Are We Ready for this Disaster?
Towards Location Mention Recognition from Crisis Tweets

Reem Suwaileh[*] Muhammad Imran[†] Tamer Elsayed[*] Hassan Sajjad[†]

[*]Computer Science and Engineering Department, Qatar University
{rs081123, telsayed}@qu.edu.qa

[†]Qatar Computing Research Institute, Hamad Bin Khalifa University
{mimran, hsajjad}@hbku.edu.qa

Abstract

The widespread usage of Twitter during emergencies has provided a new opportunity and timely resource to crisis responders for various disaster management tasks. Geolocation information of pertinent tweets is crucial for gaining situational awareness and delivering aid. However, the majority of tweets do not come with geoinformation. In this work, we focus on the task of location mention recognition from crisis-related tweets. Specifically, we investigate the influence of different types of labeled training data on the performance of a BERT-based classification model. We explore several training settings such as combing in- and out-domain data from news articles and general-purpose and crisis-related tweets. Furthermore, we investigate the effect of geospatial proximity while training on near or far-away events from the target event. Using five different datasets, our extensive experiments provide answers to several critical research questions that are useful for the research community to foster research in this important direction. For example, results show that, for training a location mention recognition model, Twitter-based data is preferred over general-purpose data; and crisis-related data is preferred over general-purpose Twitter data. Furthermore, training on data from geographically-nearby disaster events to the target event boosts the performance compared to training on distant events.

1 Introduction

Twitter has shown to be an effective medium for gaining situational awareness and performing urgent needs assessment of the affected population during sudden onset disasters (Vieweg, 2012; Hughes and Palen, 2009; Purohit et al., 2018). Furthermore, the platform often breaks events and thus considered a low-latency source for timely access to information when other traditional sources are not available. Despite these advantages, one major issue that hinders the usability of Twitter data is the lack of geolocation information. Only 1-3% of tweets has GPS-coordinates (Huang and Carley, 2019). Response authorities and humanitarian organizations heavily rely on geolocation information for both situational awareness and response tasks. While extensive research has been conducted on processing tweets for humanitarian aid, limited focus has been given to infer and extract geolocation information from them.

In this work, we focus on extracting toponyms, i.e., place or location names from tweets. We refer to this as a *Location Mention Recognition* (LMR) task. Two main factors that influence the robustness of a LMR system are: (i) the dataset used to train the classifier, and (ii) the learning model. In this work, we explore how the choice of a training dataset influences the performance of a LMR system in the domain of humanitarian crises where the cost and time of acquiring training data should be minimized.

For this purpose, one well-established approach is to use a standard Named Entity Recognition (NER) system trained on a general-purpose NER dataset such as CoNLL-2003 (Sang and De Meulder, 2003). Standard NER datasets provide annotated entities such as location, organization, and person from news articles or other formal web documents. However, the general-purpose NER system may not effectively extract toponyms from Twitter messages due to the fact that tweets often contain informal language, misspellings, grammar mistakes, shortened words, and slangs (Han et al., 2013). Moreover, entities

Proceedings of the 28th International Conference on Computational Linguistics, pages 6252–6263
Barcelona, Spain (Online), December 8-13, 2020

mentioned in tweets may have inconsistent capitalization, which is one of the main features standard NER systems rely on (Zheng et al., 2018).

An alternative choice is to train a system using Twitter-specific NER datasets. Moreover, since the focus of this work is to identify and extract toponyms from crisis-related tweets, one obvious choice is to train a system specifically on location entities and drop other entity types (i.e., ORG and PER). Furthermore, we seek to determine the performance difference between a system trained on a general-purpose Twitter data versus a system trained on disaster-specific Twitter data. We also examine the difference in effectiveness when using labeled data from past disasters compared with labeled data from the current (target) disaster. As different types of disasters, such as floods and earthquakes occur in different parts of the world, we investigate whether combining labeled data from different types versus the same type help and whether having labeled data from events occurring in close-proximity versus far-away from the target event has any effect on the performance of a LMR system. Considering all these diverse settings, we formulate our research questions as follows:

- **RQ1**: How effective is the LMR system when trained on the *web-based general-purpose* NER datasets with all types of entities (LOC, ORG, PER) versus *Twitter general-purpose* datasets?
- **RQ2**: How effective are the web-based general-purpose datasets compared with Twitter general-purpose datasets when using *only location entities* (i.e., without ORG and PER)?
- **RQ3**: Does training on *crisis-related* Twitter datasets improve the performance of the LMR system compared to the general-purpose Twitter datasets?
- **RQ4**: Does training on combined data from different types of crisis events yield better performance than training on data from the same type of events?
- **RQ5**: How does the geospatial proximity of training events to the target event affect the performance?

The research on the LMR task is currently lacking answers to all those questions. In this work, we perform extensive experiments in an effort to provide answers to them. We fix our learning model to a state of the art model (i.e., BERT-based) and use a variety of datasets, i.e., web-based general-purpose, Twitter general-purpose, and Twitter crisis-specific. Our findings suggest that the general-purpose datasets are not suitable for LMR in crisis tweets. Moreover, the types of entities (e.g., PER or ORG) used to train a model make a difference. Specifically, training using only LOC entities gives better performance than using all entity types. Furthermore, while Twitter datasets are preferred over general-purpose datasets, we observe that Twitter crisis-related datasets help achieve better performance. While labeled data from the target event yield the best performance, we note that using labeled data from disasters happened in close proximity is helpful when the target labeled data is not available.

The rest of the paper is organized as follows. We summarize related work in Section 2. We present an overview of the LMR problem and define it formally in Section 3. We discuss the experimental setup in Section 4. We thoroughly analyze the results, answer the research question, and discuss the lessons we learned in Section 5. We finally conclude and list some future directions in Section 6.

2 Related Work

Several past studies focusing on the LMR task exploit different techniques and features to extract Location Mentions (LMs) from text (Zheng et al., 2018). Most of these proposed approaches are gazetteer-based in which public location gazetteers are employed such as Geonames[1] (Sankaranarayanan et al., 2009; Malmasi and Dras, 2015; Zhang and Gelernter, 2014), OpenStreetMap[2] (Malmasi and Dras, 2015), Foursquare[3] (Li and Sun, 2014; Li and Sun, 2017), Official New Zealand gazetteer[4] (Gelernter and Balaji, 2013), and Alexandria Digital Library Gazetteer[5] (Abdelkoui and Kholladi, 2017), among others.

[1] http://www.geonames.org/

[2] http://www.openstreetmap.org/

[3] https://foursquare.com/

[4] http://www.linz.govt.nz/placenames/find-names/nzgazetteer-official-names

[5] https://www.library.ucsb.edu/map-imagery-lab/alexandria-digital-library-gazetteer

Although the gazetteer-based models achieve relatively high precision as they verify the candidate LMs by matching entries occur in gazetteers, their main drawback is the inability to detect toponyms that do not appear in the gazetteers. Additionally, the mismatch between the noisy Twitter stream and non-noisy gazetteer entries is a major issue. In this work, we aim to explore the effectiveness of exploiting available in-, cross-, and out-domain training data to build LMR models that learn the patterns of location in tweets without relying on gazetteers.

To tackle the challenge of noisy Twitter stream, Sultanik and Fink (2012), used Information Retrieval based approach. They indexed gazetteers' entries by their phonetic encodings using a multidimensional binary search tree (or k-d tree, where k is the dimensionality of the search space) (Bentley, 1975). This technique mitigates the misspellings challenge efficiently. Furthermore, Li and Sun (2014) and Li and Sun (2017), constructed noisy gazetteers using cross-posts on Twitter from Foursquare check-ins. To detect LMs, they developed a linear-chain CRF model and trained it over lexical, grammatical, and geographical features.

Differently, Ghahremanlou et al. (2014) and Yin et al. (2014) retrain StandfordNER using tweet dataset, as it was originally trained on newswire articles (CoNLL-2003, MUC 6[6], and MUC 7[7]), to effectively identify the location mentions in tweets. More recently, Al-Olimat et al. (2018) proposed a statistical approach to construct regional language models. Their tagger identifies the LMs by traversing a tree of n-grams while matching them against region-specific gazetteers.

Furthermore, in 2014, the topic of the fifth Australasian Language Technology Association ALTA shared task was on identifying LMs in tweets (Molla and Karimi, 2014). Participants explored several techniques such as feature engineering, ensemble classifiers, rule-based classification, knowledge infusion, CRFs sequence labelers, semi-supervision. As for features, they used different features including geospatial, structural, and lexical. StanfordNER was also used but after retraining it using tweet dataset.

Among all the related work, there is no single study that explores the setups we propose in this paper. Typically, the proposed approaches are trained and tested on target events, assuming the training data is available at the onset of a disaster event, which is often not true. Furthermore, none of the existing studies investigate the usefulness of labeled data from past events as well as whether geospatial proximity plays any role when choosing past events for training purposes.

3 Problem Overview

At the onset of a disaster event, response organizations and first-responders relying on Twitter need geolocation information of reports or tweets about the crisis event in general as well as those seeking immediate help. In this case, the expectation is to find the mention of one or more locations in the textual content of a tweet reporting an event or asking for help. This is different than looking at the location information present in the Twitter user profile, which is often used to find the user's home location.

Table 1 shows a few tweets with different types of location mentions taken from real-world disaster events. Tweet #1 from Chennai floods in 2015 is requesting a boat to a very specific location (in this case a street name). Similarly, tweet #2 is an important situational awareness report about a bridge being collapsed. The author mentions the name of the bridge i.e., "Adayar Bridge Saidapet", which represents a very specific fine-grained location information. Tweet #3 is a situational report about casualties caused by the Christchurch earthquake in 2012. The location mention in this tweet is at the city-level. Similarly, tweet #4 reports flooding on the roads of "Ocean city, New Jersey" caused by Hurricane Sandy. We can observe that the geolocation granularity of place names mentioned in these tweets varies from coarse-grained to fine-grained. Although fine-grained locations are considered more actionable, in this work, we do not distinguish between them. We aim to recognize and extract all types of toponyms from tweets.

Accordingly, we define the **Location Mention Recognition (LMR)** task as *the automatic extraction of toponyms (i.e., places or location names) from text.* In this work, we limit the scope from two angles; we focus on *tweets*, and more specifically *crisis-related* tweets that are shared *during emergencies and natural disasters*.

[6] https://catalog.ldc.upenn.edu/LDC2003T13

[7] https://catalog.ldc.upenn.edu/LDC2001T02

Tweet #1	`[user_mention] Dear Friends, Pl help by sending boat to 54 and 58, Vivekananda Nagar Street, Nesapakkm, Chennai [...]`
Tweet #2	`[user_mention] Fear bridge being washed away. Adayar Bridge Saidapet. Hope TVK bridge is holding up fine at Malhar [url]`
Tweet #3	`65 dead in earthquake, probably more, according to John Key (prime minister) on the news #Christchurch #earthquake`
Tweet #4	`All roads into and out of Ocean City, New Jersey are closed due to flooding that has cut off the popular Jersey... [url]`

Table 1: Tweets from real-world disaster events with location mentions (highlighted)

The problem is formally defined as follows: given a tweet t that is related to a disaster event e, the LMR system aims to identify all location mentions $L_t = \{l_i; i \in [1, n_t]\}$ in the tweet t, where l_i is the i^{th} location mention and n_t is the total number of location mentions in t if any. Each location mention may span one or more *tokens*. In this work, we follow the *BILOU* annotation scheme with 5 classes[8], due to its better performance over the commonly adopted *BIO* scheme (Ratinov and Roth, 2009; Dai et al., 2015; Yang et al., 2018). In the *BIO* scheme, labels identify the position of every term in LM: "B" denotes the beginning token of an LM, "I" denotes a token inside LM, and "O" denotes a token outside of LM. The *BILOU* scheme extends the *BIO* scheme to more positional tags: "L" denotes the last token in LM and "U" denotes the only token of a single-token LM. Therefore, we define the LMR as a multi-class classification task on the token level.

4 Experimental Setup

In this section, we describe the details of our experimental setup. We present the datasets in Section 4.1 and the experimental configurations in Section 4.2. We then discuss the base LMR model in Section 4.3 followed by the evaluation measures in Section 4.4.

4.1 Datasets

To answer the research questions listed in Section 1, we mainly need three types of datasets (i) *general-purpose NER dataset* (ii) *Twitter NER dataset*, and (iii) *Crisis-related Twitter dataset*. Table 2 shows various statistics of the datasets used in our experiments, which are described below.

- **General-purpose NER dataset:** A well-known candidate for this category is the CoNLL-2003 NER dataset (Sang and De Meulder, 2003), which comprises of newswire text from Reuters, tagged with four different entity types, namely PER, LOC, ORG, and MISC. Overall, the dataset contains 22,137 sentences and 35,089 entities. We used the standard training segment for training.
- **Twitter NER dataset:** We use the Broad Twitter Corpus (BTC) as our Twitter NER dataset (Derczynski et al., 2016). It consists of 9,515 tweets, which are tagged with three entity types, namely PER, LOC, and ORG. The dataset has a broad coverage of spatial, temporal, and social aspects. Various segments in the dataset represent different types of data collection and annotation methodologies. For instance, *Segment A* comprises of random samples of UK tweets about "New Year". We used all segments for training in our experiments.
- **Crisis-related Twitter dataset:** As the main focus of this work is to guide the development of a robust LMR system for toponym extraction from crisis-related tweets, we use several Twitter datasets from real-world disasters to perform extensive experimentation. In total, we use five datasets in this category; three of them represent floods, one hurricane, and one earthquake. The floods datasets consist of 4,500 tweets from *Chennai floods 2015*, *Louisiana floods 2016*, and *Houston floods 2016* (Al-Olimat et al., 2018). The tweets in these datasets are tagged using several location-related tags. In this work, we only use inLOC and outLOC, which indicate if the location is within or outside the disaster affected areas respectively. We further filter out all tracking hashtags used to

[8] BILOU: beginning, inside, last, outside, unit

Dataset	Country	# tweets/	# locs	Annotations				
		sentences		B	I	L	U	O
CoNLL-2003	Global	22,137	7,140	1,041 (69)	116 (70)	1,041 (69)	6,099 (67)	250,660 (68)
BTC	Global	9,383	2,869	668 (100)	295 (100)	668 (100)	2,201 (100)	169,568 (100)
HRC Sandy	US	1,996	735	665 (69)	70 (66)	665 (69)	595 (70)	32,525 (70)
ChCh EQK	NZ	1,999	291	220 (68)	71 (66)	220 (68)	544 (69)	27,633 (71)
Chennai FLD	IND	1,500	2,226	840 (80)	275 (78)	840 (80)	1386 (80)	22,196 (70)
Houston FLD	US	1,500	1,701	508 (81)	155 (84)	508 (81)	1193 (81)	22,114 (70)
Louisiana FLD	US	1,500	1,396	227 (81)	77 (78)	227 (81)	1169 (81)	24,621 (69)

Table 2: Statistics of the datasets used in our experiments. The numbers in parentheses shows the percentage of training data. HRC, EQK and FLD refer to Hurricane, Earthquake, and Floods respectively.

collect the datasets, thus limiting their effect towards biasing the model. The other two datasets in this category are adopted from Middleton et al. (2013). The original source contains multilingual tweets; however, in this work, we only use English tweets. In total, 3,995 tweets from two disaster events, namely *Hurricane Sandy 2012* and *Christchurch earthquake 2012*, are used.

4.2 Experimental Configurations

We used several training and testing configurations in our experiments. In this section, we first define the adopted terminology, then discuss the different generic experimental configurations.

We define the "**source dataset**" as the dataset (or the combination of datasets) that we use to *train* our LMR model and the "**target dataset**" as the dataset on which we *test* our LMR model. The source dataset can be of any document type (e.g., web articles or tweets) and of any topic type (e.g., general or event-oriented); however the target dataset is *always* a crisis-related Twitter dataset.

Furthermore, we use different terminologies to articulate the *match* between the source and target datasets in our experiments. We use "**domain**" to refer to the domain of the target dataset, which is always of a specific disaster type. We use "**in-domain**" to denote the case when the source and target datasets are of the same disaster type, e.g., a hurricane. We use "**cross-domain**" to denote the case when the source and target datasets are both disasters *but* of different types (e.g., earthquake vs. flood). We use "**out-domain**" to denote the case when the source dataset is not a disaster dataset (e.g., general tweets or web articles).

Using the above terminologies, we define different configurations based on the source and target datasets as follows:

- <*source dataset*>`.ner` denotes the case when we use the NER source dataset with all entity types (e.g., LOC, PER, ORG, and MISC) in the BILOU scheme.
- <*source dataset*>`.loc` denotes the case when we use the NER source dataset with only the LOC entity and discard all other entity types (e.g., PER, ORG, and MISC). By doing so, we convert the LOC entity into the BILOU scheme and the non-LOC entities are labelled as "O".
- `DIS_`<*source_area*>`.others` denotes the case when the target disaster happens in a different geographical area than the *source_area*, which (in our experiments) can be either India (IND), United States (USA), or New Zealand (NZ).
- `DIS_`<*source_type*>`.others` denotes the case when the target disaster is of different type than the *source_type*, which (in our experiments) can be either Floods (FLD), Hurricane (HRC), or Earthquake (EQK).
- `DIS.others` denotes the case when the source dataset includes all disaster datasets, regardless of the type, except the target dataset. For example, if the target event is *Chennai floods*, then we use the other two floods events (i.e., *Louisiana floods* and *Houston floods*) in addition to the hurricane and the earthquake datasets for training.
- "Combined" denotes the case when we use different document types (i.e., web and tweets) in our source dataset. In this case, we use "joint" ("seq") to denote the case when we feed the different types together as one stage (sequentially in two stages) while training our model.

4.3 LMR Model

Pretrained models, such as BERT, have shown impressive performance in the sequence modeling tasks including the NER task (Devlin et al., 2018). In this work, we employ *BERT-LARGE-CASED* model in all experiments. We added a linear classification layer on top of the BERT model and finetune it using the source dataset. For training, we used the recommended setting of hyper-parameters for token classification and did not tune them as the performance is almost stable (Devlin et al., 2018). We set epochs to 3, batch size to 8, learning rate (Adam) to 2e-5, and the maximum sequence length to 128. For Twitter datasets, we preprocessed the tweets to remove 'RT', user mentions, non-ASCII characters, and URLs. We also segmented the hashtags using the word segment library[9], since some location mentions appear as subtokens of hashtags in the datasets.

4.4 Evaluation Measures

To measure the effectiveness of the LMR model over different setups, we compute Precision (P), Recall (R), and their harmonic mean (F1 score) for each entity (i.e., location mention) using the *seqeval (v0.0.12)* package,[10] which adopts the evaluation scripts used to evaluate the chunking tasks (e.g., named-entity recognition) in CoNLL-2000 NER shared task (Sang and Buchholz, 2000). The package evaluates the model's output on entity-level rather than token-level. We use the default micro-average metric to account for the class imbalance issue in our datasets (see class distributions in table 2).

5 Results and Analysis

In this section, we thoroughly discuss the results of our experiments to answer our research questions. We explore the usefulness of exploiting "out-domain" training data with either multiple entity types (such as person and organization) alongside the location entity (Section 5.1) or with location entity alone (Section 5.2). We further study the performance when training on "in-domain", "cross-domain", and "out-domain" data in Sections 5.3 and 5.4. We finally discuss the effectiveness of considering the geographic proximity of disaster events when choosing the training data in Section 5.5.

5.1 General-Purpose (Out-Domain) Training with Multiple Entities (RQ1)

Due to the limited location labeled data, we study the effectiveness of using general-purpose NER datasets to train our LMR model. We hypothesize that since general-purpose NER data is larger in size and has location as one of the entity types, using it may be sufficient to classify toponyms effectively. This is useful in emergencies where time is critical and acquiring new training data is time-consuming and expensive. The delay in response may negatively affect relief actions.

To this end, we explore the usefulness of the general-purpose NER dataset vs. Twitter NER dataset for the LMR task. We use the following training settings:

- *CoNLL*.`ner`: using the CoNLL-2003 dataset with all entity types (LOC, PER, ORG, and MISC) for training.
- *BTC*.`ner`: using the BTC dataset with all entities (LOC, PER, and ORG) for training.

We test our LMR model on each crisis-related Twitter datasets (refer to Section 4.1). Figure 1 presents the results (the second and third bars from left in all charts). In all the cases, the *BTC*.`ner` model outperforms the *CoNLL*.`ner` model, suggesting that the general-purpose datasets that are built on documents written in formal language are not suitable for tweets. To answer **RQ1**, we conclude that Twitter NER datasets are more effective than general-purpose NER datasets for training an LMR model for toponyms recognition in tweets.

5.2 General-Purpose (Out-Domain) Training with Location Entities (RQ2)

Similar to RQ1, we aim to determine the effectiveness of an LMR model trained on general-purposes (out-domain) datasets, but this time *without* non-location entities such as PER, ORG, etc. To this end, we adapt the following training settings:

[9] `http://www.grantjenks.com/docs/wordsegment/`

[10] `https://pypi.org/project/seqeval/`

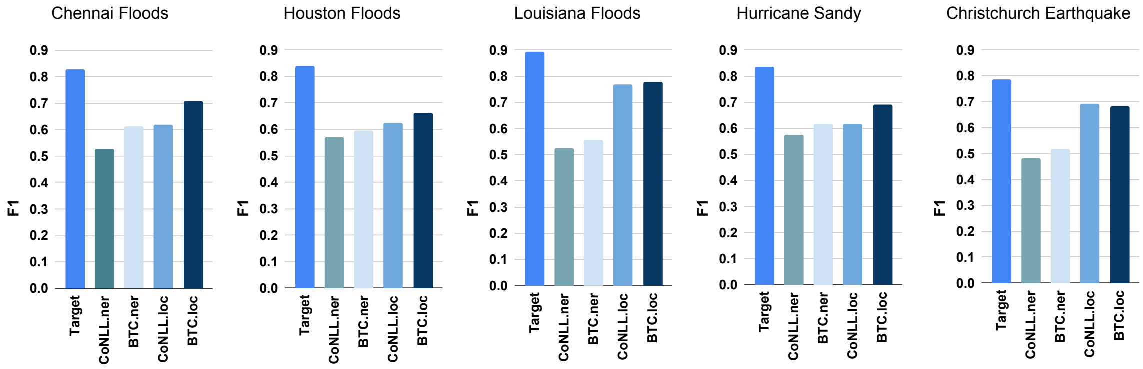

Figure 1: The results of exploiting out-domain general purpose datasets for training an LMR model

- *CoNLL*.loc: using the CoNLL-2003 dataset with only the LOC entity.
- *BTC*.loc: using the BTC dataset with only the LOC entity.

According to the results in Figure 1 (considering also the fourth and fifth bars from left in all charts), training the LMR model using only LOC entity improves the performance by 3.5-23.8% and 5.2-23.2% across the different disasters for CoNLL and BTC, respectively. We noticed that the improvement is clearly evident in precision but not recall (refer to results in Table 3 in Appendix.A), suggesting that focusing the training on locations only significantly improves the precision of recognizing locations with little or no degradation in recall. We anticipate the reason to be the distinct patterns of LMs compared to other entities in the data. For instance, different from other types of entities, location mentions are usually attached to their category (e.g., LOC street, LOC city, etc.) or surrounded by adpositions such as "near", "at", or "10Km away from". Additionally, the modeling of LOC only maps the problem to a binary classification problem, which is an easier task than a multiclass classification task.

To summarize, this result answers **RQ2**, i.e., the location-specific datasets are better for training the LMR model compared to the general-purpose NER datasets.

5.3 Crisis-Related Training (RQ3)

Thus far, we confirmed our need for location-specific data to train the LMR system. However, the location mentions in the general stream, in contrast to disaster-specific streams, might appear in different patterns. To clarify, people might tend to use more accurate and full addresses of locations when reporting incidents happening during emergencies, aiming to help responders make immediate actions. To investigate further, we train using a combination of *BTC*.loc dataset (as using it achieved the best F1 score earlier) and the available crisis-related datasets. By this, we aim to address **RQ3**: *Does training on crisis-related Twitter datasets improve the performance of the LMR system compared to the general-purpose Twitter datasets?*

An interesting aspect to explore in this context is the effect of combining the in-, cross-, and out-domain data. To address this, we train an LMR model using crisis-related datasets as follows.

- DIS.others: combining all disaster datasets except the target disaster for training.
- *Combined*.joint: combining in-, cross-, and out-domain datasets for training. Specifically, we use *BTC*.loc and all DIS.others for training. All the datasets are merged before training.
- *Combined*.seq: using *BTC*.loc and DIS.others for training; however, we first train a model using the former and then fine-tune it using the latter.

We show the results of these runs in Figure 2. Generally, the results are not consistent across disasters, hence we cannot draw a clear conclusion on which setup is clearly the best. As references, we compare the results with the case when we train on the target dataset (denoted as Target in Figure 2), and with *BTC*.loc (as using it achieved the best F1 among the non-target data). It is evident that using training data other than the target data shows significant degradation in performance with respect to the Target

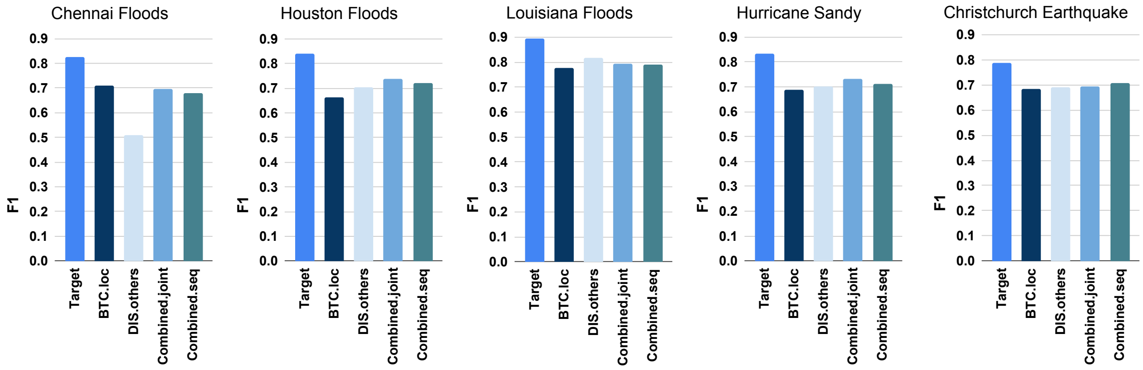

Figure 2: The F1 results of exploiting in- and out-domain data for training an LMR model

model. This finding emphasizes the importance of providing in-domain (i.e., "target") data to achieve better effectiveness. Additionally, employing only in- and cross-domain data (i.e., "DIS.others") shows comparable results to *BTC*.`loc`, except for the Chennai floods. These results confirm the potential of using in- and cross-domain data for a better performance.

Moreover, combining in-, cross-, and out-domain training data provides reasonable performance for early location extraction when a sudden disaster happens. In a worst scenario, such a reasonable model can be employed to automatically augment labeled data to improve the performance over time. This can be achieved by exploiting active learning, automatic labeling, among other known data augmentation techniques. Furthermore, the *Combined*.`joint` setup is better than the *Combined*.`seq` setup by approximately 1.5% on average across all datasets, except for the Christchurch earthquake. Upon investigation, we found the hurricane and earthquake datasets (Middleton et al., 2013) suffer from the missing-locations issue (Middleton et al., 2018). Furthermore, in the hurricane dataset, we found 15.3% of the unique locations that appear in the test data do not exist in the training set. Similarly, in the flood datasets, we found 15.2%, 12.8%, and 7.02% of the total unique locations appear in the test data but they do not appear in training datasets of Chennai, Houston, and Louisiana, respectively. These reasons probably led to a degraded performance of the "seq" model when using such datasets for subsequent fine-tuning of the models trained on *BTC*.`loc`, which is the case in the three floods target datasets. That negative effect is a bit mitigated when merging them with *BTC*.`loc` in the "joint" models.

5.4 Cross-Domain Training (RQ4)

In contrary to our expectation, using disaster-related training data does not improve the LMR model significantly. We anticipate the problem to be the difference in disaster types that we employed for training. Consequently, we study the effect of training on "cross-domain" data, i.e., training on data from previous disasters but of a different type than the target, compared to the case when both the source and target disasters are of the same type. In this section, we address **RQ4**: *Is training on combined data from different types of crisis events (cross-domain) better than training on data from the same type of events (in-domain)?* To this end, we use the following training settings:

- `DIS_FLD.others`: using data from all flood events for training and testing on other disasters (in this case, other disasters are of type FLD, HRC, and EQK).
- `DIS_HRC.others`: using data from the hurricane event for training and testing on other disasters (in this case, other disasters are of type FLD and EQK).
- `DIS_EQK.others`: using data from the earthquake event for training and testing on other disasters (in this case, other disasters are of type FLD and HRC).

Figure 3 shows the results. The missing bars in the case of Hurricane Sandy and Christchurch earthquake are due to the fact that we only have one hurricane and one earthquake events.

Looking at the results when the target type is floods (the first three sub-figures), training on disasters of the same type as the target (FLD) achieves better performance compared to training on HRC and

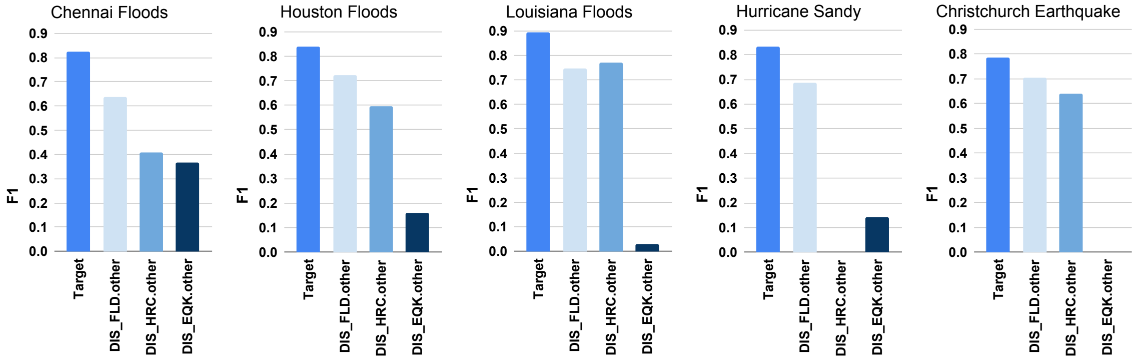

Figure 3: The F1 results of training on cross-domain data. Missing bars means there is no more one disaster dataset of target type.

EQK data, except for the Louisiana floods. We suspect the reason to be the close geographical proximity between the affected areas of Hurricane Sandy and Louisiana Floods which enhances the model's ability to detect more LMs.

We also notice that training on EQK data in consistently the worst across all disasters. Upon investigation, we found that the location mention "Christchurch" constitutes 262 out of 527 locations in the training data, and 84 out of 156 locations in the test data. Moreover, 68% of the tweets constituting this dataset have no locations. For this reason, we believe that this dataset is inadequate for training compared to other datasets. To further understand these results, next we explore how the geospatial proximity of source events to the target event affects the performance.

5.5 Geo Proximity-based Training (RQ5)

The location mentions from within the affected areas of a target disaster are expected to emerge in the tweets stream over time. However, such locations may not be seen by LMR models trained on past disasters data. We anticipate that employing an LMR model trained on the closer geographical area as the target disaster (within the same country in our experiments) can alleviate this issue. A concrete example of this is the case of Louisiana floods when trained on hurricane Sandy data (refer to previous section). To elaborate, not all countries exhibit the same naming formats (e.g., using street numbers in contrast to names) and administrative levels (e.g., states, counties, etc.). In this section, we address **RQ5**: *How does the geospatial proximity of source events to the target event affect the performance?*

To address this question, we use the following training settings:

- DIS_US.other: combining all events from USA except the target for training. For example, if the target disaster is Hurricane Sandy, we train on Houston and Louisiana floods.
- DIS_IN.FLD: training on Chennai Floods happened in India.
- DIS_NZ.EQK: training on Christchurch Earthquake happened in New Zealand.

Due to the lack of diverse disaster-specific labeled data for the LMR task, we could conduct experiments only on target datasets of disasters that happened in the US; for other areas (NZ and IN), we do not have more than one disaster-specific dataset. Nonetheless, the results in Figure 4 indicate that training on source disasters happened in close proximity areas (with respect to the target event) achieve the best performance regardless of the type of disaster.

6 Conclusion

This work contributes towards a crucial task, i.e., *Location Mention Recognition* from crisis-related tweets. We formulated several research questions for which evidence-based answers were unknown. We designed an extensive experimental setup where several experiments investigate the effectiveness of training on general-purpose NER datasets from news articles and tweets. We demonstrate how the performance of a LMR model varies when trained on formal language (new articles) compared to informal

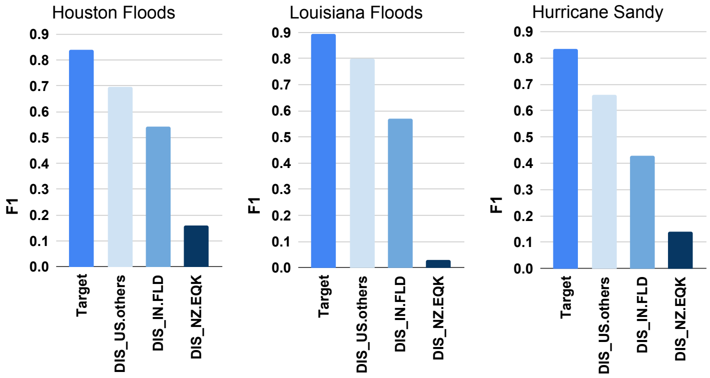

Figure 4: The F1 results of training on geo-proximity-based data.

language (tweets) as well as when trained on past disasters. Our findings suggest that Twitter-based NER labeled data is preferred over general-purpose data; and crisis-related labeled data is preferred over general-purpose Twitter data. Furthermore, our results suggest that training on disaster events data from geographically-nearby events to the target event boosts the performance compared to training on distant events. Overall, we remark that our findings will help shape future directions in this line of research.

We consider multiple future directions. We plan to extend our study to other languages (e.g., Arabic and Italian). We plan to study the cost of acquiring labeled data over time, i.e., during an emergent disaster, using incremental training. Additionally, our current usage of BERT is only for representing text; we plan to explore the effect of fine-tuning it, modifying the classification layer, trying other learning models, and applying advanced domain adaptation and transfer learning techniques.

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

Feriel Abdelkoui and Mohamed-Khireddine Kholladi. 2017. Extracting criminal-related events from Arabic tweets: A spatio-temporal approach. *Journal of Information Technology Research (JITR)*, 10(3):34–47.

Hussein S. Al-Olimat, Krishnaprasad Thirunarayan, Valerie Shalin, and Amit Sheth. 2018. Location name extraction from targeted text streams using gazetteer-based statistical language models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1986–1997. Association for Computational Linguistics.

Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

Hong-Jie Dai, Po-Ting Lai, Yung-Chun Chang, and Richard Tzong-Han Tsai. 2015. Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *Journal of cheminformatics*, 7(S1):S14.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter Corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Judith Gelernter and Shilpa Balaji. 2013. An algorithm for local geoparsing of microtext. *GeoInformatica*, 17(4):635–667.

Lida Ghahremanlou, Wanita Sherchan, and James A. Thom. 2014. Geotagging Twitter messages in crisis management. *Computer Journal*, pages 1937–1954.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):1–27.

Binxuan Huang and Kathleen M Carley. 2019. 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.

Amanda Lee Hughes and Leysia Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International journal of emergency management*, 6(3-4):248–260.

Chenliang Li and Aixin Sun. 2014. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 43–52. ACM.

Chenliang Li and Aixin Sun. 2017. Extracting fine-grained location with temporal awareness in tweets: A two-stage approach. *Journal of the Association for Information Science and Technology*, 68(7):1652–1670.

Shervin Malmasi and Mark Dras. 2015. Location mention detection in tweets and microblogs. In *Conference of the Pacific Association for Computational Linguistics*, pages 123–134. Springer.

Stuart E Middleton, Lee Middleton, and Stefano Modafferi. 2013. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17.

Stuart E. Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. Technical Report 4.

Diego Molla and Sarvnaz Karimi. 2014. Overview of the 2014 ALTA shared task: Identifying expressions of locations in tweets. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pages 151–156.

Hemant Purohit, Carlos Castillo, Muhammad Imran, and Rahul Pandev. 2018. *Social-EOC: Serviceability model to rank social media requests for emergency operation centers*. ASONAM.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning*, pages 147–155. Association for Computational Linguistics.

Erik F Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. 2009. Twitterstand: News in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM.

Evan A Sultanik and Clayton Fink. 2012. Rapid geotagging and disambiguation of social media text via an indexed gazetteer. *Proceedings of ISCRAM*, 12:1–10.

Sarah Elizabeth Vieweg. 2012. *Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications*. Ph.D. thesis, University of Colorado at Boulder.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. *arXiv preprint arXiv:1806.04470*.

Jie Yin, Sarvnaz Karimi, and John Lingad. 2014. Pinpointing locational focus in microblogs. In *Proceedings of the 2014 Australasian document computing symposium*, page 66. ACM.

Wei Zhang and Judith Gelernter. 2014. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):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):1652–1671, sep.

Appendix A. Full Results

Detailed results, including out-domain training (Table 3), in/out-domain training (Table 4), cross-domain training (Table 5), and training based on geo-proximity of events (Table 6).

	Chennai Floods			Houston Floods			Louisiana Floods			Hurricane Sandy			ChCh Earthquake		
Source Data	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Target	0.80	0.86	0.83	0.82	0.86	0.84	0.90	0.89	0.89	0.81	0.86	0.83	0.76	0.81	0.79
CoNLL.ner	0.52	0.54	0.53	0.57	0.57	0.57	0.41	0.73	0.53	0.47	0.73	0.57	0.36	0.72	0.48
BTC.ner	0.56	0.68	0.61	0.60	0.59	0.60	0.45	0.74	0.56	0.52	0.76	0.62	0.39	0.76	0.52
CoNLL.loc	0.73	0.54	0.62	0.81	0.51	0.62	0.88	0.68	0.77	0.59	0.64	0.62	0.68	0.71	**0.69**
BTC.loc	0.76	0.66	**0.71**	0.76	0.59	**0.66**	0.84	0.73	**0.78**	0.64	0.75	**0.69**	0.63	0.75	0.68

Table 3: Out-domain general purpose training results. Best F1 scores of non-target training setups are boldfaced.

	Chennai Floods			Houston Floods			Louisiana Floods			Hurricane Sandy			ChCh Earthquake		
Source Data	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Target	0.80	0.86	0.83	0.82	0.86	0.84	0.90	0.89	0.89	0.81	0.86	0.83	0.76	0.81	0.79
BTC.loc	0.76	0.66	0.71	0.76	0.59	0.66	0.84	0.73	0.78	0.64	0.75	0.69	0.63	0.75	0.68
DIS.others	0.69	0.40	0.51	0.78	0.64	0.70	0.86	0.78	**0.82**	0.66	0.74	0.70	0.62	0.78	0.69
Combined.joint	0.76	0.64	**0.70**	0.82	0.67	**0.74**	0.84	0.75	0.79	0.65	0.84	**0.73**	0.63	0.78	0.70
Combined.seq	0.80	0.59	0.68	0.80	0.66	0.72	0.86	0.73	0.79	0.67	0.76	0.71	0.63	0.81	**0.71**

Table 4: In & out-domain training results. Best F1 scores of non-target training setups are boldfaced.

	Chennai Floods			Houston Floods			Louisiana Floods			Hurricane Sandy			ChCh Earthquake		
Source Data	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Target	0.80	0.86	0.83	0.82	0.86	0.84	0.90	0.89	0.89	0.81	0.86	0.83	0.76	0.81	0.79
DIS.FLD.others	0.67	0.61	**0.64**	0.79	0.67	**0.72**	0.71	0.79	0.75	0.66	0.72	**0.69**	0.62	0.81	**0.70**
DIS.HRC.others	0.62	0.31	0.41	0.72	0.51	0.59	0.85	0.70	**0.77**	-	-	-	0.64	0.65	0.64
DIS.EQK.others	0.54	0.28	0.37	0.48	0.10	0.16	0.36	0.01	0.03	0.39	0.09	0.14	-	-	-

Table 5: Cross-domain training results. Best F1 scores of non-target training setups are boldfaced. Missing values mean when there is one target event type.

	Houston Floods			Lousiana Floods			Hurricane Sandy		
Source Data	P	R	F1	P	R	F1	P	R	F1
Target	0.82	0.86	0.84	0.90	0.89	0.89	0.81	0.86	0.83
DIS_US.others	0.79	0.62	**0.70**	0.86	0.75	**0.80**	0.62	0.70	**0.66**
DIS_IN.FLD	0.75	0.43	0.54	0.80	0.44	0.57	0.48	0.39	0.43
DIS_NZ.EQK	0.48	0.10	0.16	0.36	0.01	0.03	0.39	0.09	0.14

Table 6: The results of training based on geo-proximity. Best F1 scores of non-target training setups are boldfaced.