Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7164–7173
Marseille, 20-25 June 2022

Identification of Fine-Grained Location Mentions in Crisis Tweets

Sarthak Khanal, Maria Traskowsky, Doina Caragea
Kansas State University
Manhattan, KS 66506
{sarthakk, mariatraskowsky, dcaragea}@ksu.edu

Abstract

Identification of fine-grained location mentions in crisis tweets is central in transforming situational awareness information extracted from social media into actionable information. Most prior works have focused on identifying generic locations, without considering their specific types. To facilitate progress on the fine-grained location identification task, we assemble two English tweet crisis datasets and manually annotate them with specific location types. The first dataset contains tweets from a mixed set of crisis events, while the second dataset contains tweets from the global COVID-19 pandemic. We investigate the performance of state-of-the-art deep learning models for sequence tagging on these datasets, in both in-domain and cross-domain settings.

Keywords: Crisis management, Social-media, Location identification, Deep-Learning

1. Introduction

We have witnessed a large number of crisis situations in recent years, from natural disasters to man-made disasters and also to deadly animal and human health crises, culminating with the ongoing COVID-19 public health crisis. Affected individuals often turn to social media (e.g., Twitter or Facebook) to report useful information, or ask for help (Sakaki et al., 2010; Vieweg et al., 2010; King, 2018). Information contributed on social media by people on the ground can be invaluable to emergency response organizations in terms of gaining situational awareness, prioritizing resources to best assist the affected population, addressing concerns, and even saving lives (King, 2018).

Many recent studies have focused on identifying informative tweets posted by individuals affected by a crisis, and classifying those tweets according to situational awareness categories useful for crisis response and management (Imran et al., 2015). However, for situational awareness information extracted from social media to be actionable, knowing the corresponding geographic location is of key importance. For example, location information enables responders to perform fast assessment of the damage produced by a natural disaster (Villegas et al., 2018), or to respond to requests for help coming from affected individuals or institutions (e.g., hospitals or schools). In the case of COVID-19 health crisis, location information can also be used to identify trends by locations (e.g., stance of a community towards various health recommendations) (Mutlu et al., 2020; Miao et al., 2020), and subsequently employ that information to prevent dissemination of misinformation and rumors, and resurgence of the novel coronavirus.

Unfortunately, only a very small percentage of tweets are geotagged (Mahmud et al., 2012). Furthermore, even when geolocation information is available, that location may not be the location mentioned in the tweet text (Ikawa et al., 2013). According to Vieweg et al. (2010), the location in the tweet text is usually the location needed for monitoring and/or responding to an emergency. Table 1 shows several examples of tweets posted during recent hurricanes (first three tweets) and during the COVID-19 crisis (last three tweets). As can be seen, locations are mentioned at different levels of granularity, from region and landmark to city, state and country. Furthermore, the same location name, in our COVID-19 examples - *New York*, can be associated with different location types, such as *city* (tweet 4) and *state* (tweet 6). Information about the tags of the ambiguous entities can be used to disambiguate the corresponding locations and link them to physical locations. Therefore, tools for identifying fine-grained locations directly from the texts of crisis tweets are greatly needed.

Location identification has been frequently addressed as part of the broader named entity recognition (NER) task (Goyal et al., 2018; Li et al., 2020). Some studies have focused specifically on the task of identifying generic location mentions (without considering the type of location) in tweet text (Hoang et al., 2017), and even disaster tweet text (Kumar and Singh, 2019). Other studies have focused on identifying fine-grained points-of-interest (POI), useful for location-based services (Li and Sun, 2014; Malmasi and Dras, 2015; Ji et al., 2016; Xu et al., 2019).

To the best of our knowledge, there are no publicly available, manually annotated datasets that can facilitate progress on the task of identifying fine-grained locations (including, city, state, country, region, landmark) in crisis tweets, despite the benefits provided by the use of social media data in monitoring and responding to a crisis. To address this need, we have assembled two datasets for identifying fine-grained locations in crisis tweets. The first dataset, called *MIXED*, consists of tweets crawled during five crisis events, specif-

No.	Tweet text
1	Roads in Calhoun County are underwater, access to the Port Lavaca Causeway is flooded, the bridge is closed.
	O O B-ctc I-ctc O O O O O B-lan I-lan I-lan O O O O O O
2	Very extensive damage sustained throughout Wilmington NC from Hurricane Florence
	O O O O O B-ctc B-sta O O O
3	Big tree fell on power lines and blocking Brown Ave near Washington St in Orlando s Thornton Park
	O O O O O O O O B-oth I-oth O B-oth I-oth O B-sta O B-lan I-lan
4	There are now more confirmed cases of coronavirus in New York City than there are in all of South Korea
	O O O O O O O O O B-ctc I-ctc I-ctc O O O O O O B-con I-con
5	South Asia is quickly marching towards being the new epicenter of covid 19
	B-reg I-reg O O O O O O O O O O O
6	The difference in COVID 19 cases and deaths between New York and California continues to be astounding
	O O O O O O O O O B-sta I-sta O B-sta O O O O

Table 1: Examples of crisis tweets tagged with fine-grained location types. The subsequences representing location mentions are highlighted with pink, and their corresponding tags (in BIO format) are highlighted with blue.

ically, Nepal Earthquake, Queensland Floods, Srilanka Bombing, Hurricane Michael and Hurricane Florence. The second dataset, called *COVID*, consists of a set of coronavirus-related tweets crawled between February 27 and April 7, 2020. We used Amazon Mechanical Turk (AMT)[1] to annotate the datasets using six location types (Country, State, Region, City, Landmark and Others).

Given the success of deep learning approaches for NER tasks (Li et al., 2020), we use different state-of-the-art models to establish baseline results on the dataset. In summary, the contributions of this work are as follows:

- We create two datasets of tweets from a mixed set of crisis events and from COVID-19, respectively. The tweets are manually annotated with fine-grained location types, including city, state, country, region, landmark. We make this data publicly available along with annotation tool and preprocessing script.[2] The final annotated datasets are provided in the form of tweet-id and corresponding location annotations in that tweet.

- We use state-of-the-art models including a contextual encoder coupled with a tag decoder in a multi-task learning setting, and a model based on contextualized word and entity representations, combined with entity-aware self-attention to establish baseline results for our datasets.

- We perform extensive experiments on the *MIXED* and *COVID* datasets, respectively, in both *in-domain* and *cross-domain settings* to understand the usefulness of the data from the domain of interest, as well as the transferability of the models from one domain to another.

[1] https://www.mturk.com/

[2] https://github.com/sarthakksu/finegrained-location-data

Given this introduction, we proceed with a discussion of related work in the next section, followed by the description of the datasets constructed, and then background and approaches, experimental setup, results and error analysis, and finally, conclusions and an ethics statement.

2. Related Work

We organize the related work based on several categories relevant to the research in this paper. Specifically, we first briefly discuss the location mention identification as a specific task in the area of NER. Subsequently, we review works on fine-grained location types, followed by approaches used for identifying locations, and finally, other existing and relevant location datasets.

2.1. NER and location mention identification

NER is a well-researched problem in natural language processing (NLP) (Goyal et al., 2018; Li et al., 2020). Text-based location identification has been traditionally addressed as part of the broader NER task, although some works focus specifically on location identification (Lingad et al., 2013; Han et al., 2014; Kumar and Singh, 2019; Magge et al., 2019). Most of the works that identify locations simply tag location mentions, as opposed to identifying fine-grained location types (Li et al., 2020). For example, Lingad et al. (2013) aim to identify mentions of locations (including geographic locations and points of interest) in disaster tweets, by using standard NER taggers (pre-trained or retrained), and report best performance using retrained Stanford NER (Finkel et al., 2005). Also in the context of emergencies, Kumar and Singh (2019) use a convolutional neural network (CNN) approach to identify location references in crisis tweets, regardless of their specific types.

Dataset	Event	Keywords	Size
MIXED	Nepal Earthquake and Queensland Floods (Alam et al., 2018a)	N/A	167
	Srilanka Bombing (ours)	Sri lanka attack, Sri lanka terror, Sri Lanka horror, Sri Lanka easter	1171
	Hurricane Michael and Hurricane Florence (ours)	hurricane michael, hurricanemichael, hurricane florence, hurricaneflorence	2758
COVID	COVID-19 (ours)	#coronavirus, corona virus, #Coronavid19, #coronavirususa, #covid19, #covid-19, #quarantinelife, #socialdistancing	5243

Table 2: Keywords used to collect tweets and the number of tweets from each event in the *MIXED* and *COVID* datasets.

2.2. Fine-grained location types

Some recent works have considered fine-grained location types, such as city, state, country (Inkpen et al., 2015; Anand et al., 2017; Lal and others, 2019; Qazi et al., 2020). While focused on COVID-19 tweets, Qazi et al. (2020) use a gazetteer approach to infer the geolocation of tweets, based on user and tweet information. Closest to our goal of identifying fine-grained locations in disaster tweet texts, Inkpen et al. (2015) propose a CRF-based approach to identify countries, states/provinces and cities using a Twitter dataset annotated according to guidelines provided in (Mani et al., 2010). They make use of hand-crafted features, including gazetteer features, to train a CRF model. As opposed to (Inkpen et al., 2015), we use a larger set of location types and approaches that preclude the need for manually crafted features and gazetteers.

Other works on fine-grained location focus on identifying point of interests locations, such as restaurants, hotels, parks, etc. and linking them to pre-defined location profiles (Li and Sun, 2014; Ji et al., 2016; Xu et al., 2019). Li and Sun (2014) build a point-of-interest (POI) inventory (which can be seen as a noisy version of a gazetteer), and a time-aware POI tagger. The time-aware POI tagger is a CRF trained to extract and disambiguate fine-grained POIs. Ji et al. (2016) extend the POI tagger in Li and Sun (2014) by proposing a joint framework that achieves POI recognition and linking to pre-defined POI profiles simultaneously. Xu et al. (2019) address the same problem of identifying fine-grained POIs and linking them to location profiles. However, they use a deep learning model (specifically, BiLSTM-CRF) to avoid the need for manually designed features, and subsequently use a collection of location profiles to perform the linking. The definition of fine-grained POI tagging is different from our definition of fine-grained location tagging - we aim to assign specific types/tags to location entities, as opposed to identifying generic (yes/no) POI tags, and then linking the tags to pre-defined profiles, as in prior works (Li and Sun, 2014; Ji et al., 2016; Xu et al., 2019). Moreover, we want to avoid the use of gazetteers to ensure that the models are resilient to the informal nature of the language used in tweets. Similar to (Xu et al., 2019), we also want to avoid the need for manually designed features, and thus focus on deep learning approaches.

2.3. State-of-the-art approaches for NER

State-of-the-art approaches for NER, in general, and location identification, in particular, are sequence labeling type approaches based on deep learning language models (Li et al., 2020). More specifically, competitive architectures consist of three components: distributed representations of the input, a context encoder model, and a tag decoder model. Both character-level and word-level embeddings (or their combination) have been used to represent the NER input in recent works (Goyal et al., 2018), with BERT (Devlin et al., 2018) contextual embeddings being among the most successful (Li et al., 2020). In terms of context encoders and tag decoders, recurrent neural networks, most often, BiLSTM networks (short for Bidirectional Long Short-Term Memory) (Hochreiter and Schmidhuber, 1997), and CRF (short for Conditional Random Fields) (Lafferty et al., 2001), respectively, contribute to some of the best results on benchmark NER datasets (Luo et al., 2019; Baevski et al., 2019; Liu et al., 2019; Jiang et al., 2019). Given these successful architectures for the NER task, one of our baseline models consists of three components: BERT, BiLSTM and CRF, for the input representation, context encoder and tag decoder, respectively. As another strong baseline, we investigate a recent state-of-the-art architecture, called LUKE, (Yamada et al., 2020), based on a bidirectional transformer architecture pre-trained to output both word and entity contextualized representations. LUKE uses an entity-aware self-attention to identify entities.

2.4. Existing location datasets

Most previous works on location identification in tweet texts are focused on general tweets (Liu et al., 2014; Inkpen et al., 2015) with a few notable exceptions of works focused on crisis tweets (Lingad et al., 2013; Kumar and Singh, 2019; Qazi et al., 2020). However, the datasets used in these works are not all available (Lingad et al., 2013; Kumar and Singh, 2019). Even when available, the datasets focus on identifying location mentions without specifically identifying the fine-grained type of the location mentions (Liu et al., 2014).

Qazi et al. (2020) used a gazetteer-only approach to annotate tweets with geolocations, and the resulting annotations are not very accurate. While not specifically focused on crisis tweets, the dataset published by (Inkpen et al., 2015) is the closest to our dataset in terms of fine-grained location types used (which include city, country, state or province, etc.). However, most locations in their dataset are not mentioned in the tweet, but are inferred from auxiliary information. Specifically, only about 3% of the tweet texts in their dataset have location entities, for a total of only 220 different location entities. Furthermore, they also used a gazetteer approach to annotate most of the tweets, and performed manual annotations just for a small subset of their dataset. Given the above-mentioned differences between existing datasets and our datasets, it is not possible to directly use the existing datasets to transfer information to our tasks in a cross-domain setting.

3. Datasets

One main contribution of our work is to construct two benchmark datasets for identifying fine-grained locations (see Table 3) useful for crisis monitoring and response. The datasets cover events that are different in nature, to enable studies in both in-domain and cross-domain settings.

3.1. Data collection

The first dataset, called *MIXED*, contains tweets posted during four natural disasters and one man-made disaster that happened in specific geographical regions. The second dataset, called *COVID*, contains tweets posted during the COVID-19 pandemic, and thus has worldwide coverage. More specifically, the tweets in the *MIXED* dataset were crawled during the following events: Nepal Earthquake, Queensland Floods, Srilanka Bombing, Hurricane Michael and Hurricane Florence. The tweets from Nepal Earthquake and Queensland Floods were obtained from (Alam et al., 2018a). Tweets from Srilanka Bombing, Hurricane Michael and Hurricane Florence were crawled locally using the Twitter streaming API. A random sample of unique English tweets was included in the *MIXED* dataset that was annotated using AMT. More than 133 million tweets from COVID-19 pandemic were also crawled locally between February 27th and April 7th, 2020. A random sample of unique English tweets was included in the *COVID* dataset for AMT annotation. The keywords used to crawl the tweets and the final number of tweets included in the dataset for each event are listed in Table 2.

In addition to the *MIXED* and *COVID* datasets that are annotated as part of this work, we also used a large number of unlabeled mixed crisis and COVID-19 tweets to further pre-train the BERT-base-uncased (Devlin et al., 2018) models and obtained crisis-specific embeddings. In particular, to further pre-train the BERT-base-uncased model for the *MIXED* dataset, we collected a larger set of tweets pertaining to various crisis events from prior works (Imran et al., 2016; Nguyen et al., 2017; Alam et al., 2018b; Alam et al., 2018a; Olteanu et al., 2014; Olteanu et al., 2015) in addition to the locally crawled tweets. For the *COVID* dataset, however, we only used the locally crawled tweets to pre-train the existing BERT-base-uncased model.

Type	Descr.	MIXED Distr.		COVID Distr.	
		#	%	#	%
con	Country	1,763	28.31	1,819	53.06
sta	State	1,242	19.95	396	11.56
reg	Region, Continent	764	12.27	158	4.61
ctc	City, Town, County	797	12.80	518	15.11
lan	Building, Landmark	1,190	19.11	391	11.42
oth	Other	471	7.56	146	4.24
All	Entities	6,227	100.00	3,428	100.00

Table 3: Location types and their descriptions, together with type distribution (as raw numbers # and percentages %) in the *MIXED* and *COVID* datasets, respectively.

3.2. Data annotation and quality assessment

To prepare the tweets for annotation, the following pre-processing was performed. User mentions were anonymized by replacing them with a generic `user` keyword, and links were removed from the tweet text. Special characters, including -= !#$% ^ &*()+[]{};\':"|<>?, and non-printable ASCII characters were also removed. The tweet text was tokenized to enforce annotation at the token level and avoid accidental annotation of token fragments. Tweet tokens were annotated with six location types using the BIO scheme (where B stands for Beginning, I stands for Inside and O stands for Outside of a location entity). The location types together with their brief descriptions are shown in Table 3. Examples of annotated tweets are shown in Table 1, where the first three tweets are representative of the *MIXED* dataset, and the last three are representative *COVID*.

We used feedback from a local annotator to iteratively develop and improve a custom annotation tool for our task. The tool was subsequently deployed to AMT. Annotators were provided with definitions of the location types included in our study, together with precise instructions for annotation, and examples of annotated tweets, such as those in Table 1. We selected external AMT workers based on agreement with our local annotator on a small subset of tweets (approximately 500 tweets). Subsequently, each tweet was annotated by 3 external AMT workers. Only entities where two or more annotators agreed were included in the final datasets. The Cohen's Kappa scores that we obtained for inter-annotator agreement were 0.63 and 0.62, and the average pairwise F1-scores for inter-annotator agreement were 68.87 and 65.86 for the *MIXED* and *COVID* datasets, respectively. According to Cohen (1960), these scores represent substantial

Dataset	No.	Train	Test	Dev	Total
MIXED	Tweets	2,620	820	656	4,096
	Tokens	73,622	23,253	18,511	115,386
	Entities	4,001	1,237	989	6,227
	Entity type distribution				
	con	1,135	360	268	1,763
	sta	752	267	223	1,242
	reg	514	144	106	764
	ctc	522	141	134	797
	lan	768	238	184	1,190
	oth	310	87	74	471
COVID	Tweets	3,355	1,049	839	5,243
	Tokens	103,646	32,674	25,798	162,118
	Entities	2,206	656	566	3,428
	Entity type distribution				
	con	1,162	347	310	1,819
	sta	264	75	57	396
	reg	101	34	23	158
	ctc	328	103	87	518
	lan	265	62	64	391
	oth	86	35	27	146

Table 4: Statistics for the number of tweets, tokens and the number of location entities in the `train/test/dev` subsets of the *MIXED* and *COVID* datasets, respectively. The entity type distribution in the `train/test/dev` subsets is also shown for each dataset.

agreement.

The distributions of the location entities over the six location types included in our study are shown in Table 3. As can be seen, the annotated entities are more evenly distributed over the types considered in the *MIXED* dataset, while more than half of the entities are of type *country* in the *COVID* dataset. The datasets also show differences in terms of the number of entities per tweet, with the *MIXED* dataset containing a majority of tweets with one or two entities (and a small number of tweets with more than two entities), and *COVID* containing mostly tweets with one entity (and a small number of tweets with two or more entities). Such differences emphasize specific characteristics and challenges in the two domains, and are useful in studying the transferability of the models from one domain to another.

3.3. Benchmark Datasets

To enable progress on fine-grained location identification in crisis tweets, and facilitate comparisons between models developed for this task (in-domain and cross-domain), we created benchmark datasets by randomly splitting our *MIXED* and *COVID* datasets into training (`train`), development (`dev`) and test (`test`) subsets, respectively. We use the training subset to train our models, the development subset to select hyperparameters and the test subset to evaluate the final performance of the models. Statistics for the *MIXED* and *COVID* datasets in terms of number of tweets, tokens, entities in the `train`, `test` and `dev` subsets, respectively, are shown in Table 4. The benchmark datasets, together with the pre-processing script, are made publicly available. More specifically, to comply with Twitter's Developer Agreement and Policy[3], the datasets is made available as pairs of tweet ID and corresponding locations. The locations are specified as a list of location-phrases and corresponding location types. Given that the pre-processing script will also be made available, the index of the location phrase should precisely match the index of the tweet tokens.

[3] `https://developer.twitter.com/en/developer-terms/agreement-and-policy`

4. Background and Approaches

The task of identifying fine-grained locations in tweet text can be formulated as follows: Given a set of (X, Y) pairs, where $X = \{x_1, \cdots, x_n\}$ is a text sequence/tweet with n tokens, and $Y = \{y_1, \cdots, y_n\}$ is a tag sequence with n location tags/types (in BIO format) corresponding to the tokens in the sequence X; our sequence tagging task is to find a mapping $f_\theta : \mathbf{X} \rightarrow \mathbf{Y}$ (with parameters θ) from input sequences to output sequences of fine-grained location types.

4.1. Baseline Models

We have experimented with three classes of models to establish baseline results on the location identification task.

4.1.1. Feature-Engineered Baseline.

Stanford NER (Finkel et al., 2005) uses an arbitrary order linear chain CRF model over a set of predefined word and character level features extracted from the input. The model has been used as a strong baseline for many NER models. We retrain the model with both *MIXED* and *COVID* datasets, respectively, to learn fine-grained location types.

4.1.2. Character and Word Embedding Baselines.

One model architecture in this category consists of a distributed representation layer learning the embeddings at character and word level followed by an LSTM-based context-encoder layer and a CRF tag-decoder. The model is referred as **CNN-GloVe-BiLSTM-CRF** in what follows. Considering the recent success of transformer-based models, we also experiment with a similar model where BERT is used as the embedding layer instead of CNN+GloVe. We call this model **BERT-BiLSTM-CRF**. For both CNN-GloVe-BiLSTM-CRF and BERT-BiLSTM-CRF models, we employ a multitask learning approach (Caruana, 1997), in which the main task of fine-grained location tagging is learned simultaneously with the auxiliary task of a generic yes/no location tagging. We refer this model using the *-MTL* suffix in what follows.

4.1.3. Word and Entity Embedding Baseline.

In addition to using contextualized word embeddings learned from a transformer-based language model, **LUKE** (Yamada et al., 2020) also learns contextualized entity embeddings and subsequently uses an entity-aware self-attention mechanism to perform tasks such as entity typing, relation classification, NER, etc. The LUKE approach has achieved state-of-the-art results on standard NER datasets (among others). We fine-tune

the pre-trained *LUKE-base* model with the *COVID* and *MIXED* datasets, respectively. The LUKE model selects candidate entity spans before making the entity type category predictions, a task that is comparable to the auxiliary task in the MTL models discussed earlier. Hence, we do not use the multitask learning setting for LUKE.

5. Experimental Setup

In this section, we discuss the metrics used in the evaluation, implementation details and experiments performed.

5.1. Metrics

We use standard metrics, including precision (Pr), recall (Re) and F1 measure (F1), to evaluate the performance of the models trained.

5.2. Implementation details

We performed a grid-search with 5 trials and used the development subsets to identify best-overall hyperparameter values. We used the best-overall values in the experiments. We used the Glorot uniform initializer (Glorot and Bengio, 2010) to initialize the model weights. The optimization was performed using the AdamW optimizer (Loshchilov and Hutter, 2019), with a learning rate of $1e{-}3$, weight decay of $1e{-}2$, and gradient clipping with max norm of 5. We used a dropout of 0.5 and mini-batch size of 32 in all the experiments. We set a patience of 5 epochs on the development F1-measure, as early stopping of training. All experiments are run on NVIDIA Tesla V100 GPU.

5.3. Experiments

We conducted experiments in two settings, *in-domain* and *cross-domain*. In the **in-domain setting**, models were trained and tested on the same dataset (e.g., models were trained on MIXED-train, tuned on MIXED-dev, and tested on MIXED-test). The goal was to study:

1. the performance of the deep learning models by comparison with the traditional Stanford NER model;
2. the effect of the auxiliary task in the MTL framework;
3. the effect of different types of embeddings.

In the **cross-domain setting**, we used the best in-domain model to investigate several ways to perform transfer of information between domains:

1. a *zero-shot* transfer setting, where models trained on one dataset were tested on the other dataset (e.g., models trained on MIXED-train, tuned on MIXED-dev and tested on COVID-test);
2. an *embedding-level* transfer, where the transformer block fine-tuned on one dataset (e.g., MIXED) was used as a starting point for the transformer block of the model trained/tuned/tested on the other dataset (e.g., COVID);
3. a *model-level* transfer, where the model trained/tuned on a dataset (e.g., MIXED-train, MIXED-dev) is used as the starting point of the model for the other dataset (e.g., COVID-train, COVID-dev, COVID-test, respectively).

Dataset	**MIXED**		
Model	**Pr**	**Re**	**F1**
Stanford NLP (retrained)	**82.52**	65.64	73.12
CNN-GloVe-BiLSTM-CRF	80.26	65.20	71.95
CNN-GloVe-BiLSTM-CRF-MTL	76.92	59.30	66.97
BERT-BiLSTM-CRF	74.98	70.07	74.52
BERT-BiLSTM-CRF-MTL	74.58	**74.75**	74.67
LUKE	80.71	73.08	**76.71**
Dataset	**COVID**		
Model	**Pr**	**Re**	**F1**
Stanford NLP (retrained)	**85.71**	57.62	68.92
CNN-GloVe-BiLSTM-CRF	77.27	68.81	71.66
CNN-GloVe-BiLSTM-CRF-MTL	78.64	52.74	63.14
BERT-BiLSTM-CRF	73.41	70.74	72.05
BERT-BiLSTM-CRF-MTL	77.06	69.43	73.04
LUKE	78.49	**71.12**	**74.66**

Table 5: In-domain results. Comparison of the following models: Stanford NER, CNN-GloVe-BiLSTM-CRF/BERT-BiLSTM-CRF and their MTL variants, and LUKE.

6. Results and Discussion

We first present and discuss the in-domain results, followed by the cross-domain results. In addition, we also perform error analysis and discuss the robustness of the models.

6.1. In-domain Setting

Table 5 shows the in-domain results of the models. As can be seen in Table 5, the entity-embedding based LUKE model is the best overall in terms of F1-measure for both *MIXED* and *COVID* datasets, with a relatively high recall compared to most of the other models. Specifically, the F1-measure is 76.71% for the *MIXED* dataset and 74.66% for the *COVID* dataset. While the Stanford NLP has the highest precision overall, we argue that in the context of disaster monitoring and response, recall is more important than precision, as the final results will be reviewed by humans before any action is taken. Comparing the results for the *MIXED* and *COVID* datasets, we can see that the models have slightly better performance on the *MIXED* dataset. While this dataset contains a variety of crisis events, the events are relatively localized to specific geographical regions, which may make it easier for the models to identify the locations. As opposed to that, the *COVID* dataset has a big variety of locations as it covers a global pandemic. Nevertheless, the F1 score of the LUKE model on *COVID* is 8.3% higher than the score of the Stanford NLP model, which uses manually designed features for training. We can also observe that the contextualized word and/or entity embeddings obtained from transformer architectures are

Datasets		COVID→MIXED		
Model	**Transfer style**	**Pr**	**Re**	**F1**
BERT-BiLSTM-CRF-MTL	zero-shot	77.05	54.80	64.05
	embedding-level	76.40	71.20	73.71
	model-level	79.86	71.05	75.20
LUKE	zero-shot	50.65	47.41	48.98
	embedding-level	78.13	**74.21**	76.12
	model-level	**81.32**	73.57	**77.25**
Datasets		**MIXED→COVID**		
Model	**Transfer style**	**Pr**	**Re**	**F1**
BERT-BiLSTM-CRF-MTL	zero-shot	36.23	45.85	40.41
	embedding-level	67.19	68.85	68.01
	model-level	66.44	71.47	68.86
LUKE	zero-shot	**79.94**	43.17	56.06
	embedding-level	76.95	**73.78**	75.33
	model-level	78.08	73.32	**75.63**

Table 6: Cross-domain results. Comparison between three transfer styles: zero-shot, embedding-level and model-level.

Model level COVID→MIXED transfer				
Model	**Test**	**Pr**	**Re**	**F1**
BERT-BiLSTM-CRF-MTL	MFT	**79.86**	**71.05**	**75.20**
	INV	67.78	52.29	59.03
	DIR	47.48	31.34	37.76
LUKE	MFT	**81.32**	**73.57**	**77.25**
	INV	70.52	50.71	58.99
	DIR	56.87	34.41	42.88
Model level MIXED→COVID transfer				
BERT-BiLSTM-CRF-MTL	MFT	**66.44**	**71.47**	**68.86**
	INV	57.64	59.35	58.48
	DIR	40.26	33.80	36.75
LUKE	MFT	**78.08**	**73.32**	**75.63**
	INV	69.86	53.43	60.55
	DIR	49.47	29.28	36.79

Table 7: Error analysis tests (MFT, INV and DIR) for the capability of the model-level transfer approach to generalize the concept of a location entity.

better than both the engineered features in Stanford NLP and the character/word-embeddings in the CNN-GloVE-BiLSTM-CRF models. Finally, when comparing the BERT-BiLSTM-CRF-MTL model (with auxiliary task) to its BERT-BiLSTM-CRF variant (without the auxiliary task), the results show that the auxiliary task can help improve the F1-measure, especially in the case of *COVID*. However, for CNN-GloVe-BiLSTM-CRF, the addition of the auxiliary task decreases the F1-measure. This suggests that the transformer allows for a richer transfer of knowledge between similar tasks as compared to the CNN/GloVe architectures.

6.2. Cross-domain setting

Table 6 shows the results of the BERT-BiLSTM-CRF-MTL and LUKE models (which give the best overall results in the in-domain setting) in the cross-domain setting. Specifically, we compare three transfer styles, *zero shot*, *embedding-level*, and *model-level*, when *COVID* is used as source and *MIXED* as target, and the other way around. As expected, the model-level transfer style gives the best results overall, while the zero-shot style gives the worst results overall. Notably, in the case of the *COVID* to *MIXED* transfer, the model-level transfer improves the results of the in-domain LUKE model, from 76.71% to 77.25%. This is probably due to the diversity in the *COVID* dataset, which enables more accurate locations to be identified in the *MIXED* dataset. As opposed to that, the transfer from *MIXED* to *COVID* causes more specific locations to be identified, which improves the recall but negatively affects the precision (and the overall F1-measure).

6.3. Error analysis

We performed error analysis of the model-level transfer from Table 6 for both BERT-BiLSTM-CRF-MTL and LUKE (specifically, model-level transfer from *COVID* to *MIXED* and from *MIXED* to *COVID*). The analysis is based on the framework proposed by Ribeiro et al. (2020), where a model is tested for a capability using three tests: minimum functionality test (MFT), invariance test (INV) and directional expectation test (DIR). We performed the tests on the model's capability to generalize the concept of a location entity. In our case, MFT is the model's performance on the original *MIXED* or *COVID* test set, respectively. For INV, the location entities in the original test set were replaced with other randomly selected location entities of the same type from the test set. Finally, for DIR, the original location entities were replaced with randomly selected location entities of different types from the test set. The results of the analysis are shown in Table 7. The MFT score serves as a baseline for the other two tests. As can be seen, in both cases, the performance degrades when the locations are mixed up - tests INV and DIR as compared with the test MFT - suggesting that the model captures correlations between locations and their context. However, the F1 score for INV is better than the F1 score for DIR, which shows that the model expects a particular type of location in a given context.

Table 8 shows sample predictions for different tests (MFT, INV, DIR). In the first example, for the MFT test, the model makes a correct prediction for a tweet where a location entity of type *ctc* is followed by a location entity of type *sta*, which is the general convention for specifying a *city, state* location. However, for the DIR test, when the entities are replaced with others in reverse order of the type as compared to the original tweet (i.e., *sta, ctc* instead of *ctc, sta*), the model incorrectly, but not surprisingly, predicts *sta* as *ctc* and vice versa. In the second example, for the MFT test, the model correctly predicts Sri Lanka as a country (i.e., *con*). However, when *Sri Lanka* is replaced with *South Africa* in the case of the INV test, the model predicts it as *reg*. This is probably because Africa as a continent is a location of type *reg*, and also because cardinal directions are commonly associated with *reg* locations. Hence, without any external knowledge about *South Africa* as a country, *reg* is the next best prediction.

7. Conclusions and Future Work

In this paper, we introduced two new crisis tweet datasets manually tagged with specific fine-grained location types. These are the first manually annotated

No.	Test	Tweet text
1	MFT	NEW Hurricane Florence has made landfall near Wrightsville Beach ctc→ctc, North Carolina sta→sta at 7 15 AM EDT
	DIR	NEW Hurricane Florence has made landfall near WI, sta→ctc Panama ctc→sta at 7 15 AM EDT
2	MFT	On this Easter Sunday my thoughts are with Sri Lanka con→con following the horrific attacks on worshippers there.
	INV	On this Easter Sunday my thoughts are with South Africa con→reg following the horrific attacks on worshippers there.

Table 8: Examples of location predictions for different error analysis test settings (MFT, INV and DIR). The examples are from the *MIXED* datasets and the predictions are made using model-level transfer from *COVID* to *MIXED*. The locations are highlighted with pink. The labels and predictions for entities are shown as a subscript to the corresponding locations using the convention *gold label→model prediction*.

datasets for fine-grained location identification in crisis tweet texts, and can foster research in this area of great importance for crisis monitoring and response. The two datasets are different in nature, with one of them focused on mixed natural and man-made crisis events, which are generally localized to specific regions, and the second one focused on the worldwide COVID-19 pandemic. The different nature of the two datasets enables studies on location identification for localized and global events, as well as studies on the transferability of information between localized and global events.

In addition to introducing these datasets, we reported baseline results for the fine-grain location identification task using state-of-the-art models based on different embedding styles. Our results suggest that the entity-embedding style of the LUKE model gives the best results. We also used MTL to incorporate an auxiliary task in one of the models and showed its effectiveness in transferring information between datasets. As part of future work, we plan to improve the results of the models by including other crisis-related tagging and classification tasks in the LUKE/MTL settings.

8. Ethics and Impact Statement

The dataset that we plan to share will not provide any personally identifiable information, as only the tweet IDs and human annotated location tags will be shared. Thus, our dataset complies with Twitter's Developer Agreement and Policy[4] in terms of privacy. Furthermore, in compliance with the Twitter's Developer Agreement and Policy, Section III.E, the location information is used only in conjunction with the tweet content, and, as allowed by Twitter, we "`only use such location data and geographic information to identify the location tagged by the Twitter Content.`" In terms of impact, the research enabled by this dataset has the potential to help officials and health organizations identify actionable information useful for fast response during a crisis situation, or facilitate the health organizations to aggregate information relevant to COVID-19 by locations (which in turn can be useful in preventing a serious resurgence of the novel coronavirus in a particular region). However, we want to emphasize that we do not use any of the information in Twitter content, in particular the location information, to infer any sensitive information about the user, and most importantly our models do not infer any information about users' health[5]. The models are simply trained to identify location tags in tweets (as explicitly allowed by Twitter) and nothing more. Also important, our pre-processing script removes any user mentions from the tweet content before feeding the tweets to the models for training.

[4] `https://developer.twitter.com/en/developer-terms/agreement-and-policy`

[5] `https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases`

Acknowledgements

We thank the National Science Foundation and Amazon Web Services for support from grant IIS-1741345, which supported the research and the computation in this study.

9. Bibliographical References

Alam, F., Joty, S., and Imran, M. (2018a). Domain adaptation with adversarial training and graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alam, F., Ofli, F., and Imran, M. (2018b). Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 465–473. AAAI Press.

Anand, A., Awekar, A., et al. (2017). Fine-grained entity type classification by jointly learning representations and label embeddings. *arXiv preprint arXiv:1702.06709*.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. (2019). Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh et al., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May. PMLR.

Goyal, A., Gupta, V., and Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29.

Han, B., Yepes, A. J., MacKinlay, A., and Chen, Q. (2014). Identifying twitter location mentions. In *Proceedings of the Australasian Language Technology Association Workshop 2014*.

Hoang, T. B. N., Moriceau, V., and Mothe, J. (2017). Predicting locations in tweets. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CLICLing 2017)*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9.

Ikawa, Y., Vukovic, M., Rogstadius, J., and Murakami, A. (2013). Location-based insights from the social web. In *Proceedings of the 22nd international conference on World Wide Web*.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47.

Imran, M., Mitra, P., and Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *LREC*.

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., and Ghazi, D. (2015). Detecting and disambiguating locations mentioned in twitter messages. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 321–332, Cham. Springer International Publishing.

Ji, Z., Sun, A., Cong, G., and Han, J. (2016). Joint recognition and linking of fine-grained locations from tweets. In *Proceedings of the 25th International Conference on World Wide Web*.

Jiang, Y., Hu, C., Xiao, T., Zhang, C., and Zhu, J. (2019). Improved differentiable architecture search for language modeling and named entity recognition. 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 3585–3590, Hong Kong, China, November. Association for Computational Linguistics.

King, L. (2018). Social media use during natural disasters: An analysis of social media usage during hurricanes harvey and irma. In *Proceedings of the International Crisis and Risk Communication Conference*, volume 1, pages 20–23, 03.

Kumar, A. and Singh, J. P. (2019). Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction*, 33.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01.

Lal, A. et al. (2019). Sane 2.0: System for fine grained named entity typing on textual data. *Engineering Applications of Artificial Intelligence*, 84.

Li, C. and Sun, A. (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*, SIGIR '14, page 43–52, New York, NY, USA. Association for Computing Machinery.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Lingad, J., Karimi, S., and Yin, J. (2013). Location extraction from disaster-related microblogs. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 1017–1020, New York, NY, USA. Association for Computing Machinery.

Liu, F., Vasardani, M., and Baldwin, T. (2014). Automatic identification of locative expressions from social media text: A comparative analysis. In *Proceedings of the 4th International Workshop on Location and the Web*, pages 9–16.

Liu, Y., Meng, F., Zhang, J., Xu, J., Chen, Y., and Zhou, J. (2019). GCDT: A global context enhanced deep transition architecture for sequence labeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Luo, Y., Xiao, F., and Zhao, H. (2019). Hierarchical contextualized representation for named entity recognition. *arXiv preprint arXiv:1911.02257*.

Magge, A., Weissenbacher, D., Sarker, A., Scotch, M., and Gonzalez-Hernandez, G. (2019). Bi-directional recurrent neural network models for geographic location extraction in biomedical literature. In *PSB*.

Mahmud, J., Nichols, J., and Drews, C. (2012). Where is this tweet from? inferring home locations of twitter users. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Malmasi, S. and Dras, M. (2015). Location mention detection in tweets and microblogs. In *Conference of the Pacific Association for Computational Linguistics*. Springer.

Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S., and Clancy, S. (2010). Spatiaiml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280.

Miao, L., Last, M., and Litvak, M. (2020). Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Mutlu, E. C., Oghaz, T., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., Ozmen, O., and Garibay, I. (2020). A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19. *Data in brief*, 33:106401.

Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., and Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In *ICWSM*.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *AAAI Conference on Weblogs and Social Media*.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In *CSCW*.

Qazi, U., Imran, M., and Ofli, F. (2020). Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information. *arXiv preprint arXiv:2005.11177*.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July. Association for Computational Linguistics.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1079–1088.

Villegas, C., Martinez, M., and Krause, M. (2018). Lessons from harvey: Crisis informatics for urban resilience.

Xu, C., Li, J., Luo, X., Pei, J., Li, C., and Ji, D. (2019). Dlocrl: A deep learning pipeline for fine-grained location recognition and linking in tweets. In *The World Wide Web Conference*.

Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. (2020). LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November. Association for Computational Linguistics.