IDRISI-D: Arabic and English Datasets and Benchmarks for Location Mention Disambiguation over Disaster Microblogs

Reem Suwaileh[1], **Tamer Elsayed**[1], **Muhammad Imran**[2]
[1] Computer Science and Engineering Department, Qatar University
[2] Qatar Computing Research Institute, Hamad Bin Khalifa University
Doha, Qatar
{rs081123,telsayed}@qu.edu.qa, mimran@hbku.edu.qa

Abstract

Extracting and disambiguating geolocation information from social media data enables effective disaster management, as it helps response authorities; for example, locating incidents for planning rescue activities and affected people for evacuation. Nevertheless, the dearth of resources and tools hinders the development and evaluation of Location Mention Disambiguation (LMD) models in the disaster management domain. Consequently, the LMD task is greatly understudied, especially for the *low resource languages* such as *Arabic*. To fill this gap, we introduce IDRISI-D, the largest to date English and the first Arabic public LMD datasets. Additionally, we introduce a modified hierarchical evaluation framework that offers a lenient and nuanced evaluation of LMD systems. We further benchmark IDRISI-D datasets using representative baselines and show the competitiveness of BERT-based models.

1 Introduction

The cost-effectiveness and efficiency of communication over social media platforms make them primary sources of information during disaster events and emergencies. An essential dimension that makes the data extracted from microblogging platforms (e.g., X platform, formerly Twitter) invaluable and actionable is the geolocation information. Nevertheless, users typically opt to disable the geolocation functionalities over social media platforms to preserve their own safety and privacy which necessitates the development of geolocation extraction tools *for social good*. In this paper, we focus on the Location Mention Disambiguation (LMD) task over microblogs that we exemplify by X posts. An LMD system aims at matching location mentions (LMs) appearing in microblogs to toponyms, i.e., place or location names, in a geopositioning database, i.e., gazetteer.

Unfortunately, the research community lacks access to public disaster-specific microblogging LMD datasets, especially for low-resource languages, which consequently prevents the development and comparison of robust LMD systems. For example, there are only two English LMD datasets, namely Singapore (Ji et al., 2016) and GeoCorpora (Wallgrün et al., 2018), where the former dataset is geographically confined, lacks event context, and is not publicly available, whereas the latter one (i.e., GeoCorpora (Wallgrün et al., 2018)) is public, it has the same issues of low geographical coverage, lacking disaster event context, and many relevant/informative posts that do not contain the tracking keywords (Suwaileh et al., 2023a). On the other front, there are no Arabic LMD datasets to the best of our knowledge.

In this paper, we fill this gap and release IDRISI-D datasets[1] for Arabic (IDRISI-DA) and English (IDRISI-DE) languages. IDRISI-DA is the first public human-labeled Arabic (a low-resource language) dataset, constituting 2,869 posts and 3,893 LMs. IDRISI-DE is the largest to date human-labeled English microblogging dataset in terms of number of LMs. It constitutes 5,591 posts and 9,685 LMs. Additionally, to alleviate the lack of context challenge for microblogs and toponyms, we asked annotators to judge different features such as hashtags, replies, and URLs, among others, for usefulness for the LMD task.

Furthermore, to evaluate the LMD systems, Accuracy (Acc), Precision (P), Recall (R), and the F_β score are typically computed (Zhang and Gelernter, 2014; Li et al., 2014; Ji et al., 2016; Middleton et al., 2018; Wang and Hu, 2019a; Xu et al., 2019). While these measures evaluate binary classification tasks, the LMD task is usually perceived as a multi-class classification where every LM has only one

[1]Named after Muhammad Al-Idrisi, who is one of the pioneers and founders of advanced geography: https://en.wikipedia.org/wiki/Muhammad_al-Idrisi. The "D" refers to the **d**isambiguation task. Release: The link is removed due to the blind-review policy. The dataset and evaluation script are attached as *Supplementary Materials*.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 158–169
December 7, 2023 ©2023 Association for Computational Linguistics

(or no) correct toponym in gazetteers. Moreover, distance-based methods (Wang and Hu, 2019a), are also used to evaluate LMD systems within a distance d that is commonly set to 161 KM (100 miles). For example, $Acc@d$ is the fraction of correctly predicted LMs that are within d. However, tuning the d for different location granularity was not empirically investigated.

To address these shortcomings, we propose evaluating the LMD systems using ranking evaluation measures, namely the Mean Reciprocal Rank at cut-off r ($MRR@r$) in a lenient hierarchical strategy (Mourad et al., 2019) where systems are evaluated at different location granularity such as country, city, street, etc. Indeed, the hierarchical evaluation substitutes the distance-based measures but in discrete manner.

The contributions of this work are as follows:

- We present IDRISI-DA, the first Arabic LMD dataset containing about 2,869 posts and 3,893 LMs.
- We present IDRISI-DE, the largest *manually-labeled* public English LMD dataset of about 5,461 posts and 9,685 LMs.
- We manually label and analyze the usefulness of different features, including hashtags, event context, and URLs, replies, named entities, and other LMs, to draw helpful insights for developing effective LMD systems.
- We present a modified hierarchical LMD evaluation for classification and ranking methods.
- We provide simple yet effective English and Arabic LMD baselines.

The remainder of this paper is organized as follows. We present the related work in Section 2. We then define the LMD task in Section 3. We introduce IDRISI-D datasets and analyze them in Sections 4 and 5, respectively. We then benchmark the datasets in Section 6. We next discuss the dataset use cases in Section 7. We finally conclude in Section 9.

2 Related Work

In this section, we discuss the LMD related studies and discuss their technical solutions (Section 2.1) and evaluation (Section 2.2).

2.1 Technical Solutions

There are a few studies that tackle the LMD task using machine learning and deep learning techniques. For instance, Geoparspy (Middleton et al., 2018) is a Support Vector Machine (SVM) model trained on gazetteer-based features including location type, population, and alternative names. Additionally, the disambiguation models of the toponym resolution system employed by Wang and Hu (2019a) are essentially machine learning models including (i) *DM_NLP* (Wang et al., 2019) which is a Light Gradient Boosting Machine (LightGBM) model trained on similarity scores, contextual representations, gazetteer attributes, and mention list features, (ii) *UniMelb* (Li et al., 2019) which is an SVM that uses different feature types such as the history results in the training dataset, population, gazetteer attributes, similarity, and mention neighbors features, and (iii) *UArizona* (Yadav et al., 2019) which is a heuristic-based system that favors toponyms with higher populations.

Furthermore, Xu et al. (2019) proposed an attention-based two-pairs of bi-LSTMs for matching LMs against Foursquare gazetteer. Each location profile (*lp*) in Foursquare is represented by concatenating one-hot vector for the category attribute, TF-IDF vectors for textual attributes (e.g., address attribute), and the numeric-based attributes. On the other hand, the LM is represented using its context (i.e., post) and encoded using contextual representation attended to the *lp* vector, besides the geographical distance. The two-pair networks learn the left and right contexts of the LM. Both representations then go through a fully connected layer to learn disambiguation.

2.2 Evaluation

There is a dearth of microblogging disaster-specific LMD datasets. Table 1 presents the only two LMD datasets and their statistics. GeoCorpora (Wallgrün et al., 2018) is the only available one for the research community. Wang and Hu (2019a) evaluated it using eight different datasets available through EUPEG framework (Wang and Hu, 2019b), solely one of which is a microblogging dataset that is GeoCorpora. Xu et al. (2019) used Singapore dataset (Ji et al., 2016) for evaluation.

As for the evaluation measures, the distance-based measures have been used in non-disaster-specific studies to evaluate LMD systems. For that, the distance between the GPS coordinates of the gold and predicted LMs is measured using the great circle distance. The systems' overall performance is then computed by the Median and Mean Error

Dataset	# Twt	# LM (unique)	Labeling	LM types	Public
Singapore (Ji et al., 2016)	3,611	1,542 (-)	In-house	-	×
GeoCorpora (Wallgrün et al., 2018)	6,648	3,100 (1,119)	Crowd	×	✓
IDRISI-DE	5,591	9,586 (1,601)	In-house	✓	✓
IDRISI-DA	2,869	3,893 (763)	In-house	✓	✓

Table 1: The existing LMD datasets compared to IDRISI datasets.

Distance.

Additionally, the discrete measures including Accuracy (Acc), Precision (P), Recall (R), and the F_β score are computed to evaluate systems (Zhang and Gelernter, 2014; Li et al., 2014; Ji et al., 2016; Middleton et al., 2018; Wang and Hu, 2019a; Xu et al., 2019), however, they provide a bird's-eye view of systems' performance neglecting the nuance in their techniques. To overcome this shortcoming, Karimzadeh (2016) proposed using Cross Entropy (CE) that considers the probabilities of systems rather than their ranks, Root Mean Square Error (RMSE) that quantifies the average great circle distance between predicted and gold toponyms, and Eccentricity that combines both CE and RMSE.

Acc, P, R, and F_β can also be computed within a distance d that is commonly set to 161 KM (100 miles). For example, Acc@d is the fraction of correctly predicted LMs within d.

While these measures evaluate binary classification tasks, the LMD is typically modeled as a multi-class classification task making them inappropriate for evaluation.

3 Problem Definition

The LMD System, as illustrated in Figure 1, is given the following inputs:

- A post (a microblog) p that is related to a disaster event e,
- A set of location mentions (LMs): $L_p = \{l_i; i \in [1, n_p]\}$ in post p, where l_i is the i^{th} location mention and n_p is the total number of location mentions in p, if any.
- A geo-positioning database G (i.e., gazetteer) that consists of a set of toponyms: $T = \{t_j; i \in [1, k]\}$, where t_j is the j^{th} toponym, and k is the number of toponyms in G.

The LMD system aims to match every location mention l_i in the post p to one of the toponyms t_j in G that accurately represents l_i, if exists. Otherwise, the system must abstain and declare that l_i is unresolvable (or unlinkable).

4 Dataset Construction

In this section, we discuss the constructing process of IDRISI-D datasets. We start by describing IDRISI-R datasets. We then present the sampling strategy and the annotation process.

IDRISI-R Datasets: We extend IDRISI-R Location Mention Recognition (LMR) English (IDRISI-RE) (Suwaileh et al., 2023a) and Arabic (IDRISI-RA) (Suwaileh et al., 2023b) datasets that are originally sampled from HumAID (Alam et al., 2021) and Kawarith (Alharbi and Lee, 2021) datasets, respectively. We select these datasets due to their unique characteristics as described below.

IDRISI-RE is the largest to date LMR microblogging English dataset. It exhibits unique diversity (domain and location types), coverage (temporal and geographical), and generalizability (domain and geographical), compared to all existing datasets of its kind. It comprises around 20k human-labeled (gold) and 57k machine-labeled (silver) posts from 19 disaster events of diverse types covering wide geographical areas. The events capture the critical periods of disaster events. The annotations include spans of location mentions in the textual content alongside their location types (e.g., country, city, street). Empirically, IDRISI-RE is the best domain and geographical generalizable dataset against all existing English datasets.

IDRISI-RA is the first Arabic LMR microblogging dataset. It contains 22 disaster events of different types that happened in Arab countries, covering various dialects reasonably. It contains 4.6K manually-annotated (gold) posts sampled from 7 disaster events,[2] and 1.2M automatically-annotated (silver) posts sampled from the entire dataset. Both versions are labeled for location mentions and location types. Empirically, the LMR models trained on IDRISI-RA showed decent generalizability to unseen events and acceptable domain and geographical generalizability.

[2]These events are labeled for informativeness in Kawarith dataset.

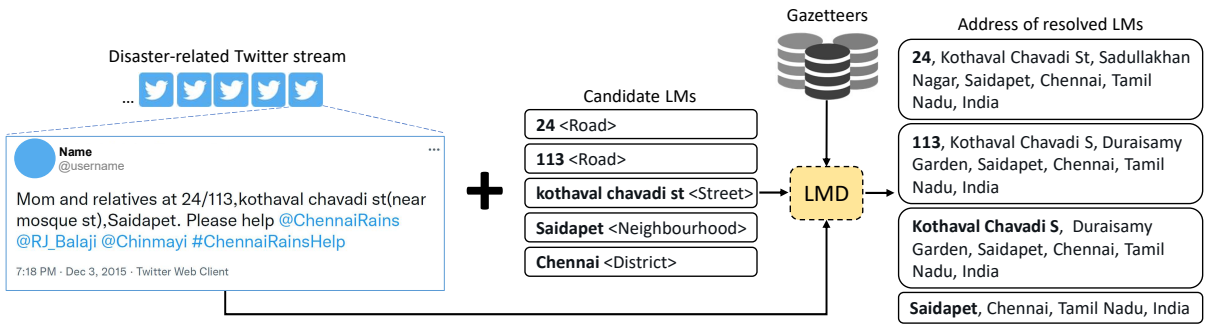

Figure 1: High-level overview of the Location Mention Disambiguation (LMD) task.

Dataset Sampling: Constrained by not overwhelming the *volunteered* annotators, we sampled a set of posts from every disaster event in IDRISI-RE while maintaining the distribution of LM types, but covering all fine-grained LMs including neighborhoods, streets, and POIs. In total, we sampled 8,224 posts containing 11,023 LMs. On the other hand, the IDRISI-RA gold version was labeled entirely, including seven events containing 2,974 having LMs (the remaining 1,618 posts do not contain LMs) and 5,236 LMs.

Dataset Annotation: The LMD annotation removes the ambiguity of geo/geo entities (in contrast to the geo/non-geo LMR annotations). We collected the LMD annotations in 3 phases to increase the reliability of annotations with the minimum burden on the expert annotators:

P1. Two in-house annotators are assigned for every event with the condition of having a good familiarity with the country of the disaster event. When one of the annotators declares a low confidence for a specific LM or both disagree, the LM is forwarded to a meta annotator in *Phase 2*.

P2. A meta annotator resolves the disagreement from *Phase 1* and labels the low-confident examples. She has a solid understanding of the LMD task; hence, she verifies the doubtful annotations. When she fails to disambiguate an LM, it goes to experts in *Phase 3*.

P3. Expert annotators disambiguate the hard unresolved LMs from *Phases 1* and *2*. Experts are residents of the countries where the disaster events took place.

In all phases, annotators attentively read the post online alongside replies and the linked web pages. Next, they (1) disambiguate the LMs by searching OpenStreetMap (OSM) gazetteer through Nominatim search engine[3] to find the best matching toponym, (2) assign a confidence score between 1-3 for their annotation, and (3) judge the usefulness of features for disambiguation ("Yes", "No", or "None"). The features we investigate their usefulness include:

- Event: The disaster event name.
- Hashtags: The set of posts having the same hashtag as the target post within their text.
- Replies: The thread or responses to the post.
- Other LMs: Other location mentions appearing within the same post text.
- URLs: The linked web pages or media within the post text.
- Entities: Named entities that appear within the post text.

We define the usefulness as *whether a feature helps the annotator to accurately find the correct toponym from the OSM that best matches the candidate LM being annotated*.

Additionally, to avoid propagating human errors from IDRISI-R, we asked the annotators to modify LMs, add new LMs, or drop LMs in certain cases. In Table 2, we show example posts and elaborate on them in the following:

Modifying LMs: Several cases require modification, such as separating multiple LMs (Posts #1 and #6), fixing LM boundaries (Posts #2 and #7), and fixing LM type (changing "Street" to "City" in Post #3 and "City" to "POI" in Post #8), to list a few. Annotators modified 15 and 154 LMs in both IDRISI-DA and IDRISI-DE, respectively. IDRISI-RA is cleaner than IDRISI-RE as it was labeled in-house.

Dropping LMs: Annotators dropped LMs when they violate the LMR annotations guidelines. Cases include organization or person entities (Posts #4

[3] https://nominatim.openstreetmap.org/

and #9), ambiguous LMs (Posts #5), nationalities, and locational descriptions, among others. In total, we dropped 212 and 1,986 mentions, 97 and 435 of which are unique, from IDRISI-DA and IDRISI-DE datasets, respectively.

Adding new LMs: Annotators added unlabeled LMs if they are resolvable. For example, the "Pontagea Health Centre" in Post #10. This resulted in adding 27 new LMs to IDRISI-DE while no LMs are added to IDRISI-DA.

Adding LMs to OSM: Annotators added 171 and 27 new toponyms to OSM for IDRISI-DA and IDRISI-DE, respectively.

We ran the annotation task for ten weeks and obtained the final IDRISI-DE and IDRISI-DA datasets. Table 1 presents their statistics.

5 Dataset Analysis

IDRISI-D datasets inherit the geographical, domain, location types, temporal, informativeness, and dialectical (for Arabic) coverage from IDRISI-R datasets. In this section, we analyze the reliability of annotations and the usefulness of post features for the LMD task.

Reliability: To evaluate the reliability of annotations in Phase 1, we compute the Inter-Annotator Agreement (IAA) using Cohen's Kappa (Cohen, 1960). We measure the IAA for the ability to resolve LMs, i.e., whether an LM is resolvable or not. We also compute the agreement percentage on the extracted toponyms from gazetteers by the annotators for all LMs. The annotators in Phase 1 achieved substantial and almost perfect Cohen's Kappa scores of approximately 0.90 and 0.83 for IDRISI-DA and IDRISI-DE datasets, respectively. The raw agreement percentages are around 97.98% and 93.50% for IDRISI-DA and IDRISI-DE datasets, respectively. These results statistically demonstrate the high quality and reliability of annotations of IDRISI-D datasets. To further increase the quality of the datasets, we resolved the disagreement cases in the subsequent annotation phases 2 and 3.

Usefulness of Features: Table 4 shows the percentages of features' presence in posts and the percentages of useful features. We show the statistics for: (i) "ALL": all types of LMs in the datasets, (ii) "Coarse": the coarse-grained LMs including countries, cities, states, counties, districts, and neighborhoods, and (iii) "Fine": the fine-grained LMs including streets, natural POIs, human-made POIs. Apparently, the "event", "other LMs", and "hashtags" are the most useful features for LMD, especially for fine-grained LMs.

Looking carefully at the annotations of features' usefulness, we make different observations through examples in Table 3:

Event: Knowing the event place helps in narrowing the search space over OSM. Consequently, annotators can mitigate the "Toponymic homonymy" challenge (Suwaileh et al., 2022). In Post #1, all results for "شارع كورنيش النيل" ("Corniche El Nile Street") in Post #1 are not within "Cairo" where the "Cairo BMB 2019" event took place. Thus, searching toponyms within the affected area results in accurate annotations.

Other LMs: The geo-vicinity between co-occurring LMs usually represents inclusion and containment relationships, making the coarse-grained LMs useful to disambiguate the fine-grained LMs. For instance, in Post #2, "بيروت" ("Beirut" is a city) which is also a hashtag is helpful for accurately disambiguating "مستشفى القديس جاورجيوس" ("Saint George Hospital" is a human Point-of-Interest). Similarly, in Post #4, "Nebraska" (State) was useful to distinguish "Elkhorn River" (Human Point-of-Interest) from another part of the river located in "West Virginia" (State). Different reasons cause the low usefulness percentages of "other LMs". To elaborate, in cases where the same LM appears multiple times in the same post, the duplicates are useless for disambiguating each other.

Hashtags: As most hashtags indicate the disaster event (e.g., "#انفجار_معهد_الأورام" and "#انفجار_المرفأ" in posts #1 and #2), they are equally important to the "Event" feature.

Replies: Typically, a small number of posts get the community attention. Hence, replies are rarely useful for LMD.

URLs: Linked web pages are useful if they elaborate on the geographical context of the reported information in the post. For example, the linked web page in Post #2 was useful for locating the hospital. Also, "Lake Butler" in Post #4 is challenging LM. The linked Facebook page contains "Lake Butler, FL, United States" and "Keystones Heights" that helped the annotator to successfully resolve this LM by their geo-proximity. The importance percentage of URLs is low as many URLs are already broken or require a paid subscription.

#	Change	Post text
1	Separate LMs	…حملة نكرم موتانا لإعادة دفن القبور التي كشفتها سيول الأمطار في **مقبرتي الصليبيخات و صبحان**
2	Modify offsets	…طقس غير مستقر … على السواحل الشمالية الغربية **ووسط سيناء وجنوب سيناء وشمال الصعيد**
3	Modify type	الطرق التالية مغلقة بسبب الغبار الكثيف وتدني مدى الرؤية: * … باتجاه **الزرقاء**$_{Street \rightarrow City}$
4	Drop ORG	إصابة ٤٠ موظفاً في **جمعية كيفان** بفيروس كورونا - تم إغلاق السوق المركزي وجميع الأفرع …
5	Drop undefined	… ولازالت ٢٨ حالة مصابة تحت العلاج بـ **مستشفى العزل** بمعبر رفح
6	Separate LMs	Please join us for Hurricane Maria relief this Saturday on Melrose St btwn **Buchwick** & **Broadway** ...
7	Modify offsets	The University of **Nebraska** **Omaha** Love Your Melon Crew sure knows how to make kids happy ... #MealsThatHeal
8	Modify type	Amidst applause, Canadas rescue team arrives in **Mexico City Airport**$_{City \rightarrow POI}$ on Saturday #earthquake #CASDDA via [user_mention]
9	Drop ORG	**Rosen Hotels & Resorts** in Orlando announces availability of 30 guestrooms at [user_mention] for #HurricaneIrma evacuees...
10	Add LM	Pontagea Health Centre in Beira, #Mozambique, was partially destroyed by #CycloneIdai, ...

Table 2: Examples of issues and corrections in LMD annotations. **Bold** text is the annotated LMs in IDRISI-R. Underlined text is the corrected LMs in IDRISI-D.

#	Useful features	Post text
1	Event, Other LMs, Hashtag.	أغلقت الإدارة العامة للمرور شارع كورنيش النيل (خلف **جاردن سيتي**) على خلفية اندلاع حريق بـ **معهد الأورام** في **#المنيل** **#انفجار_معهد_الأورام** ...
2	Other LMs, Hashtag & URL	**#انفجار_المرفأ** ... يتسبب في دمار كبير بـ مستشفى القديس جاورجيوس في **#بيروت** **https://t.co/7SdALOhviW**
3	None	فيديو يوضح وجود مفرقعات … داخل أحد المستودعات قبل حصول إنفجار في #بيروت ...
4	Other LMs	Human remains discovered along Elkhorn River after flooding, sheriff says https://buff.ly/2CEShla #**Nebraska**
5	URL	In the wake of Hurricane Irma, we've planned a food distribution event in Lake Butler to help anyone affected by... **fb.me/2fbe0b4YE**
6	None	Labatt to help those affected by Fort McMurray wildfire [...] #FortMcMurray #LCBO

Table 3: Example posts showing the usefulness of different features for the LMD annotation. Underlined and **bold** text indicate the LMs and features, respectively.

It is worth noting here that the coarse-grained LMs are usually easy to disambiguate without exploiting any features (e.g., posts #3 and #6).

6 Benchmarking Experiments

In this section, we discuss the experimental setup and results of benchmarking IDRISI-D.

	Loc type	Event	Hashtags	URLs	Replies	Other LMs	Entities
IDRISI-DE							
Exist?	All	100.0%	63.9%	37.0%	0.4%	67.3%	31.2%
	Fine	100.0%	64.0%	34.3%	2.7%	65.5%	31.9%
	Coarse	100.0%	63.9%	37.2%	0.3%	67.7%	31.2%
Useful?	All	98.4%	32.7%	3.9%	5.0%	38.3%	5.6%
	Fine	94.0%	54.7%	28.2%	0.0%	66.9%	12.3%
	Coarse	98.8%	30.9%	2.1%	32.1%	36.0%	5.1%
IDRISI-DA							
Exist?	All	100.0%	56.6%	41.9%	27.7%	42.7%	34.8%
	Fine	100.0%	77.5%	53.5%	59.8%	74.6%	63.8%
	Coarse	100.0%	50.6%	38.4%	17.8%	32.7%	25.8%
Useful?	All	63.2%	22.2%	2.6%	0.9%	23.1%	2.0%
	Fine	89.8%	21.2%	3.6%	0.6%	19.8%	1.0%
	Coarse	54.4%	22.4%	2.0%	1.2%	24.8%	2.5%

Table 4: Statistics of the LMD features in IDRISI-D dataset.

6.1 Evaluation Setup

This section presents the learning models and the evaluation strategy we used to benchmark our IDRISI-D datasets.

6.1.1 Learning models

We train our own BERT-based models. We further employ retrieval- and heuristic-based off-the-shelf LMD baselines.

BERT$_{LMD}$: We fine-tuned the BERT-LARGE-CASED (Devlin et al., 2019) and MARBERT (Abdul-Mageed et al., 2021) models in sequence classification mode for English and Arabic LMD, respectively. To augment negative examples, we issue every gold LM against OSM and pick the top toponym that does not match it. We add only one negative example to balance the training data.

NOMINATIM (NOMIN): A search engine to search OSM data by name and address. We note that none of the existing studies compare their approaches against gazetteer search APIs (Nominatim, 2023).

GEOLOCATOR2 (GEOL2): CMU-geolocator is an off-the-shelf LMP system that considers the hierarchy of location mentions in posts when resolving them (Zhang and Gelernter, 2014).

GEOLOCATOR3 (GEOL3): An improved version of CMU-geolocator that uses the population to post-filter retrieved results from Nominatim (Zhang and Gelernter, 2014).

GEOPARSEPY (GEOPY): A trained SVM model on gazetteer-based features including location type, population, and alternative names (Middleton et al., 2018).

It is worth mentioning that GEOL and GEOPY employ NOMIN and apply post-filters on top of it. Additionally, when benchmarking IDRISI-DA, we exclude GEOPY as it is incapable of processing Arabic text. We also note that we could not employ the disaster-specific LMD models, except GEOPY, as they are nonpublic. Re-implementation is not handy due to the lack of several technical details and the unavailability of their evaluation datasets (Ji et al., 2016; Xu et al., 2019).

6.1.2 Evaluation Measures and Strategy

Inspired by the evaluation of user geolocation task (Mourad et al., 2019), we leniently evaluate LMD systems using hierarchical evaluation; however, we adopt three major changes. First, we use exhaustive locational levels including country, state, county, city, district, neighborhood, street, and POI. Second, we propagate errors from higher to lower levels. Third, we compute ranking evaluation measures, i.e., $MRR@r$ not classification or distance-based measures. In this work, we set $r = 1$,[4] but we can use different values when perceiving the task as ranking.

6.2 Results and Discussion

In this section, we benchmark IDRISI-D using off-the-shelf LMD models and our own BERT$_{LMD}$

[4]The $MRR@1$ is equivalent to the accuracy measure for classification since for every LM, we have only one correct toponym.

model. Table 5 shows the $MRR@1$ results over IDRISI-D datasets.

System	CRY	STA	CON	CTY	STR	POI
IDRISI-DA						
GEOL2	**0.45**	0.08	0.00	0.03	0.00	0.01
GEOL3	0.44	0.07	0.00	0.02	0.00	0.01
NOMIN	0.43	0.22	0.03	0.17	0.13	0.11
BERT$_{LMD}$	**0.45**	**0.49**	**0.10**	**0.34**	**0.42**	**0.28**
IDRISI-DE						
GEOL2	**0.85**	0.60	0.32	0.24	0.02	0.02
GEOL3	0.83	0.61	0.31	0.24	0.02	0.02
GEOPY	0.64	0.32	0.14	0.09	0.00	0.00
NOMIN	0.81	**0.66**	**0.38**	**0.36**	**0.24**	**0.07**
BERT$_{LMD}$	0.73	0.61	0.29	0.28	0.14	**0.07**

Table 5: The results for the LMD models on IDRISI-DE and IDRISI-DA datasets. "CRY," "STA," "CON," "CTY," "STR," and "POI" refer to COUNTRY, STATE, COUNTY, CITY, STREET, and POINT-OF-INTEREST evaluation levels, respectively

Arabic LMD: The GEOL systems show high performance at COUNTRY level. However, their performance is comparable to the BERT$_{LMD}$ model. GEOL systems fail at the fine-grained evaluation levels as they employ the GeoNames gazetteer that does not support Arabic for fine-grained locations. The NOMIN baseline is showing the best results among baselines, but it fails to outperform the BERT$_{LMD}$ at all evaluation levels.

English LMD: It is evident that the post-filters that are employed by GEOL and GEOPY are not effective for all evaluation levels, except for the COUNTRY level making the raw results from NOMIN more accurate. GEOL systems show the best results for the COUNTRY level, but their performance decreases against the BERT$_{LMD}$ model at finer evaluation levels including STATE, CITY, STREET and POI. NOMIN is the top model at almost all evaluation levels. The BERT$_{LMD}$ model managed to compete with NOMIN at only the POI evaluation level, which counts for the BERT$_{LMD}$ as the fine-grained LMs are harder to disambiguate and they are of interest to the response authorities in the disaster domain (Kropczynski et al., 2018). The results also confirm that disambiguating fine-grained LMs is more challenging than coarse-grained LMs.

7 Research Use Cases

Releasing IDRISI-D enables research on *disaster-specific* and *generic* geolocation applications that we discuss below:

Event/incident detection: While LMs indicate *where* events and incidents took place (Hu and Wang, 2021), IDRISI-D datasets with their resolved LMs could serve event/incident detect models that exploit geospatial features.

Relevance filtering: While LMs increase the likelihood of microblogs being relevant and informative with regard to the disaster events (De Albuquerque et al., 2015), IDRISI-D can enable research on relevance filtering approaches that utilize geospatial information.

Geolocation applications: While the LMP tasks play a key role in tackling all of the geolocation tasks (e.g., predicting post location (Ozdikis et al., 2019), inferring user location (Luo et al., 2020), modeling user movement (Wu et al., 2022), etc.) that employ textual features (Zheng et al., 2018), IDRISI-D is an invaluable resource for tackling all these tasks.

Geographical retrieval: The geographical information retrieval (GIR) systems are concerned with extracting spatial information alongside the relevant multimodal data to the user information need. IDRISI-D could empower the GIR retrieval techniques that rely on applying LMP tasks over queries and documents (García-Cumbreras et al., 2009).

8 Challenges

Compared to gazetteers, posts over social media contain informal language, misspellings, grammar mistakes, shortened words, and slang, causing the so-called mismatch challenge (Han et al., 2013). Table 6 presents different types of issues in the following with examples in Table 6:

Nicknames: Some places have common nicknames used by locals. For example, in Post #1, "مستشفى القديس جاورجيوس" is named "مستشفى الروم". Also, *Chennai* is nicknamed "The Detroit of India" in Post #2. The nicknames often do not exist in the gazetteers.

Abbreviations: Short names of places are prevalent on Twitter due to the character limit of posts. For example, "المملكة" (Kingdom) in Post #3 is abbreviation of المملكة العربية السعودية (Kingdom of Saudi Arabia). Also, "T. Nagar" and "GM Chetty Road" are abbreviations of "Theagaraya Nagar" and "Gopathi Narayanaswami Chetty", respectively, in Post #4.

Misspellings: Misspellings and grammar mistakes are common over Twitter. For instance,

T#	Challenge	Post text
1	Nicknames	الوضع كارثي في مستشفى الروم وهناك ضحايا في المستشفى #لبنان _ينهار #بيروت
2		#ChennaiFloods sad to see the state of city. Detroit of India is suffering. Hv personal experienced.
3	Abbreviations	نظراً للأعداد المتزايدة بالإصابة بفيروس كورونا في المملكة ... فقد أصدرت وزارة الداخلية عقوبات على كل من يخالف أوامر الحظر ...
4		Anyone around T. Nagar, needing shelter or food, can approach the Gurudwara on GM Chetty Road #Chennai
5	Misspelling	امطار حفرالباطن غريق بحي النهضه #حفرالباطن _الان ...
6		Medical students of shri ramchandra medical college in chennai stranded without supplies. Need help.
7	Shortcuts	إغلاق ط. صلاح سالم عند نفق العروبة في الاتجاهين وعند ك. الفنجري اتجاه ...
8		sm 1 help providing water 50 children @Lawrence Charitable Trust.safe.2/4,1st cross st,3rd avenue,AshokNagar-LakshmanSruti #ChennaiFloods

Table 6: Example posts illustrating the challenges of processing user-generated content for the LMD task. LMs with issues are underlined in text.

"بحي النهضه" and "حفرالباطن" in Post #5 should be written as "بحي النهضة" (with ة taa marbuta letter) and "حفر الباطن" (with space), respectively. Also, "**shri** ramchandra medical college" in Post #6 should be written as "**sri** ramchandra medical college".

Shortcuts: Users tend to use shortened words due to the character limit of posts. For example, "ط." and "ك." in Post #7 refer to "طريق." (road in English) and "كوبري." (bridge in English), respectively. Also, using "st" instead of "road", in Post #8. Also, using "@" symbol instead of the literal "at" prepositions in the same post.

Capitalization: Users tend to ignore capitalization when writing posts (e.g., "chennai" instead of "Chennai" in Post #6.

Dialectics and varieties: "كوبري." (bridge in English) in Post #7 is the dialectical (e.g., Egyptian) form of جسر in Modern Standard Arabic (MSA).

9 Conclusion

This paper contributes towards a crucial task, i.e., *Location Mention Disambiguation* in the crisis management domain. We introduced IDRISI-D, the first Arabic and the largest to date English LMD datasets.The LMD annotations that are of high reliability indicating the usefulness of the dataset. A key characteristic of IDRISI-D is the annotations of features' usefulness that we anticipate to guide the development of LMD tools. Our benchmarking results show the competitiveness of simple exact matching (NOMINATIM) and the promising performance of contextual features (BERT$_{LMD}$) for learning LMD. We release the datasets and the evaluation script for the research community. The future directions are two-fold: (i) enhancing the representation of LMs and toponyms for robust LMD learning, and (ii) employing advanced learning algorithms.

Limitations

There are a few shortcomings that we discuss below:

Twitter API Accessibility: Recently, X platform have re-envisioned its business model imposing more restrictions on the API accessibility for the research community. Although X data is extremely useful for disaster management, we expect less attention from the academic researchers to develop LMD systems that are specific for X platform. Nevertheless, IDRISI-D is invaluable resource for developing LMD systems that process user-generated content, specifically the data from microblogging platforms.

Underrepresented fine-grained LMs: Although we had chosen a careful sampling method, akin to the existing LMD datasets, the fine-grained LMs are yet underrepresented which forms a major limitation in IDRISI-D.

Temporary locations: Temporary facilities (i.e., medical camps, shelters, etc.) are constructed during emergencies to provide resources and support for the affected people. The names of these locations could change during emergencies. For example, allocating a specific school as a shelter and giving it a new expressive name (e.g., "main shelter"). Once the disaster event is over, the school will return to providing its original services. The difficulty of these temporary locations lies in their need for context when resolved. Although they are important for the affected people and response authorities, not all of them are labeled in IDRISI-D.

Ethics Statement

Although the X platform allows users to disable the geo-tagging features to protect their privacy, "even well-informed and rational individuals cannot appropriately self-manage their privacy" (Solove, 2012). There are situations where extracting geolocation data can be justified for the greater good such as during natural disasters when the focus is on saving lives and providing essential support. Therefore, any resources and tools must preserve the users' privacy and safety, especially during critical situations that could risk people's lives (e.g., conflicts and wars). Consequently, we have de-identified the data to protect users' privacy.[5] We further release the data for research purposes only under the Creative Commons Attribution 4.0 International License. Above all, we affirm that systems developed using IDRISI-D datasets must implement appropriate mechanisms to safeguard user privacy.

[5] NewTab.html

Acknowledgements

This work was made possible by the Graduate Sponsorship Research Award (GSRA) #GSRA5-1-0527-18082 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 7088–7105. https://doi.org/10.18653/v1/2021.acl-long.551

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021. HumAID: Human-Annotated Disaster Incidents Data from Twitter. In *15th International Conference on Web and Social Media (ICWSM)*. AAAI Press, Palo Alto, California, USA, 933–942.

Alaa Alharbi and Mark Lee. 2021. Kawarith: an Arabic Twitter Corpus for Crisis Events. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 42–52. https://aclanthology.org/2021.wanlp-1.5

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

Joao Porto De Albuquerque, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International journal of geographical information science* 29, 4 (2015), 667–689.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. Association for Computational Linguistics (ACL), Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Miguel Á García-Cumbreras, José M Perea-Ortega, Manuel García-Vega, and L Alfonso Ureña-López. 2009. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing & Management* 45, 5 (2009), 605–614.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical Normalization for Social Media Text. *ACM Transactions on Intelligent Systems and Technology* 4, 1 (Feb. 2013), 1–27. https://doi.org/10.1145/2414425.2414430

Yingjie Hu and Jimin Wang. 2021. How Do People Describe Locations During a Natural Disaster: An Analysis of Tweets from Hurricane Harvey. In *11th International Conference on Geographic Information Science (GIScience 2021) - Part I (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 177)*, Krzysztof Janowicz and Judith A. Verstegen (Eds.). Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 6:1–6:16. https://doi.org/10.4230/LIPIcs.GIScience.2021.I.6

Zongcheng Ji, Aixin Sun, Gao Cong, and Jialong Han. 2016. Joint Recognition and Linking of Fine-Grained Locations from Tweets. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1271–1281. https://doi.org/10.1145/2872427.2883067

Morteza Karimzadeh. 2016. Performance Evaluation Measures for Toponym Resolution. In *Proceedings of the 10th Workshop on Geographic Information Retrieval* (Burlingame, California) *(GIR '16)*. Association for Computing Machinery, New York, NY, USA, Article 8, 2 pages. https://doi.org/10.1145/3003464.3003472

Jessica Kropczynski, Rob Grace, Julien Coche, Shane Halse, Eric Obeysekare, Aurelie Montarnal, Frederick Benaben, and Andrea Tapia. 2018. Identifying Actionable Information on Social Media for Emergency Dispatch. In *ISCRAM Asia Pacific 2018: Innovating for Resilience – 1st International Conference on Information Systems for Crisis Response and Management Asia Pacific*. ISCRAM Digital Library, Wellington, New Zealand, p.428–438. https://hal-mines-albi.archives-ouvertes.fr/hal-01987793

Guoliang Li, Jun Hu, Jianhua Feng, and Kian-lee Tan. 2014. Effective location identification from microblogs. In *2014 IEEE 30th International Conference on Data Engineering*. Institute of Electrical and Electronics Engineers (IEEE), Chicago, IL, USA, 880–891. https://doi.org/10.1109/ICDE.2014.6816708

Haonan Li, Minghan Wang, Timothy Baldwin, Martin Tomko, and Maria Vasardani. 2019. UniMelb at SemEval-2019 Task 12: Multi-model combination for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 1313–1318. https://doi.org/10.18653/v1/S19-2231

Xiangyang Luo, Yaqiong Qiao, Chenliang Li, Jiangtao Ma, and Yimin Liu. 2020. An overview of microblog user geolocation methods. *Information Processing & Management* 57, 6 (2020), 102375.

Stuart E. Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Transactions on Information Systems* 36, 4 (June 2018), 1–27.

Ahmed Mourad, Falk Scholer, Walid Magdy, and Mark Sanderson. 2019. A practical guide for the effective evaluation of twitter user geolocation. *ACM Transactions on Social Computing* 2, 3 (2019), 1–23.

Nominatim. 2023. Nominatim Documentation. https://nominatim.org/release-docs/develop/

Ozer Ozdikis, Heri Ramampiaro, and Kjetil Nørvåg. 2019. Locality-adapted kernel densities of term co-occurrences for location prediction of tweets. *Information Processing & Management* 56, 4 (2019), 1280–1299.

Daniel J Solove. 2012. Introduction: Privacy self-management and the consent dilemma. *Harv. L. Rev.* 126 (2012), 1880.

Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. 2022. *Role of Geolocation Prediction in Disaster Management*. Springer Nature Singapore, Singapore, 1–33. https://doi.org/10.1007/978-981-16-8800-3_176-1

Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. 2023a. IDRISI-RE: A generalizable dataset with benchmarks for location mention recognition on disaster tweets. *Information Processing & Management* 60, 3 (2023), 103340. https://doi.org/10.1016/j.ipm.2023.103340

Reem Suwaileh, Muhammad Imran, and Tamer Elsayed. 2023b. IDRISI-RA: The First Arabic Location Mention Recognition Dataset of Disaster Tweets. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 16298–16317. https://doi.org/10.18653/v1/2023.acl-long.901

Jan Oliver Wallgrün, Morteza Karimzadeh, Alan M. MacEachren, and Scott Pezanowski. 2018. GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science* 32, 1 (2018), 1–29.

Jimin Wang and Yingjie Hu. 2019a. Are We There yet? Evaluating State-of-the-Art Neural Network Based

Geoparsers Using EUPEG as a Benchmarking Platform. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities* (Chicago, Illinois) *(GeoHumanities '19)*. Association for Computing Machinery, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/3356991.3365470

Jimin Wang and Yingjie Hu. 2019b. Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS* 23, 6 (2019), 1393–1419. https://doi.org/10.1111/tgis.12579

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*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 917–923. https://doi.org/10.18653/v1/S19-2156

Junhang Wu, Ruimin Hu, Dengshi Li, Lingfei Ren, Wenyi Hu, and Yilin Xiao. 2022. Where have you been: Dual spatiotemporal-aware user mobility modeling for missing check-in POI identification. *Information Processing & Management* 59, 5 (2022), 103030.

Canwen Xu, Jiaxin Pei, Jing Li, Chenliang Li, Xiangyang Luo, and Donghong Ji. 2019. DLocRL: A Deep Learning Pipeline for Fine-Grained Location Recognition and Linking in Tweets. In *The World Wide Web Conference, WWW 2019*. ACM, San Francisco, CA, USA, 3391–3397. https://doi.org/10.1145/3308558.3313491

Vikas Yadav, Egoitz Laparra, Ti-Tai Wang, Mihai Surdeanu, and Steven Bethard. 2019. University of Arizona at SemEval-2019 Task 12: Deep-Affix Named Entity Recognition of Geolocation Entities. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 1319–1323. https://doi.org/10.18653/v1/S19-2232

Wei Zhang and Judith Gelernter. 2014. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science* 2014, 9 (2014), 37–70.

Xin Zheng, Jialong Han, and Aixin Sun. 2018. A Survey of Location Prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (sep 2018), 1652–1671. https://doi.org/10.1109/TKDE.2018.2807840 arXiv:1705.03172v2