

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

Tessa Masis (*they/them*)    Brendan O'Connor  
University of Massachusetts Amherst, MA, USA  
{tmasis, brenocon}@cs.umass.edu

## Abstract

Geo-entity linking is the task of linking a location mention to the real-world geographic location. In this paper we explore the challenging task of geo-entity linking for noisy, multilingual social media data. There are few open-source multilingual geo-entity linking tools available and existing ones are often rule-based, which break easily in social media settings, or LLM-based, which are too expensive for large-scale datasets. We present a method which represents real-world locations as averaged embeddings from labeled user-input location names and allows for selective prediction via an interpretable confidence score. We show that our approach improves geo-entity linking on a global and multilingual social media dataset, and discuss progress and problems with evaluating at different geographic granularities.

## 1 Introduction

The real-world geographic location of social media users is valuable data for many computational social science tasks, including disaster response (Kumar and Singh, 2019), disease surveillance (Lee et al., 2013), analyzing language variation (Huang et al., 2016), and comparing regional attitudes (Rosenbusch et al., 2020). Many studies have used Twitter (now known as X) data for such analyses, focusing on geo-tagged tweets where each tweet is associated with latitude and longitude coordinates. However, geo-tagging with coordinates was deprecated in June 2019 and even before then only a small percentage of tweets ( $< 2\%$ ) was geo-tagged (Kruspe et al., 2021).

It has thus become increasingly necessary to infer location from user profiles and especially from the free text Location field, in which a user may enter anything they want to identify their location. This field is frequently specified, with at least 40% of users providing recognizable locations in over 60 different languages (Huang and Carley, 2019b).

The task of linking a location reference to the actual geographic location is known as *geo-entity linking* (see Table 1 for examples). There are few open-source multilingual geo-entity linking tools available and existing ones are often rule-based (Alex et al., 2016; Dredze et al., 2013), which may break easily in noisy social media settings, or LLM-based (Zhang et al., 2023), which are too expensive for large-scale datasets.

In this paper, we investigate the task of geo-entity linking for noisy, multilingual user-input location references. Our work makes the following contributions:

- We propose a method for geo-entity linking of noisy and multilingual user input by representing real-world locations with averaged embeddings from labeled user-input location names. Unlike previous methods, ours enables selective prediction via an adjustable threshold for cosine similarity scores, which we analogize with confidence scores (§4).
- We compare performance of multiple variations of our proposed method on a global and multilingual dataset, and show that all of them outperform the leading baseline (§5).
- Through a manual annotation experiment, we approximate accuracy upper bounds on our dataset and show that our method is near the upper bound at country- and administrative-levels but quite far below at the city-level. We discuss problems with geo-entity linking social media data at the city level (§6).

## 2 Related Work

**Geo-entity linking**, also known as toponym resolution, seeks to link some mention of a geographic entity to the correct entity in a target database. Previous approaches typically use some combination of the mention’s text and/or context; knowledge

User-input location	Real-world location	Type of noise
TURKEY/SINOP	Sinop, Sinop, TR	Uncommon punctuation use
福島県いわき市	Iwaki, Fukushima, JP	Non-Latin script
Catskills	Hyde Park, New York, US	Informal/alternative name
where the wild things are	N/A	Not a real location

Table 1: Examples of user-input location references, the real-world locations they should be linked with, and the type of noise that the geo-entity linking model must be able to handle.

bases (e.g. gazetteer, Wikipedia) which contain features such as population, location type, etc.; and coordinates/geometric features. They may use some mix of rule-based, unsupervised, and supervised methods. The majority of prior work on geo-entity linking has focused on data in the English language and in the domain of news articles (Lieberman and Samet, 2012; Speriosu and Baldrige, 2013; Kamaloo and Rafiei, 2018; Cardoso et al., 2019; Kulkarni et al., 2021; Cardoso et al., 2022; Sá et al., 2022; Li et al., 2022; Sharma et al., 2023; Zhang and Bethard, 2023). We note that the geospatially grounded model GeoLM – which was trained only on English data – was evaluated on geo-entity linking in a way similar to our proposed method, by ranking locations by cosine similarity between each candidate and the query location (Li et al., 2023). This is similar to our proposed NameGeo method, although the authors did not explore using cosine similarity thresholds for selective prediction or any of the other variations that we investigate.

Some prior work has examined geo-entity linking in historical texts (Smith and Crane, 2001; Ardanuy and Sporleder, 2017; Ardanuy et al., 2020), which includes English, Spanish, Dutch, and German data; and in web pages (Moncla et al., 2014), which includes French, Spanish, and Italian data.

Most relevant to the current work are previous studies which have examined geo-entity linking in social media data. Alex et al. (2016) uses a rule-based English system, Dredze et al. (2013) uses a multilingual rule-based system, and Zhang et al. (2023) uses a multilingual LLM-based system.

**Entity-linking** is the broader task of linking some mention of an entity – which could be a person, place, or organization – to the correct entity in a target database. We note a similarity between our proposed UserGeo method and one introduced for entity-linking in FitzGerald et al. (2021), in that they both represent entities using all mentions in the training data. However, our method is simpler as it does not involve model training. Addition-

ally, UserGeo represents an entity by averaging mention embeddings instead of having a separate embedding for each mention, which induces a more holistic entity representation that can better handle noisy mentions present in social media data.

**Predicting user location**, also known as user geolocation, is a task distinct from geo-entity linking social media data in that it seeks to determine the location of a user using both text data and user metadata, including post content, user bio, user language, time zone, or social networks (Han et al., 2012; Jurgens et al., 2015; Rahimi et al., 2015; Huang et al., 2016; Rahimi et al., 2017; Izbicki et al., 2019; Huang and Carley, 2019a; Luo et al., 2020). The Location field is one of many features used for location prediction, if used at all.

**Geoparsing** is the task of both identifying and linking geographic entities in unstructured text (Wang and Hu, 2019), essentially combining toponym recognition (Hu et al., 2023) and geo-entity linking. Geoparsers focused on standard English texts (including news articles, Wikipedia, or scientific papers) include CLIFF-CLAVIN (D’Ignazio et al., 2014), TopoCluster (DeLozier et al., 2015), CamCoder (Gritta et al., 2018), and DM\_NLP (Wang et al., 2019). Other geoparsers include the Edinburgh geoparser for historical English text (Grover et al., 2010), GeoTxt for English social media data (Karimzadeh et al., 2013, 2019), and Perdido for French texts (Moncla et al., 2014).

### 3 Task and Data

#### 3.1 Geo-entity linking task

Given a target location database  $D$ , a training set  $T$  containing user-input location name and ground truth location pairs, and a test set of user-input location names  $I$ , for each  $i \in I$  we model

$$\arg \max_{d \in D} \text{score}(d, T, i)$$

to predict best matching geographic entity  $d$ ,

which is represented by a triple containing a city name, primary administrative region name (e.g. state, province), and country name. (The city and administrative region names may be empty strings if the entity is of a higher granularity, e.g. a country) The `score()` function evaluates the quality of  $d$  as a match to user input  $i$ , given training data  $T$ .

We note that a user-input location name  $i$  may contain multiple locations or no real locations. Predicted entity triples are allowed to be composed of only empty strings (referred to as NULL), indicating that no location could be predicted for the given user input.

### 3.2 Data

**Target location database.** We use a modified version of the GeoNames<sup>1</sup> database, which contains entity names and coordinates for over 11M countries, administrative regions, counties, and cities across the globe. We filtered this database to exclude cities with populations under 15K, since tweets are more likely to come from more populated areas. Our final target location database contains 28,767 distinct locations: 252 countries, 3,947 administrative regions, and 24,568 cities.

**Train and test dataset.** We use data from the TWITTER-GLOBAL dataset (Zhang et al., 2022). The original dataset is described as containing data from 15.3M tweets which are either tagged with geocoordinates or Twitter Place objects and are posted by users with a non-empty Location field. The tweets were posted from 2013 to 2021 and contain global and multilingual data.

We use only the 4.1M geocoordinate-tagged tweets in TWITTER-GLOBAL, because geocoordinates are much more reliable than Place objects. Geocoordinates are meant to be the exact geolocation of the user’s device and always specify a precise latitude and longitude, while Place objects are pre-defined geographic entities that the user selects from a list and may be at any granularity, from points of interest to countries (Kruspe et al., 2021). Users may assign a Place to a post simply because they are talking about it and not because they are actually there. To identify the ground truth location for each geocoordinate-tagged tweet, we used the *reverse-geocoder* library to map the coordinates to the closest city in our target location database.<sup>2</sup> There are tweets from 196 different

<sup>1</sup><https://www.geonames.org/>

<sup>2</sup><https://github.com/thampiman/reverse-geocoder>

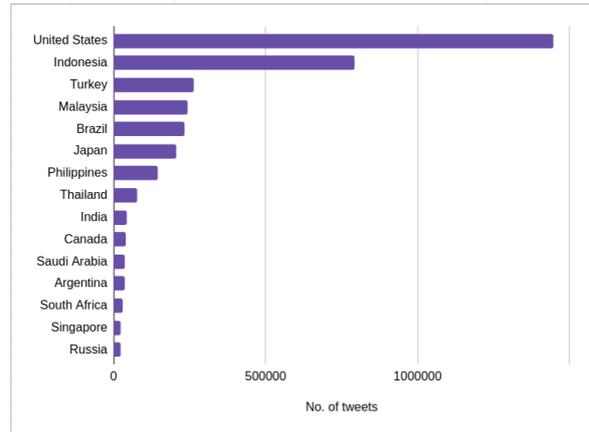


Figure 1: Tweet frequency for top 15 countries in the geocoordinate-tagged subset of TWITTER-GLOBAL.

countries, although they are not evenly distributed (see Fig. 1 for tweet distribution of the 15 most frequent countries).

We note limitations to this data. Geo-tagged tweets are not a random or representative sample of all tweets and there have been shown to be biases in who uses geo-tagging. Huang and Carley (2019b) show that less than 3% of Korean-speaking users geo-tag their posts while more than 40% of Indonesian-speaking users do. Certain countries like Turkey and Indonesia have very high percentages of coordinate-tagged tweets out of all geo-tagged tweets (53% and 67%, respectively), where most countries have 15-30%. In addition, users who have non-empty Location fields are more likely to use geo-tags, and there is strong homophily in geo-tagging behavior where users tend to connect to friends with similar geo-tagging preferences. Pavalanathan and Eisenstein (2015) show that users who geocoordinate-tag their posts versus have non-empty Location fields are measurably different groups in terms of demographics like age and gender.

It has also been shown that a non-empty Location field does not always correspond to the geo-tagged coordinates. Alex et al. (2016) find that in their dataset, 40% of users have geo-tagged coordinates within 10km of their specific Location, 70% are within 100km, and 85% are within 1000km. Despite these limitations, previous work studying geo-entity linking and user geolocation has found geocoordinate-tagged data to be useful, so we use such data here.

## 4 Methods

### 4.1 Proposed Method

We propose a method (referred to as UserGeo) that computes embeddings for each location in the target database and then, to predict a location for some user input, simply predicts the location with the closest embedding. To compute the target location embeddings, UserGeo uses training set  $T$  which contains a pair  $(x_t, y_t)$  for each tweet  $t$ , where  $x$  is the user-input Location field and  $y$  is the ground truth location triple. To help supplement locations with few user inputs, we additionally create a pair  $(s(d), d)$  for each  $d \in D$  where  $s(d)$  is a string representation of  $d$ , with comma-separated city, primary administrative region, secondary administrative region, country name, and two-letter country code, as applicable. UserGeo creates embedding representation  $Z_d$  for each location  $d \in D$  by averaging all associated  $x_t$  and  $s(d)$  embeddings:

$$Z_d = \frac{1}{|\{t : y = d\}|} \sum_{t:y=d} e(x_t)$$

where  $e(x)$  is the embedding for  $x$  from a pre-trained language model.

Then, for each user-input Location field  $i \in I$ , its predicted location triple  $\hat{d}_i$  is the location embedding  $Z_d$  that it has the highest cosine similarity with. If the cosine similarity with all location embeddings is below a given threshold  $t$ , then this is interpreted as a low confidence score and no prediction is made. In other words,

$$m = \max_{d \in D} c(Z_d, e(i))$$
$$\hat{d}_i = \begin{cases} \arg \max_{d \in D} c(Z_d, e(i)) & \text{if } m \geq t \\ \text{NULL} & \text{otherwise} \end{cases}$$

where  $c(a, b)$  is the cosine similarity between vectors  $a$  and  $b$ .

The motivation behind this method is that it leverages millions of examples of user-defined location names, essentially inducing a soft-alias location name database.<sup>3</sup> We note parallels between our framework and two other methodological classes. First, Bi-encoders obtain separate embeddings for

<sup>3</sup>We don't investigate fine-tuned models, due to two main disadvantages. First, it is unclear exactly what the fine-tuned model would learn (e.g. learning regional slang, instead of variable ways of expressing a location). Second, it is more likely for the model to overfit to the training data and perform poorly on countries that are not well-represented in it.

two sentences and then calculate cosine similarity between them. Second, Gaussian Discriminant Analysis classifies a point based on the minimum distance to clusters learned from training data.

### 4.2 Baselines

We evaluate four variations of our method as baselines. NameGeo: for each location  $d \in D$ ,

$$Z_d = e(s(d))$$

. In other words, a zero-shot version of UserGeo where each target location is represented only by its embedded string name.

Different embedding models: we evaluate embedding with three SBERT (Reimers and Gurevych, 2019) variants – the popular *all-MiniLM-L6-v2* model, the multilingually-trained *paraphrase-multilingual-miniLM-L12-v2* model, and the larger *all-mpnet-base-v2* model – and the geospatially grounded GeoLM (Li et al., 2023).

Variants: there are multiple  $s(d)$  functions to create different string representations of  $d$  that are all included in  $Z_d$ . The original  $s(d)$  represents  $d$  as only "<city>, <admin2>, <admin1>, <country>, <2-letter country code>". Here, if  $d$  is a country then it is also represented as "<country>"; if  $d$  is a primary administrative region, then it is also represented as "<admin1>", "<admin1> in <country>", and as "<country> / <admin1>"; and if  $d$  is a city, then it is also represented as "<city>", "<city> in <admin2> in <admin1> in <country>", "<admin1> / <city>", and as "<country> / <city>".

Pruning: removing outlier user inputs  $x_t$  from each  $Z_d$ . A user input is determined to be an outlier if the squared Euclidean distance between  $e(x_t)$  and  $Z_d$  is farther than a given threshold.

We also evaluate the only prior open-source tool that was explicitly created for and evaluated on broadly multilingual data. Carmen 2.0: uses a combination of regular expressions and manually curated aliases to predict real-world locations (Zhang et al., 2022).

## 5 Experiments

### 5.1 Experimental setup

**Data.** We divide the Twitter-Global data into a 90/10 split, with 3.7M examples in the training set and .4M in the test set. We evaluate at three levels of geographic granularity (city, primary administrative region, and country). A predicted triple is correct at the country level if the predicted country

Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
<i>all-MiniLM-L6-v2</i>			
NameGeo	59.8	37.7	14.3
UserGeo	<b>67.8</b>	<b>44.2</b>	14.8
<i>all-mpnet-base-v2</i>			
NameGeo	60.9	38.3	<b>14.9</b>
UserGeo	<u>67.4</u>	<u>43.7</u>	13.9
<i>paraphrase-multilingual-MiniLM-L12-v2</i>			
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

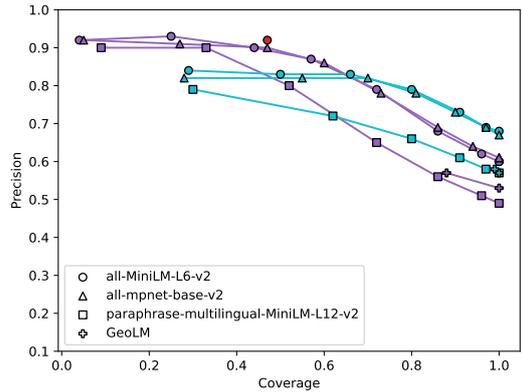
Table 2: Accuracy scores (%) at all three geographic granularities for Carmen 2.0, and NameGeo and UserGeo with different embedding models (all at threshold of 0, i.e. predictions are made for all examples). The highest score is bolded, the top 2 are underlined.

is a string match with the correct country; correct at the administrative level if both predicted country and administrative region are string matches with the correct ones; and correct at the city level if predicted country, administrative region, and city are all string matches with the correct ones. This hierarchical string matching procedure has limitations in that it does not remove all geographic ambiguity (e.g. if there are multiple cities with the same name in the same administrative region), but it should be effective in the vast majority of cases.

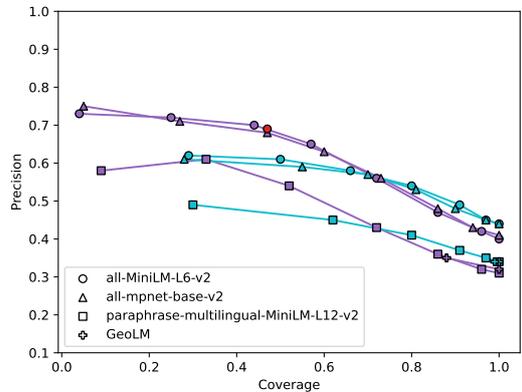
**Metrics.** We use four metrics to evaluate methods. *Accuracy* is the percentage of examples for which the method made a correct location prediction. Since many models do not always predict a location for all user inputs, we also evaluate *coverage* (the percentage of examples with a non-NULL prediction) and *precision* (the percentage of correct non-NULL predictions, following the standard definition of precision). For individual country performance we also evaluate F1-score, since most countries are uncommon.

## 5.2 Results

**NameGeo, UserGeo, and Carmen 2.0.** UserGeo achieves the highest accuracy at the country and administrative level, with gains over Carmen 2.0 of 25 and 17 points, respectively, and NameGeo achieves the highest accuracy at the city level, with gains over Carmen 2.0 of 5 points (Table 2). And while Carmen 2.0 has a competitive precision-coverage



(a) Country



(b) Admin.

Figure 2: Precision-coverage curves at the country (a) and administrative (b) levels. Red points are Carmen 2.0, purple are NameGeo, and cyan are UserGeo. NameGeo and UserGeo are shown with four different embedding models, where each point represents the precision and coverage at a threshold  $t \in \{0, .1, .2, .3, .4, .5, .6, .7, .8, .9\}$ . These curves demonstrate how thresholds can be used to tradeoff between precision and coverage with NameGeo and UserGeo.

tradeoff, especially at the country and administrative level, its overall coverage is still quite low (see Fig. 2, and Fig. 4 in Appendix). In other words, it is often correct when it makes a prediction but it does not often make a prediction. It also only has a single precision-coverage point (similar to most other prior tools). In contrast, our proposed methods demonstrate the ability to tradeoff between precision or coverage by choosing a different threshold and can achieve a higher precision or coverage than Carmen 2.0 at certain thresholds. The threshold can be used to adjust the precision/coverage balance if, for a given application, it is more important to get predictions correct or if it is more important

User-input	Carmen 2.0	NameGeo @0.5	UserGeo @0.5
TURKEY/SINOP	NULL	"" , Sinop, TR	Boyabat, Sinop, TR
福島県いわき市	NULL	Zhongshu, Yunnan, CN	Iwaki, Fukushima, JP
Catskills	NULL	Catalca, Istanbul, TR	Greenburgh, New York, US
where the wild things are	NULL	NULL	NULL

Table 3: Error analysis of the same user-input examples as in Table 1 (see Table 1 for corresponding real-world locations). Results from NameGeo and UserGeo are using the *all-MiniLM-L6-v2* SBERT model. Empty strings indicate that the model did not make a prediction at that geographic granularity, and NULL indicates that no prediction was made.

to make more predictions.<sup>4</sup>

We conducted a manual error analysis comparing Carmen 2.0, NameGeo, and UserGeo and observed trends in the types of errors made by each method (see Table 3). First, Carmen 2.0 rarely makes predictions for user-inputs with unexpected punctuation or in non-Latin scripts. Second, NameGeo often incorrectly predicts locations that look superficially similar to the user input (e.g. it predicts a location in China for a user input written in Japanese, and a location named ‘Catalca’ for the user-input ‘Catskills’). Third, UserGeo often correctly predicts locations for non-Latin inputs and alternate/informal location names. And lastly, all three models are frequently able to identify user-inputs that are not real locations.

We compare country-level F1-scores<sup>5</sup> across countries for Carmen 2.0, NameGeo, and UserGeo, for the 23 countries with more than 1,000 examples in our test set, in order to investigate geographic bias in our models (Liu et al., 2022). We observe that the number of examples per country in the training set – which may differ by multiple orders of magnitude – does not appear to influence UserGeo performance (see Table 6 in Appendix). This suggests that an unbalanced training set doesn’t negatively impact performance as it might for a traditional supervised learning method.

**Different embedding models.** Across SBERT bases, we find that the *all-MiniLM-L6-v2* model surprisingly performs better than the multilingually-trained *paraphrase-multilingual-MiniLM-L12-v2* model and it performs comparably with the larger *all-mpnet-base-v2* model (Table 2, Fig. 2). Anecdotally, we found that the multilingual model per-

forms worse for examples in Latin script; for example, NameGeo with the multilingually-trained model incorrectly predicts ("" , Pukapuka, CK) for user input "Kucukyali izmir" while NameGeo with *all-MiniLM-L6-v2* correctly predicts (Izmir, Izmir, TR). On the other hand, it performs better for non-Latin scripts; UserGeo with the multilingually-trained model correctly predicts (Dnipro, Dnipropetrovsk Oblast, UA) for "Днепрпетровск" , while UserGeo with *all-MiniLM-L6-v2* only partially correctly predicts (Pidhorodne, Dnipropetrovsk Oblast, UA) . We hypothesize that part of why the improved performance for non-Latin scripts does not outweigh the decrease in performance for Latin script is that non-Latin user inputs are a minority in this dataset; for a different dataset, it’s possible that the performance of the multilingually-trained model may be better than that of *all-MiniLM-L6-v2*.

Regarding the GeoLM model, the only SBERT model it outperforms is the multilingual one. Additionally, the cosine similarity threshold does not work effectively for GeoLM because the cosine similarities between a user input and each target location are very close; for example, the average cosine similarity between a correct prediction and an incorrect prediction for NameGeo with GeoLM at the country-level is .95 and .92, respectively. There is therefore a very limited range of values for which to have a threshold. Fig. 2 demonstrates this limited range with very short precision-coverage curves for GeoLM.

Since the *all-MiniLM-L6-v2* model performs the best out of all the embedding models, we use it for the rest of the experimental results.

**Variants.** Adding additional variants of the location name improves performance for NameGeo, and in fact NameGeo+variants does better than UserGeo at the city-level (see Table 4). However, adding additional variants does not improve perfor-

<sup>4</sup>We note that these precision-coverage curves are valid for only this dataset. Users would likely have to reevaluate precision and coverage on a new dataset or domain in order to choose an appropriate threshold.

<sup>5</sup>We use F1-score instead of accuracy here in order to better represent rare countries.

mance for UserGeo, and UserGeo+variants does marginally worse than UserGeo.

We investigate the impact of number of location mentions in the training data on model performance. Fig. 3 contains a plot of the number of mentions in the training data per location versus NameGeo/UserGeo accuracy. NameGeo does not use any user inputs from the training data so it can be considered a control for how training data affects performance in UserGeo. We see that while some training data is better than none (UserGeo consistently outperforms NameGeo when number of mentions is less than 1,000), it is not true that more training data always continues to improve performance (UserGeo and NameGeo have comparable performance when number of mentions is greater than 1,000). This suggests that having more variants or examples in a location’s averaged embedding is helpful, but only to an extent. Few-shot is better than zero-shot, but after a certain point there are diminishing returns.

**Pruning.** We defined an outlier as an embedding that was more than the average distance from its centroid, where the average is calculated using all embeddings associated with the given location. On average, 38% and 42% of embeddings in each cluster were pruned for UserGeo and UserGeo+variants, respectively.<sup>6</sup>

In general, removing outliers did not improve performance and frequently made it worse (Table 4). This suggests that having more variety in the averaged embeddings is good for performance, even if it comes at the cost of noise.

We observed via error analysis that it is often the case that user inputs close to the centroid will be standard English spellings of the location, ones within 1x average distance will include the location spelled in different scripts or nearby/related locations, and ones more than 1x average distance will not be very related to the location. For example, for the country Armenia, user inputs less than .5x average distance include "Yerevan, Armenia" and "ARMENIA"; user inputs less than 1x average distance include "Armenia | Հայաստան | Армения" and "Azerbaijan, Baku"; and user inputs more than 1x average distance include "Paris/Singapore" and "Worldwide".

<sup>6</sup>We determined through preliminary analyses that the average distance was an appropriate threshold, e.g. with a threshold of 2x the average distance, only 3% of embeddings in each cluster were pruned on average and there was no improvement in performance.

Method	Country	Admin.	City
NameGeo	59.8	37.7	14.3
+variants	62.0	40.9	<b>17.0</b>
UserGeo	<b>67.8</b>	<b>44.2</b>	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

Table 4: Accuracy scores (%) for variations of NameGeo and UserGeo (all at threshold of 0); the highest score is bolded, the top 2 are underlined. The *all-MiniLM-L6-v2* SBERT model was used for all results.

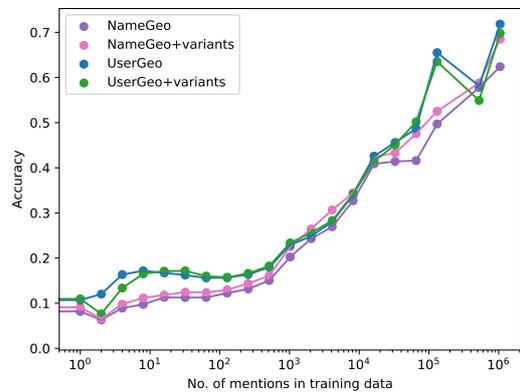


Figure 3: Average accuracy for a given number of mentions in training data (bucketed by  $\lfloor \log_2 \# \text{mentions} \rfloor$ ), for NameGeo and UserGeo with and without adding location name variants. Location entities at all 3 geographic granularities are present in the plot.

However, it is also frequently the case that user inputs one may want to exclude are less than 1x average distance and inputs one may want to include are more than 1x average distance. For example, "Moscow" is less than 1x average distance and "ARMENI, ABOBYAN." is more than 1x average distance. User inputs that are semantically relevant (i.e. refer to the location) but stylistically dissimilar (e.g. contain emojis or uncommon punctuation uses) are often farther from the centroid than user inputs that are less semantically relevant but more stylistically similar (e.g. "New York" is within 1x average distance of the centroid for the country Gibraltar, while "G I B R A L T A R" is not). These results suggest that the model is not able to effectively differentiate between good noise and bad noise, and thus it is better to not do pruning and to keep as many user inputs in the training data as possible.

## 6 Accuracy Upper Bounds

To approximate an upper bound for accuracy scores on our dataset, we conducted an analysis of a random sample of user inputs to determine how many contained references to actual locations. The percentage of examples which contain a location at each granularity suggest an upper bound for accuracy, indicating the approximate proportion that a geo-entity linker could be expected to connect to a real geographic location. The first author manually annotated 120 random examples from the training set for whether or not the user input contained an identifiable reference to a location at the country, administrative, and/or city levels.<sup>7</sup>

We include results from this annotation in Table 5 as well as a comparison with our best performing models’ accuracy scores. UserGeo is only 5 points below the country-level upper bound and 14 points below the administrative-level upper bound, indicating that our current performance is fairly close to the upper limit in terms of accuracy. In contrast, NameGeo+variants is more than 30 points below the city-level upper bound.

We discuss a few problems with evaluating geo-entity linking at the city level for social media data. First, the assumption of the geocoordinates as ground truth is frequently untrue and especially so at the fine-grained city level. Unlike news articles or Wikipedia data, the place that a user puts on social media is not necessarily the place where they actually are. As Alex et al. (2016) have shown, only 70% of users have geocoordinates within 100km of the place specified in their Location field. Second, since we define ground truth as the closest city to the geocoordinates with a population higher than 15K, there will always be a mismatch between ground truth and Location field if a user puts a city with a lower population. Third, there is a disconnect between what is considered a city by a database such as GeoNames versus by an everyday person. Istanbul, Jakarta, Moscow, Gaborone, Lampung, and Santa Ana are all names of both cities and administrative regions in GeoNames – when users put them in their Location field, it’s unclear which one they are referring to.

We hypothesize that these problems contribute to the poor performance of geo-entity linking tools

<sup>7</sup>We did not count coordinates as containing a location, as we wanted to investigate the ability of the model to identify natural language location references. If a location was named that could be either an administrative region or a city, then the example was marked as containing both.

	Country	Admin.	City
Upper bound	72.5	58.3	49.2
NameGeo+variants	62.0	40.9	<b>17.0</b>
UserGeo	<b>67.8</b>	<b>44.2</b>	14.8

Table 5: Accuracy upper bounds and scores (%) for our two best models, with highest scores bolded.

at the city level. We additionally acknowledge how inferring location at the city level is more invasive than inferring at the country or administrative level because it predicts more fine-grained information about the user. Therefore, due to mismatches between geocoordinates and user-defined Location fields as well as the privacy concerns, we suggest that researchers only predict city-level location if it’s necessary for a given application and otherwise use country- or administrative-level predictions.

## 7 Conclusion and Future Work

In this paper we introduced new methods for geo-entity linking noisy, multilingual social media data with selective prediction. Of our two best performing methods, one does not require any training data (NameGeo+variants) while the other achieves state-of-the-art performance at the country and administrative levels (UserGeo). We also discussed problems with geo-entity linking at the city level for social media data, and suggested against doing this unless necessary for the application.

This work is not without limitations. We do not compare performance with LLM prompting methods, as they are too expensive for large-scale datasets and it’s currently unclear how much prompt testing is necessary for a robust evaluation. However, it would be a useful comparison to know how LLMs perform at zero- or few-shot geo-entity linking; we leave this for future work. We also acknowledge that our methods rely heavily on SBERT models, and thus are reliant on a third party for sustaining them.

We plan to release a version of the UserGeo location embeddings in future, although we will not release the current version due to concerns about the TWITTER-GLOBAL data (e.g. unbalanced across countries, unclear sampling methodology). We also hope to further evaluate our method on other domains with noisy location references, such as historical data, and to explore extensions of our method so it may be used for the broader task of geoparsing unstructured text.

## 8 Ethical Considerations

We discuss here ethical issues that may arise from using any geo-entity linking tool on social media data for a downstream application. When using a geo-entity linking tool on social media data, there is always the risk of de-anonymizing users through the inference of sensitive location information. It is generally recommended to use the lowest granularity necessary for the application simply because it is easier to correctly obtain than more fine-grained data (Kruspe et al., 2021), but it is also true that lower geographic granularities protect individuals more due to being part of a larger aggregate (Dupre et al., 2022). We also acknowledge the increasing importance of geomasking techniques, which aim to protect the privacy of individuals while preserving spatial information in geodata. Lorestani et al. (2024) survey the privacy risks of geocoded data and present a taxonomy of current geomasking techniques, and Gao et al. (2019) specifically examine the efficacy of geomasking techniques for protecting the privacy of Twitter users.

### Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and feedback. This work was supported by a National Science Foundation Graduate Research Fellowship (No. 1938059).

### References

- Beatrice Alex, Clare Llewellyn, Claire Grover, Jon Oberlander, and Richard Tobin. 2016. Homing in on twitter users: Evaluating an enhanced geoparser for user profile locations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3936–3944.
- Mariona Coll Ardanuy, Kasra Hosseini, Katherine McDonough, Amrey Krause, Daniel van Strien, and Federico Nanni. 2020. A deep learning approach to geographical candidate selection through toponym matching. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 385–388.
- Mariona Coll Ardanuy and Caroline Sporleder. 2017. Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 175–180.
- Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2019. Using recurrent neural networks for toponym resolution in text. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part II 19*, pages 769–780. Springer.
- Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2022. A novel deep learning approach using contextual embeddings for toponym resolution. *ISPRS International Journal of Geo-Information*, 11(1):28.
- Grant DeLozier, Jason Baldrige, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.
- Catherine D’Ignazio, Rahul Bhargava, Ethan Zuckerman, and Luisa Beck. 2014. CLIFF-CLAVIN: Determining Geographic Focus for News Articles. *NewsKDD: Data Science for News Publishing*.
- Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, volume 23, page 45. Citeseer.
- Sam Dupre, Lindsay Spell, and Paul Jung. 2022. Geospatial data disclosure avoidance and the census. *Select Topics in International Censuses (STIC 1, 1 (2022))*, 1–9.
- Nicholas FitzGerald, Dan Bikel, Jan Botha, Dan Gillick, Tom Kwiatkowski, and Andrew McCallum. 2021. Moleman: Mention-only linking of entities with a mention annotation network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 278–285.
- Song Gao, Jinmeng Rao, Xinyi Liu, Yuhao Kang, Qunying Huang, and Joseph App. 2019. Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of twitter users. *Journal of Spatial Information Science*, (19):105–129.
- Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Which melbourne? augmenting geocoding with maps. Association for Computational Linguistics.
- Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889.
- Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062.

- Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2023. Location reference recognition from texts: A survey and comparison. *ACM Computing Surveys*, 56(5):1–37.
- Binxuan Huang and Kathleen Carley. 2019a. A hierarchical location prediction neural network for Twitter user geolocation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4732–4742, Hong Kong, China. Association for Computational Linguistics.
- Binxuan Huang and Kathleen M Carley. 2019b. A large-scale empirical study of geotagging behavior on twitter. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 365–373.
- Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding us regional linguistic variation with twitter data analysis. *Computers, environment and urban systems*, 59:244–255.
- Mike Izbicki, Vagelis Papalexakis, and Vassilis Tsotras. 2019. Geolocating tweets in any language at any location. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 89–98.
- David Jurgens, Tyler Finethy, James McCorrison, Yi Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 188–197.
- Ehsan Kamaloo and Davood Rafiei. 2018. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1287–1296.
- Morteza Karimzadeh, Wenyi Huang, Siddhartha Banerjee, Jan Oliver Wallgrün, Frank Hardisty, Scott Pezanowski, Prasenjit Mitra, and Alan M MacEachren. 2013. Geotxt: a web api to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval*, pages 72–73.
- Morteza Karimzadeh, Scott Pezanowski, Alan M MacEachren, and Jan O Wallgrün. 2019. Geotxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23(1):118–136.
- Anna Kruspe, Matthias Häberle, Eike J Hoffmann, Samyo Rode-Hasinger, Karam Abdulahhad, and Xiao Xiang Zhu. 2021. Changes in twitter geolocations: Insights and suggestions for future usage. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 212–221.
- Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldrige, Eugene Ie, and Li Zhang. 2021. Multi-level gazetteer-free geocoding. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 79–88.
- Abhinav Kumar and Jyoti Prakash Singh. 2019. Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction*, 33:365–375.
- Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1474–1477.
- Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. 2022. Spabert: A pretrained language model from geographic data for geo-entity representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2757–2769.
- Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. 2023. Geolm: Empowering language models for geospatially grounded language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240.
- Michael D Lieberman and Hanan Samet. 2012. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 731–740.
- Zilong Liu, Krzysztof Janowicz, Ling Cai, Rui Zhu, Gengchen Mai, and Meilin Shi. 2022. Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3:9.
- Mahrokh Abdollahi Lorestani, Thilina Ranbaduge, and Thierry Rakotoarivelo. 2024. Privacy risk in geodata: A survey. *arXiv preprint arXiv:2402.03612*.
- Xiangyang Luo, Yaqiong Qiao, Chenliang Li, Jiangtao Ma, and Yimin Liu. 2020. An overview of microblog user geolocation methods. *Information processing & management*, 57(6):102375.
- Ludovic Moncla, Walter Renteria-Agualimpia, Javier Noguera-Iso, and Mauro Gaio. 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd acm sigspatial international conference on advances in geographic information systems*, pages 183–192.
- Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148.

- Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 167–176.
- Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1362–1367.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Hannes Rosenbusch, Anthony M Evans, and Marcel Zeelenberg. 2020. Interregional and intraregional variability of intergroup attitudes predict online hostility. *European Journal of Personality*, 34(5):859–872.
- Breno Dourado Sá, Ticiano Coelho da Silva, and José Antônio Fernandes de Macêdo. 2022. Enhancing geocoding of adjectival toponyms with heuristics. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 37–45.
- Praval Sharma, Ashok Samal, Leen-Kiat Soh, and Deepti Joshi. 2023. A spatially-aware data-driven approach to automatically geocoding non-gazetteer place names. *ACM Transactions on Spatial Algorithms and Systems*, 10(1):1–34.
- David A Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *International Conference on Theory and Practice of Digital Libraries*, pages 127–136. Springer.
- Michael Speriosu and Jason Baldridge. 2013. Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1476, Sofia, Bulgaria. Association for Computational Linguistics.
- Jimin Wang and Yingjie Hu. 2019. Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23(6):1393–1419.
- Xiaobin Wang, Chunping Ma, Huafei Zheng, Chu Liu, Pengjun Xie, Linlin Li, and Luo Si. 2019. DM\_NLP at semeval-2018 task 12: A pipeline system for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 917–923.
- Jingyu Zhang, Alexandra DeLucia, and Mark Dredze. 2022. Changes in tweet geolocation over time: A study with carmen 2.0. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 1–14.
- Jingyu Zhang, Alexandra DeLucia, Chenyu Zhang, and Mark Dredze. 2023. Geo-seq2seq: Twitter user geolocation on noisy data through sequence to sequence learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4778–4794.
- Zeyu Zhang and Steven Bethard. 2023. Improving toponym resolution with better candidate generation, transformer-based reranking, and two-stage resolution. *arXiv preprint arXiv:2305.11315*.

# A Appendix

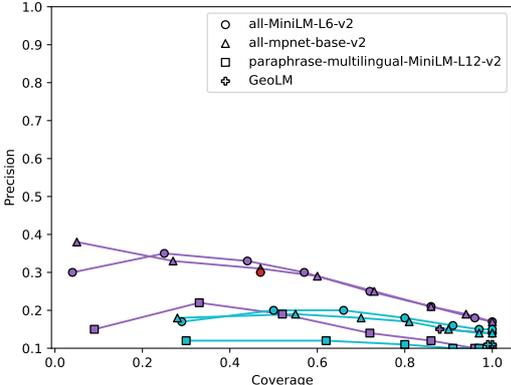


Figure 4: Precision-coverage curve at the city level. The red point is Carmen 2.0, purple points are NameGeo, and cyan are UserGeo.

Country	No. mentions in train data	Carmen 2.0	NameGeo	UserGeo
United States	1300869	.73	.77	.84
Indonesia	713967	.54	.73	.74
Turkey	237472	.57	.91	.91
Malaysia	218257	.49	.70	.70
Brazil	210428	.47	.78	.89
Japan	184647	.27	.48	.92
Philippines	131152	.57	.69	.73
Thailand	70403	.55	.67	.64
Singapore	43996	.51	.54	.53
India	40656	.80	.91	.93
Canada	36990	.54	.83	.78
Saudi Arabia	34559	.40	.48	.73
Argentina	34104	.54	.72	.84
South Africa	27228	.59	.77	.80
Russia	21820	.53	.92	.91
Kuwait	17829	.48	.65	.59
Australia	14363	.53	.85	.86
Chile	14268	.80	.87	.89
Nigeria	12482	.80	.84	.87
Spain	11699	.40	.65	.79
Egypt	10979	.57	.72	.78
UAE	10918	.43	.71	.67
Pakistan	9423	.79	.86	.89

Table 6: Per-country country-level F1-scores for the 23 countries with over 1,000 examples in the test set. Results from NameGeo and UserGeo are using the *all-MiniLM-L6-v2* SBERT model. Both NameGeo and UserGeo outperform Carmen 2.0 for each of the 23 countries. We also note that F1-score can vary significantly by country, demonstrating the importance of evaluating per-country performance when using a geo-entity linker to investigate social media users who are from or identify with particular countries.