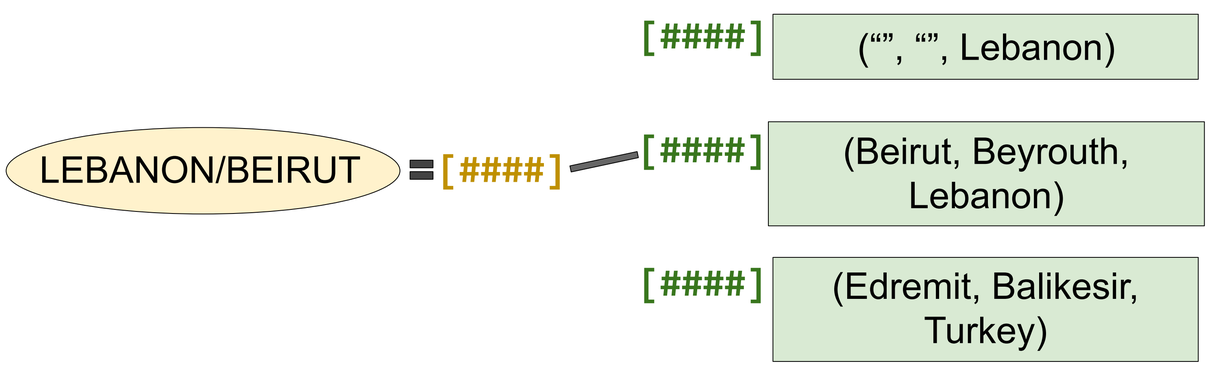
4.1M geocoordinate-tagged tweets; we link each poster's Location field to a ground truth location, defined as the closest city in GeoNames database

Proposed Method: UserGeo

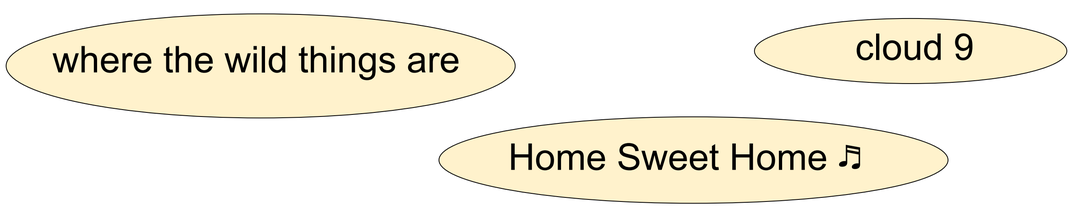
1) Training: For each location in GeoNames database, create a soft-alias location name representation by averaging SBERT embeddings of all linked Location fields in Twitter-Global



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

Qualitative Analysis

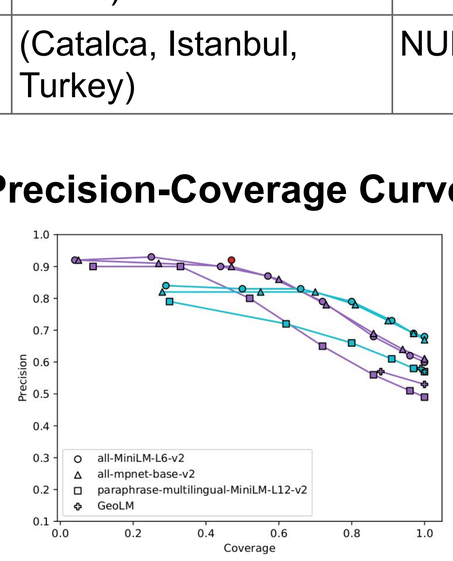
Location field	UserGeo	NameGeo	Carmen
福島県いわき市	(Iwaki, Fukushima, Japan)	(Zhongshu, Yunnan, China)	NULL
Catskills	(Hyde Park, New York, US)	(Catalca, Istanbul, Turkey)	NULL

Accuracy

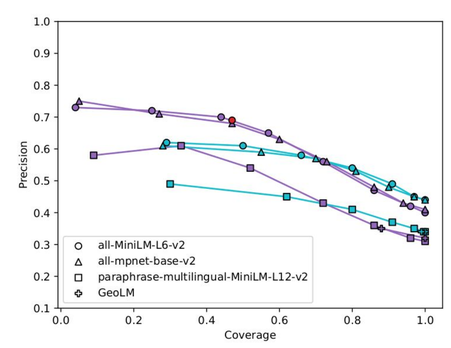
Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
all-MiniLM-L6-v2			
NameGeo	59.8	37.7	14.3
UserGeo	**<u>67.8</u>**	**<u>44.2</u>**	<u>14.8</u>
all-mpnet-base-v2			
NameGeo	60.9	38.3	**<u>14.9</u>**
UserGeo	<u>67.4</u>	<u>43.7</u>	13.9
paraphrase-multilingual-MiniLM-L12-v2			
NameGeo	48.7	28.9	8.1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

Method	Country	Admin.	City
NameGeo	59.8	37.7	14.3
+variants	62.0	40.9	**<u>17.0</u>**
UserGeo	**<u>67.8</u>**	**<u>44.2</u>**	14.8
+pruning	63.5	41.4	13.2
+variants	<u>66.0</u>	<u>43.7</u>	<u>15.3</u>
+variants+pruning	65.2	43.4	13.9

Precision-Coverage Curves



(a) Country



(b) Admin.

How Many Have a Real Location?

Manual examination of proportion of Location fields that reference an actual location; we use this as an <u>upper bound of accuracy</u>

	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

Summary & Future Work

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