On Identifying Hashtags in Disaster Twitter Data

Jishnu Ray Chowdhury[1], **Cornelia Caragea**[1], **and Doina Caragea**[2]
[1]Department of Computer Science, University of Illinois at Chicago
[2]Department of Computer Science, Kansas State University
jraych2@uic.edu, cornelia@uic.edu, dcaragea@ksu.edu

Abstract

Tweet hashtags have the potential to improve the search for information during disaster events. However, there is a large number of disaster-related tweets that do not have any user-provided hashtags. Moreover, only a small number of tweets that contain actionable hashtags are useful for disaster response. To facilitate progress on automatic identification (or extraction) of disaster hashtags for Twitter data, we construct a unique dataset of disaster-related tweets annotated with hashtags useful for filtering actionable information. Using this dataset, we further investigate Long Short-Term Memory-based models within a Multi-Task Learning framework. The best performing model achieves an F1-score as high as 92.10%. The dataset, code, and other resources are available on Github.[1]

Introduction

During disasters, affected individuals often turn to social media platforms, such as Twitter and Facebook, to find the latest updates from government and response organizations, to request help or to post information that can be used to enhance situational awareness (Rhodan 2017; MacMillan 2017; Frej 2018; Lapin 2018). Nonetheless, the value of the information posted on social media platforms during disasters is highly unexploited, in part due to the lack of tools that can help filter relevant, informative, and actionable messages (Villegas, Martinez, and Krause 2018).

According to Villegas, Martinez, and Krause (2018), more than $5,200$ rescue requests made on social media were missed by the first responders, while about 46% of the critical damage information posted on social media during Harvey Hurricane was missed by FEMA in their original damage estimates (that is almost half of the total costs of $\$125$ billion estimated for this hurricane).[2] As an official explained: "It's very labor intensive to watch [social media] and because of the thousand different ways people can *hashtag* something or *keyword* something, trying to sort out what's relevant and what's not and what's actionable is very, very difficult" (Silverman 2017).

Examples of tweets that illustrate the diverse ways in which people use hashtags to highlight information during disasters are shown in Table 1. Specifically, the user-provided hashtags, when available, are shown in blue color in the table. As can be seen, the first two tweets do not have any user-provided hashtags. The third tweet has a general disaster-name hashtag, #HurricaneIrma. While this hashtag is useful in recognizing that the tweet was posted during Hurricane Irma, it is not useful in identifying situational awareness (e.g., damage, power loss, blocked street) or the type of disaster response requests. The fourth tweet, which explicitly reports damage, has disaster-name, location, and weather as hashtags, but no specific hashtag about damage. Finally, the fifth tweet is at the other extreme, in that it has a large number of hashtags (specifically, 11), some of them representing lexical variations of the same base word.

These examples show not only that people use a variety of ways to hashtag tweets or that they may not understand or know how to hashtag tweets, but also that the user-provided hashtags tend to be either too general or too specific. Moreover, these examples are not exceptions, but rather they are representatives for a disaster-related tweet dataset. An analysis of a large corpus of tweets that was used in this work revealed that most of the hashtags in a disaster-related tweet corpus simply represent disaster names and locations, and that approximately half of the tweets do not have any hashtags at all. Thus, filtering based on user-provided hashtags is not helpful for disaster response or people on the ground to quickly find relevant information. Similarly, filtering tweets during disasters based on keyword match is not expected to work well since keywords can be ambiguous and can lead to noisy results, e.g., a search for the keyword "Harvey" will retrieve tweets about the hurricane, but also about people whose name is "Harvey."

To address these limitations of tweet retrieval based on user-provided hashtags or keywords, we envision a system that learns to identify relevant and topically informative hashtags and extract them directly from the content of the disaster tweets, capturing three main elements in a tweet: 1) disaster name; 2) location; and 3) situational awareness in-

[1]https://github.com/JRC1995/Tweet-Disaster-Keyphrase

[2]https://coast.noaa.gov/states/fast-facts/hurricane-costs.html

No.	Tweet text
1.	we need help in Houston. our apartments are surrounded with water like an island we need rescue 10373 N Sam Houston Pkwy E need help | Houston | need rescue
2.	@houstonpolice please help I'm stranded with my kids I need help fast my address is 8618 Banting st. houston tx 77078. stranded | need help | houston
3.	Big tree fell on power lines and blocking Brown Ave near Washington St in Orlando's Thornton Park neighborhood. #HurricaneIrma power lines | blocking | Orlando | #HurricaneIrma
4.	Very extensive damage sustained throughout #Wilmington, #ncwx... from #hurricane #Florence. Lots of trees split or uprooted, siding ripped from homes, powerlines down, flooding of downtown streets, etc. extensive damage | #Wilmington | #hurricane | #Florence | powerlines down
5.	I am evacuated from my house but I'm safe. #fire #CampFire #WoolseyFire #wildfire #safe #Evacuation #evacuations #EVACUATED #scary #ThousandOaks #Camarillo evacuated | #WoolseyFire | #ThousandOaks | #Camarillo

Table 1: Examples of tweets posted during disasters. The original user-provided hashtags, when available, are shown in blue color for each tweet. Relevant and topically more informative hashtags manually identified to have the potential to retrieve actionable disaster tweets are highlighted in a light red box.

formation. Examples of relevant and topically informative hashtags that represent these three elements for the tweets in Table 1 are provided with each tweet and are shown in a light red box in the table (these hashtags are extracted directly from the tweets' content). Precisely, the first two tweets may be retrieved based on a search for #Houston and #needhelp. The third tweet can be retrieved with a search for #HurricaneIrma, #Orlando, #powerlines. The fourth tweet can be retrieved with a search for #HurricaneFlorence, #Wilmington and #damage (or #powerlines). Finally, the fifth tweet can be retrieved when searching for #WoolseyFire, #ThousandOaks or #Camarillo, or #evacuated, and can be used to find information about evacuated people. Thus, together, the above three elements can be used to filter tweets of potential interest to an emergency organization, which is responding to the disaster in question, or can be used to recommend hashtags in real time as the user types.

Although there are previous works that focus on hashtag recommendation (Gong and Zhang 2016; Li et al. 2016a; Zhang et al. 2017; Li et al. 2019) and topical keyphrase extraction (Marujo et al. 2015; Zhang et al. 2016) for the general Twitter, research on identifying and extracting hashtags from the disaster Twitter data is limited. A notable exception is the work by Imran et al. (2013b), where the authors extracted short information nuggets representing "what", "where", "when", etc., for a very small number of tweets classified in specific situational awareness categories.

One potential reason that hindered progress on automatic hashtag identification (or extraction)[3] from disaster-related tweets is the lack of large publicly available social media datasets annotated with relevant and topically informative hashtags. To fill in this gap, we constructed a large and unique dataset of disaster-related tweets annotated with hashtags to enable the development of deep learning techniques for automatic hashtag identification in order to further research in this critical area.

In doing so, we first collected tweets related to multiple disasters and disaster types (e.g., hurricane, flooding) and then manually crafted a lexicon, which was used together with the hashtags from the tweets, whenever available, to annotate a large dataset. Using this dataset, we further investigated a powerful deep learning model, initially proposed by (Zhang et al. 2016) for keyphrase extraction from general tweets, and evaluate its performance capability for hashtag extraction from disaster-related tweets. This model, a joint-layer Long-Short Term Memory network trained using Multi-Task Learning (LSTM-MTL) and its variants that capture specifics of informal writing can be regarded as strong baselines on this dataset. Specifically, we make the following contributions:

1. We present a hashtag annotated dataset of more than $67,288$ tweets related to disasters of various types (e.g., hurricanes, floodings, and earthquakes) and validate the hashtag annotations using human judgements.
2. We develop an LSTM-MTL model and variants that incorporate informal writing styles in order to exploit our new dataset for automatic hashtag extraction. The dataset, code, and other resources from this work are made available on Github.
3. We conduct a thorough empirical evaluation of the LSTM-MTL model and its variants and show improvements of these variants over strong baselines.

Related Work

Mizuno et al. (2016) described two systems that can be used to analyze and summarize information posted on social media during disasters. The first system, called DISAANA,

[3]We use interchangeably hashtag identification and hashtag extraction in this paper.

can answer questions (e.g., "What is in short supply in Kumamoto?") and list problem reports identified using Twitter data (e.g., "people were burried alive"). The second system, called D-SUMM, can be used to summarize similar problem reports into a broader category. These systems work directly on Twitter, and could benefit from informative extractive hashtags in order to reduce the number of tweets they can be used on, and consequently, improve their speed.

A variety of hashtag recommendation approaches have been proposed for general tweets. For example, Li et al. (2016a) used a multi-class classification approach (i.e., each candidate hashtag represents a class), which combines the skip-gram model for finding word embeddings (Mikolov et al. 2013), with a convolutional neural network for learning sentence vectors, and a long short-term memory (LSTM) network (Hochreiter and Schmidhuber 1997a) for combining sentence vectors into a tweet vector. The tweet vectors were provided as input to a softmax layer, which was used to identify hashtags associated with the tweet, among a set of candidate hashtags. The approach was tested on a general tweet dataset, using a set of 20 popular hashtags. Experimental results showed that the proposed approach performed better than baselines that used TF-IDF or other types of recurrent neural networks.

Gong and Zhang (2016) also formulated the problem as a multi-class classification task, and proposed an approach based on convolutional neural networks, seen as a global channel, combined with the attention mechanism, seen as a local channel, to recommend hashtags. The attention mechanism was used to identify tweet words that trigger hashtags. The model was tested on a general tweet dataset with user-provided hashtags as gold-standard, and gave significant improvements over several baselines that did not use deep learning models.

Li et al. (2016b) proposed an LSTM-based approach that uses the attention mechanism to incorporate the topic information (Blei, Ng, and Jordan 2003) of the tweet into the model. Implicitly, the model finds associations between local hidden representations of the words and the global topic information of the tweet, and uses these associations to generate a representation that leads to useful topical hashtags when passed through a softmax layer. Most recently, Li et al. (2019) extended their previous approach to include a co-attention mechanism that models content and topic information simultaneously. The extended approach was inspired from another hashtag recommendation approach (Zhang et al. 2017) that used the co-attention mechanism to combine textual and visual information available in many tweets. More specifically, the tweet content was modeled using a bidirectional LSTM (Bi-LSTM) sequence encoder, while the tweet topic was modeled using the approach proposed in (Zhao et al. 2011b). Using a co-attention mechanism, a new content/topic representation is learned for each tweet. As with the other hashtag recommendation approaches, this approach was evaluated on general tweets. Experimental results showed significant improvements over several baselines, including the previous model in (Li et al. 2016b).

One common theme to the hashtag recommendation approaches reviewed above is that they formulate the problem as a multi-class classification task, where the hashtags are a priori established, and a softmax layer transforms the hidden tweet representation into a probability distribution over the hashtags. Usually, there is a small number of candidate hashtags, e.g., 20 as in (Li et al. 2016a).

While these approaches allow recommendation of hashtags that do not appear in the tweet, the requirement of preselecting a fixed number of hashtag candidates makes the classification-based approaches impractical for our purpose. In a time of disaster, it is expected that there could a new set of emerging disaster-related entities that we may want to use as hashtags. In critical times, we cannot afford to collect new sets of candidate hashtag classes and re-train the classification models. Therefore, instead of a classification-based approach, we take an extractive approach to this task. Precisely, we explore models for extracting important terms (which may serve as good hashtags) present in the tweets. As such, our task more closely aligns with keyphrase extraction.

Hashtags in tweets are closely related to keyphrases. For example, Zhang et al.(2016) used hashtags as gold keyphrases for keyphrase extraction from Twitter. Therefore, we treat the task of hashtag identification as similar to the task of keyphrase extraction. Thus, close to our research is also the work on keyphrase extraction from Twitter (Marujo et al. 2015; Zhang et al. 2016; Zhao et al. 2011a; Bellaachia and Al-Dhelaan 2012). For example, (Marujo et al. 2015) formulated the problem as binary classification and showed that word embeddings in a system such as MAUI (Medelyan, Frank, and Witten 2009) perform better than the TF-IDF (Sparck Jones 1972) for keyphrase extraction on general tweets. Zhang et al. (2016) formulated the problem as a sequence labeling task which allows the extraction of keyphrases of arbitrary lengths, without being constrained by some fixed number of classes. Zhang et al. (2018) extends the work of Zhang et al. (2016) by encoding conversational context.

We chose to focus on a sequence labeling model for identifying hashtag, and specifically a model based on the Joint-Layer-RNN proposed by Zhang et al. (2016) since it achieved state-of-the-art performance on general tweets. This approach enables the model to extract new hashtags that may not have been seen before in the training data. This is particularly important in the context of disaster tweet annotation as new disasters with specific new names, locations, requests, and needs happen all the time.

Hashtag Annotated Dataset

Previous approaches for tweet keyphrase or hashtag extraction have been used with general tweets, where the user-provided hashtags are considered to be the gold-standard. There is no large corpus for disaster tweet hashtag extraction. Furthermore, as explained above, the user-provided hashtags for disaster-related tweets do not always contain useful or sufficient information in terms of situational awareness and disaster response. Therefore, the strongest contribution of our work is to construct a benchmark dataset for disaster tweet hashtag extraction and to provide several models that can be used as strong baselines for this dataset.

Disaster	Size
Joplin Tornado (Imran et al. 2013b)	2280
Sandy Hurricane (Imran et al. 2013a)	514
Mexico Earthquake (Alam, Ofli, and Imran 2018)	904
Harvey Hurricane (Alam, Ofli, and Imran 2018)	2952
Irma Hurricane (Alam, Ofli, and Imran 2018)	3050
Nepal Earthquake (Alam, Joty, and Imran 2018)	4890
Queensland Floods (Alam, Joty, and Imran 2018)	2941
Colorado Wildfires (Olteanu, Vieweg, and Castillo 2015)	623
Costa Rica Earthquake (Olteanu, Vieweg, and Castillo 2015)	191
Guatemala Earthquake (Olteanu, Vieweg, and Castillo 2015)	133
Venezuela Refinery (Olteanu, Vieweg, and Castillo 2015)	36
Alberta Floods (Olteanu, Vieweg, and Castillo 2015)	641
Australia Bushfire (Olteanu, Vieweg, and Castillo 2015)	631
Bohol Earthquake (Olteanu, Vieweg, and Castillo 2015)	334
Boston Bombings (Olteanu, Vieweg, and Castillo 2015)	391
Colorado Floods (Olteanu, Vieweg, and Castillo 2015)	696
LA Airport Shootings (Olteanu, Vieweg, and Castillo 2015)	603
Manila Floods (Olteanu, Vieweg, and Castillo 2015)	434
Sardinia Floods (Olteanu, Vieweg, and Castillo 2015)	66
Singapore Haze (Olteanu, Vieweg, and Castillo 2015)	281
Typhoon Yolanda (Olteanu, Vieweg, and Castillo 2015)	619
West Texas Explosion (Olteanu, Vieweg, and Castillo 2015)	422
Typhoon Pablo (Olteanu, Vieweg, and Castillo 2015)	507
Philippines Floods (Olteanu, Vieweg, and Castillo 2015)	540
Pakistan Earthquake (Imran, Mitra, and Castillo 2016)	1262
California Earthquake (Imran, Mitra, and Castillo 2016)	1453
Chile Earthquake (Imran, Mitra, and Castillo 2016)	1047
Ebola Virus (Imran, Mitra, and Castillo 2016)	1468
Hurricane Odile (Imran, Mitra, and Castillo 2016)	992
India Floods (Imran, Mitra, and Castillo 2016)	1183
M.E.R.S (Imran, Mitra, and Castillo 2016)	1234
Pakistan Floods (Imran, Mitra, and Castillo 2016)	1524
Typhoon Hagupit (Imran, Mitra, and Castillo 2016)	1453
Cyclone Pam (Imran, Mitra, and Castillo 2016)	1328
Nepal Earthquake (Imran, Mitra, and Castillo 2016)	28
Chiapas Earthquake (ours)	2549
Mexico Earthquake (ours)	3742
Harvey Hurricane (ours)	4349
Irma Hurricane (ours)	4246
Maria Hurricane (ours)	12200
California Fire (ours)	2551
All Disasters Tweets (in total)	67, 288

Table 2: Number of tweets extracted for different disasters from different sources.

Collection of Disaster Tweets

To construct a large ground-truth dataset for hashtag extraction, we first collected publicly available datasets that

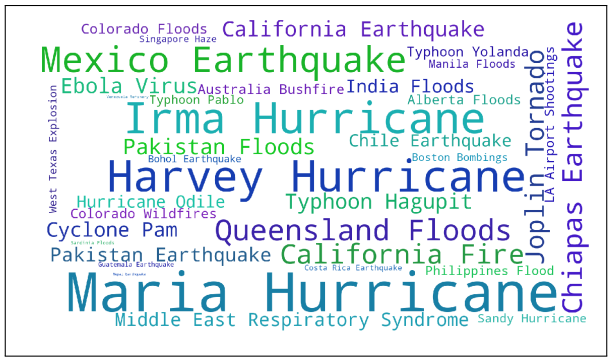

Figure 1: Wordcloud of disasters in our dataset.

contain tweets related to a big variety of disasters of different types (Imran et al. 2013a; 2013b; Olteanu et al. 2014a; Olteanu, Vieweg, and Castillo 2015; Imran, Mitra, and Castillo 2016; Alam, Joty, and Imran 2018; Alam, Ofli, and Imran 2018). All these existing datasets were annotated using crowdsourcing platforms or volunteers, with respect to relevance to a disaster and/or informativeness. We removed tweets annotated as irrelevant, as uninformative, or as related to emotional support and sympathy.

In addition to the publicly available disaster datasets, we also used a collection of tweets that we crawled, using the Twitter streaming API, during the following disasters: Hurricane Harvey, Hurricane Irma, Hurricane Maria, Mexico Earthquake, Chiapas Earthquake, and California Fire (wildfire). This collection was filtered for disaster relevance using a Naïve Bayes classifier trained on CrisisLexT6 (Olteanu et al. 2014a). Finally, we removed duplicates and non-English tweets from the combined dataset that consists of the publicly available datasets and our own collection of tweets. The resulting combined dataset comprises a large variety of disasters of different types (37 disasters in total), as illustrated in Figure 1. The size of each disaster name in the figure is proportional to the number of tweets from that disaster included in our new dataset. The overall details of our dataset including all the disasters present in it along with their corresponding source and size are provided in Table 2.

Disaster Lexicon

Following the disaster-relevant data collection, we proceeded with a disaster lexicon construction, as follows: we first sampled approximately 150 tweets from each of the 37 disasters in our collection; then, we manually identified terms that were relevant and topically informative for the respective disaster and added them to the lexicon. The types of terms and phrases that we identified as relevant and informative are: location terms, requests or help-related terms (need food, send help, etc.), situation awareness terms (building collapse, 100 dead, stranded, trapped, etc.), disaster names (Hurricane, Hurricane Maria), names of important persons, organizations, or nouns related to important vocations (related to government, president, etc.), terms related to disaster warnings and updates, and other terms associated with infrastructure damage, property loss, affected individuals, injured or missing persons, fatality rates, etc.

We only stored unigram and bigram phrases in the lexicon. In total, we extracted $2,140$ lexicon phrases from our sample tweets ($\approx 5,500$ tweets). Finally, we included some phrases from the CrisisLex lexicon (Olteanu et al. 2014b), which were not already in our lexicon. Our final lexicon contains $2,430$ unique phrases.

Dataset Annotation

We used the manually constructed lexicon to automatically annotate *hashtags* in the tweets, by matching the lemmatized version of a tweet phrase with a lemmatized phrase in the disaster lexicon. We chose to use the lemmatized version of a word instead of its stemmed version to avoid ambiguous words resulting from chopping off the end of words. While our lexicon only contain bigrams and unigrams, we chain together overlapping bigram matches in a sequence to create larger keyphrases. Consider the example where we have a subsequence: "hurricane maria recovery efforts", and we have the following bigram phrases from the lexicon: "hurricane maria", "maria recovery", and "recovery efforts". Thus, there are overlapping matches between the given subsequence and the bigram lexicon phrases. In this case, we can chain the bigram matches together, combining them into a single annotated keyphrase: "hurricane maria recovery efforts". We found that most of the time, phrases from the lexicon appear in the tweets within a similar context (due to the fact that we pre-filtered most tweets that do not occur in a disaster context). This fact mitigates the risk of annotating phrases in an unsuitable context even though our annotation approach does not take context into account.

In addition to hashtags annotated as gold-standard based on the manually constructed lexicon, we also used user-provided hashtags as gold-standard (when available), as they generally capture disaster names and locations, as mentioned before. However, we removed the # sign, and segmented all hashtagged phrases into the constituent words, before annotating them, to ensure that the model learns to distinguish hashtag-like words without relying on the # sign. We also removed user mentions and urls from the tweets.

Benchmark Dataset

To enable progress on hashtag identification in disaster tweets and facilitate models' comparisons, we created a benchmark dataset by splitting our dataset into training, validation and test subsets. The test subset consists of: (1) three disasters that are not represented in the training data, specifically, Maria Hurricane, Philippines Flood and California Fire, and (2) 7% of the data (removed from the training set) from the disasters represented in the training set. The performance of the models is evaluated on each of these four subsets separately. The validation subset consists of the whole Typhoon Pablo data, together with 15% of the data (removed from the training set) from the disasters occurring in the training set.

Quality Assessment

To assess the quality of our semi-automated lexicon annotation, we had human annotators manually inspect the lexicon-based annotations for a sample of tweets, and make an assessment about the annotation of a tweet as appropriate or not appropriate. For this task, we sampled 500 tweets from our dataset (training set), and uploaded them to Amazon Mechanical Turk. The task was to decide whether a given predicted keyphrase is appropriate (as hashtag) for a given tweet or not. The specific options were "Appropriate", "Not Appropriate", and "Unsure". Each tweet was assessed by three annotators. Approximately 89% of the keyphrases had a majority vote for "Appropriate" and only about 3% of the sampled data were consensually voted as "Not Appropriate" by all the three annotators.

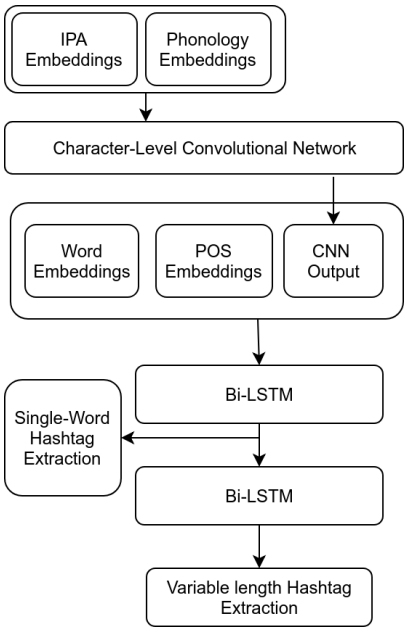

Figure 2: Diagram of the LSTM-MTL model with word embeddings, POS-embeddings, and concatenated IPA and phonological features.

LSTM and Variants

For the core modeling of our dataset, we use the Joint-layer Recurrent Neural Network (RNN) model proposed by Zhang et al. (2016) and investigate several of its variants that capture specifics of informal writing in social media. We chose this model because it achieves state-of-the-art performance on general Tweet dataset (with hashtags treated as ground truth). It significantly outperforms other models such as Marujo's (Marujo et al. 2015) variant of MAUI (Medelyan, Frank, and Witten 2009). Furthermore, traditional keyphrase extraction models such as TF-IDF (Salton and McGill 1986), TextRank (Mihalcea and Tarau 2004), or KEA (Witten et al. 1999) that rely on statistical features like word co-occurences and word counts of terms within a document, are not expected to work well on Twitter data since tweets consist of very short text. Hence, in a tweet, most candidate keyphrases will usually occur only once and a word can only co-occur with very few other different words. In-

Model	Pr	Re	F_1	Pr	Re	F_1
	Maria Hurricane			California Fire		
LSTM	89.65%	83.49%	86.46%	91.21%	85.29%	88.15%
2-layer LSTM	89.81%	83.08%	86.31%	91.09%	84.79%	87.83%
LSTM-MTL	89.87%	83.93%	86.80%	90.96%	85.40%	88.09%
LSTM-MTL+ELMo	90.48%	**85.80%**	88.08%	**92.65%**	**89.19%**	**90.89%**
LSTM-MTL+IPA,POS	89.19%	84.92%	87.00%	91.00%	86.76%	88.83%
LSTM-MTL+ELMo,IPA,POS	**90.97%**	85.59%	**88.20%**	91.88%	87.72%	89.75%
	Philippines Flood			Multiple disasters		
LSTM	85.99%	83.39%	84.67%	92.89%	88.95%	90.88%
2-layer LSTM	85.71%	83.59%	84.64%	92.39%	88.08%	90.18%
LSTM-MTL	85.29%	83.94%	84.61%	92.88%	89.27%	91.04%
LSTM-MTL+ELMo	87.30%	85.52%	86.40%	93.67%	90.55%	92.08%
LSTM-MTL+IPA,POS	87.26%	84.58%	85.90%	93.38%	89.74%	91.52%
LSTM-MTL+ELMo,IPA,POS	**87.88%**	**85.92%**	**86.89%**	**93.83%**	**90.66%**	**92.22%**

Table 3: Precision, Recall, and F1 scores on four test datasets

deed, we find them to perform considerably poorly compared to Joint-layer RNN based models on Twitter data in the work of Zhang et al. (2018). We also compare the Joint-layer RNN with other RNN models (a single layered BiLSTM, and a two layered BiLSTM). We describe Joint-layer RNN model and its variants in what follows. A diagram of the most complete variant is shown in Figure 2.

LSTM-MTL: The Joint-layer RNN is a Bi-LSTM model (Hochreiter and Schmidhuber 1997b; Graves, Jaitly, and Mohamed 2013), trained using Multi-Task Learning (MTL), which stacks two Bi-LSTM layers and jointly trains them on two related tasks. The first Bi-LSTM is trained on the task of identifying single words that are suited to be part of a hashtag (a lower level auxiliary task). The second Bi-LSTM is trained to label hashtag candidate phrases of arbitrary length (the main task, treated as a sequence labeling problem). Similar MLT approaches have been used in other contexts (Søgaard and Goldberg 2016; Liu, Qiu, and Huang 2016). For this model, we used GloVe embeddings (Pennington, Socher, and Manning 2014) pre-trained on Twitter. The embeddings were loaded using Flair (Akbik, Blythe, and Vollgraf 2018). Following Zhang et al. (2016), we represented each word in a sequence as a concatenation of three words, specifically, the current word and its immediate neighbors.

LSTM-MTL+ELMo: In this variant of the Joint-layer RNN, we concatenate the GloVe embeddings with contextualized ELMo word embeddings (Peters et al. 2018). However, to keep the number of parameters smaller, we did not use the three-word window representation as in the original model, as ELMo is already encoding the context.

LSTM-MTL+IPA,POS: This model variant of the Joint-layer RNN is aimed at better handling noise in the data. Hence, we incorporate information about the informal writing, inspired from the methods proposed by Aguilar et al. (2018). They noted how Twitter users often tend to spell words based on their pronunciations, which means it is possible to make more normalized representations of the words by using their phonetics or corresponding IPA (International Phonetic Alphabet) letters, alongside with their phonological features. Following their work, we used Epitran (Mortensen, Dalmia, and Littell 2018) to convert graphemes to phonemes (represented using IPA), and Panphon (Mortensen et al. 2016) to convert each IPA phoneme into a vector representing various phonological (articulatory) features associated with it. We also used part-of-speech (POS) taggers as provided by Owoputi et al. (2013) to explicitly add POS-tag information to this model.

We used randomly initialized embeddings for each POS-tag and IPA symbol. We directly used the phonological vector representations created with Panphon as embeddings for phonological features. We concatenated the embeddings of IPA symbols with their corresponding phonological feature vectors. The result of the concatenation, a character level representation of the phonetics and phonological features for each word, was fed to a character-level CNN (Zhang, Zhao, and LeCun 2015), followed by a global max-pooling layer, to create word level representations. Unlike Aguilar et al. (2018), we chose to use a CNN as opposed to a Bi-LSTM for creating the word level representations because the CNN is much faster. The output of the CNN was concatenated with the POS-tag embeddings and pre-trained GloVe embeddings. The resultant representation was then fed to the stacked Bi-LSTM model. This model uses the three-word window representation on GloVe embeddings.

LSTM-MTL+ELMo,IPA,POS: The last variant of the Joint-layer RNN is a combination of the above two models. It uses ELMo concatenated with GloVe embeddings, POS-tag embeddings, and CNN encoded word-level representations of phonetics and phonological features. Given that ELMo captures context, this model does not use the three-word window representation.

Flood in the UST Hospital is now on the 2nd floor. No food for the patients & staff. Pls. help! #rescuePH @norescu
was so happy to see these two babies pulled out of collapsed building alive heartbreaking bawling my eyes out
Hurricane Maria came ashore in Puerto Rico this morning as a category 4 hurricane with winds of 55 mph
14 people killed and several missing after Cyclone Cleopatra hit Italian island of Sardinia , officials say
pls help: People in Hermosa, Bataan r in roofs now,there's no rescuers helping as of now #rescuePH
Maybe this will help . Please donate Hispanic Federation and direct relief online highly rated by charity naviga
#rescuePH Pregnant mom and small kids trapped in 5 feet flooded house valenzuela city .

Table 4: Examples of tweets from our test dataset. The agreement between predicted hashtags and annotated gold-standard hashtags is marked with blue. Gold-standard hashtags that are not in the set of predicted hashtags are marked with yellow, and predicted hashtags that are not annotated as gold-standard are marked with red. The predictions are made with LSTM-MTL+ELMo, POS, IPA model. (Personal information and urls were removed.)

Experimental Setup and Results

We describe our experimental setting and present results on our four test datasets in this section.

Experimental Settings

We used 100 dimensional Twitter GloVe embeddings, 1024 dimensional ELMo embeddings, 64 dimensional POS-tag embeddings, and 22 dimensional IPA embeddings. Phonological features were represented with a 22 dimensional vector. The embeddings were further fine-tuned during training. Each Bi-LSTM network had 300 hidden units. For the CNN, we used 128 filters and a kernel of size 3. We used dropouts of 0.5 on the input to the Bi-LSTM layers. For optimization, we used the nadam optimizer (Dozat 2016) with a learning rate of 0.0015. Hyper-parameters were either tuned on the validation data or selected based on values that gave good results in prior works.

Results

The results of the experiments are shown in Table 3 for the four test datasets, specifically, Maria Hurricane, California Fire, Philippines Floods, and a dataset sampled from multiple disasters that are explicitly represented in the training data (however, with no overlap between train and test). The multiple disasters dataset is used to evaluate the ability of the models to generalize between similar training and test data, and can be seen as providing an upper bound for the performance of the models. We grouped the results according to the baselines (LSTM, 2-layer LSTM, and LSTM-MTL) and our explored variants of the LSTM-MTL that capture informal writing in social networks. Underlined scores in the table are best within each group and bold scores are best overall. As can be seen from Table 3, our models can generalize well to unseen and underrepresented disasters like California Fire, Philippines Floods, and Maria Hurricane.

Regarding the performance of the models, interestingly, we find in Table 3 that the simpler LSTM models (LSTM and 2-layer LSTM) perform on par with the LSTM with multi-task setup (LSTM-MTL). However, the variants that explicitly incorporate specifics from the informal writing bring further improvements, with the LSTM-MTL+ELMo,IPA,POS model being the best overall (with the exception of California Fire dataset). Between the LSTM-MLT+ELMO and LSTM-MLT+IPA,POS variants, the LSTM-MLT+ELMO variant gives better results overall. We can also observe that the concatenation of the GloVe embeddings with the contextual ELMo embeddings capture better the context of the tweet as compared to the LSTM-MTL model.

Error Analysis and Prediction Quality

To gain insights into the hashtag predictions made by the best performing model, LSTM-MTL+ELMo,IPA,POS, Table 4 shows the predictions on several tweets from our test data, by comparison with the gold-standard annotations. The agreement between model predictions and gold-standard is shown in blue. Predicted hashtags that are not in the gold-standard are marked with red, while gold-standard annotations that are not in the set of predicted hashtags are marked with yellow. As can be seen, the agreement between model predictions and gold-standard is very high. Interestingly our model can predict certain named entities such as 'Hispanic Federation' and ''ataan,' which were beyond the lexicon, but it also misses some other named entities such as 'valenzuela city', and 'UST Hospital'. Overall we can see that the chosen model that capture informal writing is able to make good hashtag predictions for new disasters.

The models described here can be integrated into systems that can help response organizations to have a real time map of a disaster - what is happening on the ground, which could display both the physical disaster and the spikes of intense

activity in the proximity to the disaster. In time, such AI-based systems could have a strong social impact with great benefits to those affected by disasters, and could pinpoint the joy of having survived a falling tree, the horror of a bridge washing out, or the fear of looters in action. Responders will be able to use such a system to provide real time alerts of the status of the disaster and of the affected population.

Conclusion

In this paper, we introduced a new disaster-related dataset of tweets that were annotated with hashtags using a semi-automated lexicon-based approach. This is the first large-scale dataset constructed for identifying relevant and topically informative hashtags for disaster tweets. We believe that our dataset will foster research in this domain, will enable the design of deep learning models, and will help response organizations to make better use of social media data contributed by individuals affected by disasters, and will contribute to better decision-making during disasters when resources are limited. In addition to introducing a new dataset, we built an LSTM-MTL model and explored its variants to capture informal writing in social media. The results show that taking informal writing into account improves the F1-score of LSTM-MTL by up to 2%. This opens up directions for future investigation for more explicitly capturing informal writing into the modeling. Also, incorporating domain knowledge into the models would be expected to improve performance further. In the near future, we wish to bring more attention to building a multilingual setup. Hierarchical multiclass classification with fine and coarse grained labels enabled by our dataset would also be an interesting future direction.

Acknowledgments. We thank the National Science Foundation (NSF) and Amazon Web Services for support from grants IIS-1741345 and IIS-1912887, which supported the research and the computation in this study. We also thank NSF for support from the grants IIS-1526542, IIS-1423337, IIS-1652674, and CMMI-1541155. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF. We also thank our anonymous reviewers for their constructive feedback.

References

Aguilar, G.; Monroy, A. P. L.; González, F.; and Solorio, T. 2018. Modeling noisiness to recognize named entities using multitask neural networks on social media. In *NAACL*, volume 1, 1401–1412.

Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, 1638–1649.

Alam, F.; Joty, S.; and Imran, M. 2018. Domain adaptation with adversarial training and graph embeddings.

Alam, F.; Ofli, F.; and Imran, M. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *ICWSM*.

Bellaachia, A., and Al-Dhelaan, M. 2012. Ne-rank: A novel graph-based keyphrase extraction in twitter. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, 372–379. IEEE Computer Society.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Dozat, T. 2016. Incorporating nesterov momentum into adam.

Frej, W. 2018. Hurricane florence flood victims turn to social media for rescue. *HuffPost*.

Gong, Y., and Zhang, Q. 2016. Hashtag recommendation using attention-based convolutional neural network. In *IJCAI*, 2782–2788.

Graves, A.; Jaitly, N.; and Mohamed, A.-r. 2013. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding*, 273–278. IEEE.

Hochreiter, S., and Schmidhuber, J. 1997a. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Hochreiter, S., and Schmidhuber, J. 1997b. Long short-term memory. *Neural computation* 9(8):1735–1780.

Imran, M.; Elbassuoni, S.; Castillo, C.; Diaz, F.; and Meier, P. 2013a. Practical extraction of disaster-relevant information from social media. In *WWW*, 1021–1024.

Imran, M.; Elbassuoni, S. M.; Castillo, C.; Diaz, F.; and Meier, P. 2013b. Extracting information nuggets from disaster-related messages in social media. *Proc. of ISCRAM, Baden-Baden, Germany*.

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

Lapin, T. 2018. Family's tweet for help leads to rescue during hurricane florence. *New York Post*.

Li, J.; Xu, H.; He, X.; Deng, J.; and Sun, X. 2016a. Tweet modeling with lstm recurrent neural networks for hashtag recommendation. In *2016 International Joint Conference on Neural Networks (IJCNN)*, 1570–1577. IEEE.

Li, Y.; Liu, T.; Jiang, J.; and Zhang, L. 2016b. Hashtag recommendation with topical attention-based lstm. Coling.

Li, Y.; Liu, T.; Hub, J.; and Jiang, J. 2019. Topical co-attention networks for hashtag recommendation on microblogs. *Neurocomputing* 331:356 – 365.

Liu, P.; Qiu, X.; and Huang, X. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

MacMillan, D. 2017. In irma, emergency responders new tools: Twitter and facebook. *The Wall Street Journal*.

Marujo, L.; Ling, W.; Trancoso, I.; Dyer, C.; Black, A. W.; Gershman, A.; de Matos, D. M.; Neto, J.; and Carbonell, J. 2015. Automatic keyword extraction on twitter. In *IJCNLP*, 637–643.

Medelyan, O.; Frank, E.; and Witten, I. H. 2009. Human-competitive tagging using automatic keyphrase extraction. In *EMNLP*, 1318–1327.

Mihalcea, R., and Tarau, P. 2004. Textrank: Bringing order into text. In *EMNLP*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in NeurIPS*, 3111–3119.

Mizuno, J.; Tanaka, M.; Ohtake, K.; Oh, J.-H.; Kloetzer, J.; Hashimoto, C.; and Torisawa, K. 2016. Wisdom x, disaana and d-summ: large-scale nlp systems for analyzing textual big data. In *COLING 2016*, 263–267.

Mortensen, D. R.; Littell, P.; Bharadwaj, A.; Goyal, K.; Dyer, C.; and Levin, L. S. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *COLING*, 3475–3484.

Mortensen, D. R.; Dalmia, S.; and Littell, P. 2018. Epitran: Precision G2P for many languages. In chair), N. C. C.; Choukri, K.;

Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S.; and Tokunaga, T., eds., *LREC*. ELRA.

Olteanu, A.; Castillo, C.; Diaz, F.; and Vieweg, S. 2014a. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM 2014*.

Olteanu, A.; Castillo, C.; Diaz, F.; and Vieweg, S. 2014b. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*.

Olteanu, A.; Vieweg, S.; and Castillo, C. 2015. What to expect when the unexpected happens: Social media communications across crises. In *ACM CSCW*, 994–1009. ACM.

Owoputi, O.; O'Connor, B.; Dyer, C.; Gimpel, K.; Schneider, N.; and Smith, N. A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL*, 380–390.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*, 2227–2237. ACL.

Rhodan, M. 2017. 'please send help.' hurricane harvey victims turn to twitter and facebook. *Time*.

Salton, G., and McGill, M. J. 1986. *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Silverman, L. 2017. Facebook, twitter replace 911 calls for stranded in houston. *NPR*.

Søgaard, A., and Goldberg, Y. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *ACL*, 231–235.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28(1):11–21.

Villegas, C.; Martinez, M.; and Krause, M. 2018. Lessons from harvey: Crisis informatics for urban resilience. *Rice University Kinder Institute for Urban Research*.

Witten, I. H.; Paynter, G. W.; Frank, E.; Gutwin, C.; and Nevill-Manning, C. G. 1999. Kea: Practical automatic keyphrase extraction. In *4th ACM Conference on DL '99*, 254–255. ACM.

Zhang, Q.; Wang, Y.; Gong, Y.; and Huang, X. 2016. Keyphrase extraction using deep recurrent neural networks on twitter. In *EMNLP*, 836–845.

Zhang, Q.; Wang, J.; Huang, H.; Huang, X.; and Gong, Y. 2017. Hashtag recommendation for multimodal microblog using co-attention network. In *IJCAI*, 3420–3426.

Zhang, Y.; Li, J.; Song, Y.; and Zhang, C. 2018. Encoding conversation context for neural keyphrase extraction from microblog posts. In *NAACL*, 1676–1686. ACL.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *NeurIPS*, 649–657.

Zhao, W. X.; Jiang, J.; He, J.; Song, Y.; Achananuparp, P.; Lim, E.-P.; and Li, X. 2011a. Topical keyphrase extraction from twitter. In *ACL*, 379–388. Association for Computational Linguistics.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; and Li, X. 2011b. Comparing twitter and traditional media using topic models. In *Advances in IR*, 338–349. Springer Berlin Heidelberg.