DOI: 10.1111/tgis.12627

RESEARCH ARTICLE

Transactions in GIS WILEY

NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages

Jimin Wang[1] | Yingjie Hu[1] | Kenneth Joseph[2]

[1]GeoAI Lab, Department of Geography, University at Buffalo, Buffalo, NY, USA

[2]Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA

Correspondence
Yingjie Hu, GeoAI Lab, Department of Geography, University at Buffalo, Buffalo, NY 14260, USA.
Email: yhu42@buffalo.edu

Abstract

Social media messages, such as tweets, are frequently used by people during natural disasters to share real-time information and to report incidents. Within these messages, geographic locations are often described. Accurate recognition and geolocation of these locations are critical for reaching those in need. This article focuses on the first part of this process, namely recognizing locations from social media messages. While general named entity recognition tools are often used to recognize locations, their performance is limited due to the various language irregularities associated with social media text, such as informal sentence structures, inconsistent letter cases, name abbreviations, and misspellings. We present NeuroTPR, which is a Neuro-net ToPonym Recognition model designed specifically with these linguistic irregularities in mind. Our approach extends a general bidirectional recurrent neural network model with a number of features designed to address the task of location recognition in social media messages. We also propose an automatic workflow for generating annotated data sets from Wikipedia articles for training toponym recognition models. We demonstrate NeuroTPR by applying it to three test data sets, including a Twitter data set from Hurricane Harvey, and comparing its performance with those of six baseline models.

1 | INTRODUCTION

Social media messages, such as tweets, are frequently used by people during natural disasters or other emergency situations (e.g., the Boston Marathon bombing) to share real-time information and report incidents (Imran, Castillo, Diaz, & Vieweg, 2015; Silverman, 2017; Yu, Huang, Qin, Scheele, & Yang, 2019). Geographic locations are often described in these messages. Consider, for example, the following two tweets posted during Hurricane Harvey in 2017 (where the text of the tweets is slightly revised for privacy protection; see Ayers, Caputi, Nebeker, & Dredze, 2018): "Anyone with a boat in the Meyerland area! A pregnant lady named Nisa is stranded near Airport blvd & station dr #harvey"; and "Rescue: two kids are on the roof at 1010 Bohannon Rd. Please RT #Harvey". Accurately recognizing and geolocating locations from these social media messages are critical for reaching people in need, potentially helping to save human lives.

It is worth differentiating the geographic locations tagged *to*social media messages (i.e., geotagging) and those mentioned *within*the message content. Many social media platforms, including Twitter, allow a message to be associated with the current location of the user. However, the current location of the user is not necessarily the location of the incident; for example, a person may first run to a safe place before sending out a tweet. In discussing the value of tweets for situation awareness, MacEachren et al. (2011) differentiated two types of locations, namely *tweet-from* locations (i.e., geotagged locations) and *tweet-about* locations (i.e., locations mentioned in tweet content). While tweet-from locations are usually in a structured format, tweet-about locations are embedded in natural language text and can be difficult to extract, due to the informal sentence structures of social media content, the variations of a place's name, noise in user-generated text, and other factors. Tweet-about locations have become even more critical, as in June 2019 Twitter announced the removal of its precise geotagging feature. Such a change is likely to lead to a further decrease in the number of geotagged tweets (i.e., tweet-from locations), and makes the task of recognizing and geolocating locations from the content of tweets even more important.

Geoparsing is the process of recognizing place names, or toponyms, from text and identifying their corresponding spatial footprints (Freire, Borbinha, Calado, & Martins, 2011; Gelernter & Balaji, 2013; Gritta, Pilehvar, Limsopatham, & Collier, 2018). As a research topic, geoparsing has been frequently studied in the broader field of geographic information retrieval (GIR; Jones & Purves, 2008; Purves, Clough, Jones, Hall, & Murdock, 2018). A software tool developed for geoparsing is called a *geoparser*. There exist many important applications of geoparsing, one of which is extracting locations from social media messages for disaster response (Gelernter & Mushegian, 2011; Gu, Qian, & Chen, 2016; Inkpen, Liu, Farzindar, Kazemi, & Ghazi, 2017; Wang, Mao, Wang, Rae, & Shaw, 2018; Zhang & Gelernter, 2014).

A geoparser usually functions in two consecutive steps: *toponym recognition* and *toponym resolution*. The first step recognizes toponyms from text without identifying their geographic locations, and the second step resolves any possible place name ambiguity and assigns suitable geographic footprints. Figure 1 shows these two steps of geoparsing. This article focuses on the first step, namely toponym recognition.

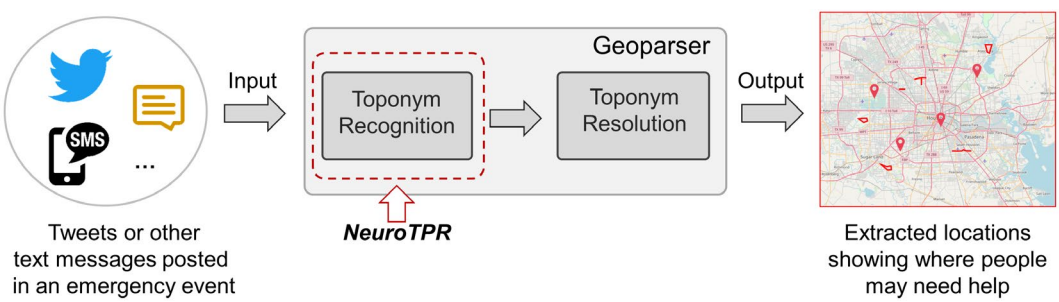

FIGURE 1 The two steps of geoparsing in the context of disaster response and our focus on toponym recognition

Existing research typically uses a *named entity recognition* (NER) tool, such as the Stanford NER (Manning et al., 2014), for toponym recognition. These off-the-shelf tools are designed to recognize locations, but also other kinds of named entities, such as people and organizations. While in theory these off-the-shelf NER tools thus solve the task of toponym recognition, both the literature (Gelernter & Mushegian, 2011; Wang et al., 2018) and our recent work (Hu, Mao, & McKenzie, 2019) have shown the limited performance of the Stanford NER in processing user-generated text that has various language irregularities such as informal sentence structures, inconsistent upper and lower case (e.g., "there is a HUGE fire near camino and springbrook rd"), name abbreviations (e.g., "bsu" for "Boise State University"), and misspellings.

In this work, we propose NeuroTPR, a Neuro-net ToPonym Recognition model for extracting locations from social media messages. NeuroTPR extends a general recurrent neural network model for toponym recognition with a number of enhancements to address language irregularities in social media messages. The contributions of this article are as follows:

1. We propose and develop NeuroTPR as a new toponym recognition model for extracting locations from social media messages that outperforms existing approaches.
2. We propose an automatic workflow for generating data sets with place name annotations for training NeuroTPR and other toponym recognition models.
3. We share the source code of NeuroTPR, the workflow for generating training data, and the annotated test data at https://github.com/geoai-lab/NeuroTPR.

The remainder of this article is organized as follows. Section 2 reviews related work on geoparsing and toponym recognition in the context of disaster response. Section 3 presents the methodological details of NeuroTPR and an automatic workflow for generating training data from Wikipedia articles. Section 4 presents the experiments for training and testing NeuroTPR and discusses the experiment results. A real-world Twitter data set from Hurricane Harvey 2017 is used as one of the three test data sets for comparing NeuroTPR with other baselines. Finally, Section 5 summarizes this work and discusses future directions.

2 | RELATED WORK

Social media messages, such as tweets, are frequently used by people in emergency situations. Crooks, Croitoru, Stefanidis, and Radzikowski (2013) examined the tweets sent after a 5.8 magnitude earthquake occurred on the east coast of the U.S. on August 23, 2011, and found that the first tweet arrived only 54 seconds after the event. Many studies have leveraged the real-time characteristics and rich content of tweets (e.g., texts, images, and geotagged locations) to support situational awareness and disaster response. One of the earliest examples was the work of Starbird and Stamberger (2010), who proposed a Twitter-based hashtag syntax to help people format their disaster-related tweets in a way that could be quickly processed by emergency response organizations. Other examples include studies on tweets from the flooding in Pakistan (Murthy & Longwell, 2013), Hurricane Sandy (Huang & Xiao, 2015), the Boston Marathon bombing (Buntain, Golbeck, Liu, & LaFree, 2016), and Hurricane Irma (Yu et al., 2019). While people in the U.S. can also call 911 for help during disasters, the phone calls of the victims may not get through due to the large volume of calls and failed emergency call centers (Seetharaman & Wells, 2017). During Hurricane Harvey, for example, National Public Radio published an article with the headline "Facebook, Twitter replace 911 calls for stranded in Houston" (Silverman, 2017), which described how these social media platforms were used by Houston residents in the flooding areas to call for help. Similarly, dedicated websites and efforts are also sometimes created and organized during disasters to help people share information. For example, after the 2010 Haiti earthquake, the Ushahidi platform was established which allowed people to send short text messages about their current locations and urgent needs

(Meier, 2010). However, such services rely on word-of-mouth knowledge for usage, compared to the already widespread use of social media.

Given the large number of social media messages posted during an emergency event, it is often necessary to perform automatic information extraction on them. Geoparsing is an effective approach for automatically extracting locations from text, and a number of geoparsers have been developed. GeoTxt, initially developed by Karimzadeh et al. (2013) and further enhanced in their recent work (Karimzadeh, Pezanowski, MacEachren, & Wallgrün, 2019), is a Web-based geoparser that leverages the Stanford NER and several other NER tools for toponym recognition and uses the GeoNames gazetteer and a set of heuristic rules for toponym resolution. TopoCluster is a geoparser developed by DeLozier, Baldridge, and London (2015) which uses the Stanford NER to recognize toponyms from text and then resolves toponyms based on the geographic profiles of the surrounding words (the geographic profile of a word quantifies how frequently this word is used in different geographic areas). Cartographic Location And Vicinity INdexer (CLAVIN) is an open-source geoparser that employs the Apache OpenNLP tool or the Stanford NER for toponym recognition and utilizes a gazetteer, fuzzy search, and heuristics for toponym resolution. The Edinburgh Geoparser was developed by the Language Technology Group at Edinburgh University (Alex, Byrne, Grover, & Tobin, 2015). It uses their in-house natural language processing tool, called LT-TTT2, for toponym recognition, and the toponym resolution step is based on a gazetteer (e.g., GeoNames) and pre-defined heuristics. CamCoder is a toponym resolution method developed by Gritta, Pilehvar, and Collier (2018b) which uses an integration of convolutional neural networks, word embeddings, and geographic vector representations of place names for toponym resolution. Gritta et al. (2018b) further converted CamCoder into a geoparser by connecting it with the spaCy NER tool for toponym recognition. There also exist studies that focus on the step of toponym resolution only (Buscaldi & Rosso, 2008; Ju et al., 2016; Overell & Rüger, 2008; Speriosu & Baldridge, 2013).

For the toponym recognition step, existing geoparsing research has often used an off-the-shelf NER tool. The rationale in doing so is that toponym recognition is often a sub-task of NER. Thus, one can save time and effort by using an existing NER tool and keeping only locations, instead of developing a new model from scratch. However, it has been shown that off-the-shelf NER tools, such as the Stanford NER, have limited performance on informal text written by general Web users (Gelernter & Mushegian, 2011; Hu et al., 2019; Wang et al., 2018). Acknowledging these limitations, scholars have begun to seek improvements over these off-the-shelf models. Most recently (in June 2019), a geoparsing competition, *Toponym Resolution in Scientific Papers*, was held as one of the SemEval 2019 tasks in conjunction with the Annual Conference of the North American Chapter of the Association for Computational Linguistics (Weissenbacher, Magge, O'Connor, Scotch, & Gonzalez, 2019). The top three winning teams all leveraged deep neural network models, such as the bidirectional long short-term memory (BiLSTM) model, to design their geoparsers (Li, Wang, Baldwin, Tomko, & Vasardani, 2019; Wang et al., 2019; Yadav, Laparra, Wang, Surdeanu, & Bethard, 2019). The model that won first place is DM_NLP, which was developed by Wang et al. (2019) and achieved an *F*-score of over 0.9 in the competition. While this competition demonstrated the power of deep learning models for geoparsing, a major limitation is that the models were tested on only a single data set with 45 research papers. The text in these 45 research papers is well formatted and contains relatively simple toponyms such as the names of major cities.

In our latest work (Wang & Hu, 2019a), we systematically tested these three winning deep learning based geoparsers on our benchmarking platform EUPEG (Wang & Hu, 2019b), using eight different corpora with both well-formatted (e.g., news articles) and ill-formatted texts (e.g., tweets and uncapitalized Web text). We compared the deep learning geoparsers (Li et al., 2019; Wang et al., 2019; Yadav et al., 2019) with the existing off-the-shelf geoparsers discussed previously on their performance on both toponym recognition and toponym resolution. Our experiment result suggested that: deep learning based models (such as the BiLSTM model adopted by all three winning teams) usually outperform traditional machine learning models in toponym recognition across different types of texts; but that while showing high performance on well-formatted text, these deep learning geoparsers performed poorly on user-generated text, in particular on data without capitalization, such as the *Ju2016* data set

(Ju et al., 2016). Our NeuroTPRmodel aims to address these limitations and improve toponym recognition from social media messages.

3 | METHODS

3.1 | Model architecture

NeuroTPR is designed based on the basic BiLSTM–conditional random field (CRF) model proposed by Lample, Ballesteros, Subramanian, Kawakami, and Dyer (2016), which achieved state-of-the-art performance on a general NER benchmarking task. With this basic model, we add a number of improvements to develop NeuroTPR. Figure 2 provides an overview of the model.

We present NeuroTPR from bottom to top, characterizing the *layers* of the neural network. *Layer 0* contains the individual words of a tweet, which are used as the input to the model. The next four layers represent each word as vectors using four different approaches. *Layer 1* and *Layer 2* use character embeddings to model each word as a sequence of characters. The case-sensitive character embeddings in Layer 1 use different vectors to represent the upper and lower cases of the same character, while the caseless character embeddings in Layer 2 use the same vector to represent a character regardless of its case. Both case-sensitive and caseless character embeddings are modeled using the BiLSTM architecture in Figure 3. Character embeddings capture the morphological features of words which can be useful for toponym recognition; for example, words with certain prefixes or suffixes may have higher probabilities of representing locations. In addition, character embeddings are good at handling misspellings in user-generated text, since the semantics of a word is still largely captured when a user misspells or misses a character when typing a word (e.g., typing "t" rather than "y", or skipping this letter completely).

Layer 3 uses pre-trained word embeddings to represent the words in a tweet. These embeddings are pre-trained on a large set of unlabeled text, and can capture the semantics of a word based on the other words that

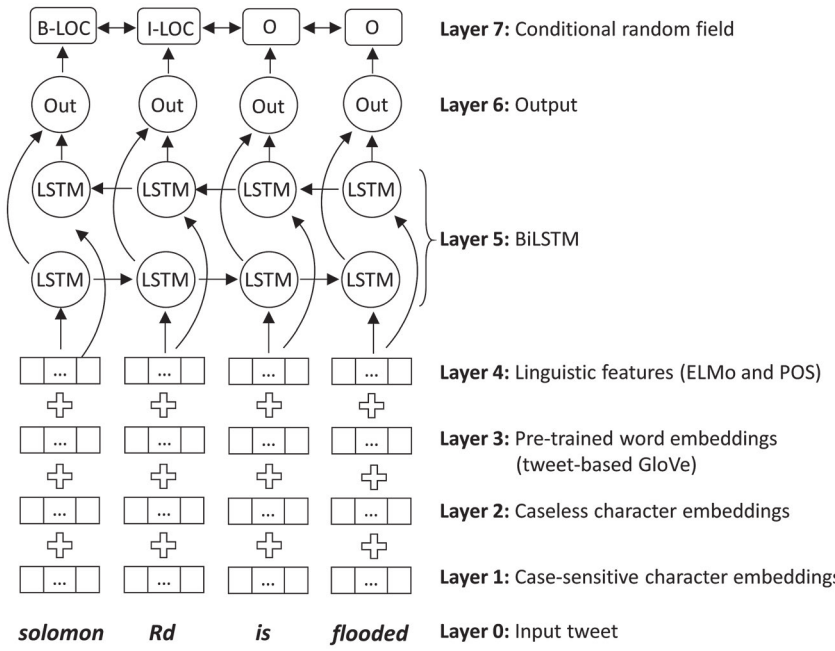

FIGURE 2 Overall architecture of NeuroTPR

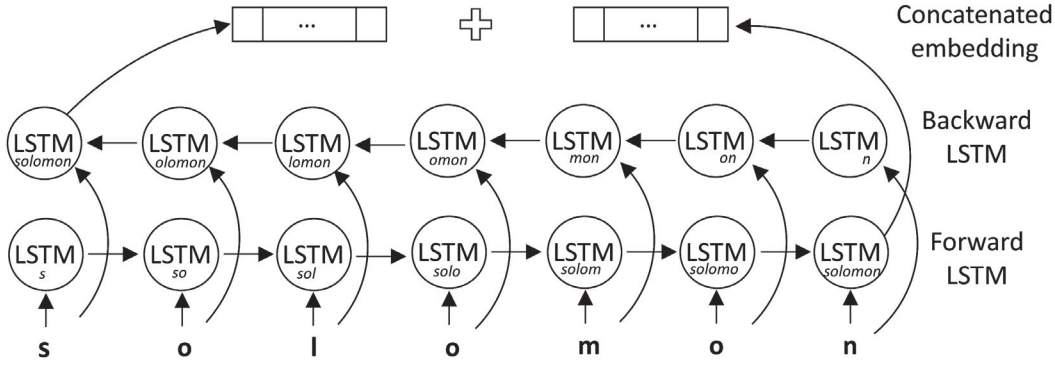

FIGURE 3 BiLSTM architecture modeling a word as a sequence of characters

typically co-occur with it. Compared with character embeddings that focus on the morphological features of a word itself, word embeddings help determine whether a word represents a location based on the typical context within which the word is used. These pre-trained word embeddings are fixed during the training process. *Layer 4* provides linguistic features derived from the words of a tweet to help recognize toponyms. In particular, we include two types of linguistic features: part-of-speech (POS) tags and a type of deep contextualized word embeddings, embeddings from language models (ELMo) (Peters et al., 2018). POS tags inform the model about the type of a word (noun, verb, adjective, preposition, etc). These POS tags help NeuroTPR learn the usage patterns related to locations, (e.g. a location phrase is often used after a preposition). ELMo captures the different semantics of a word under varied contexts. Note that the pre-trained word embeddings in Layer 3 capture the semantics of words based on their typical usage contexts and therefore provide *static* representations of words; by contrast, ELMo provides a *dynamic* representation for a word by modeling the particular sentence within which the word is used.

These four layers capture four different aspects of a word, and their representation vectors are concatenated together into a large vector to represent each input word. These vectors are then used as the input to *Layer 5*, which is a BiLSTM layer consisting of two layers of LSTM cells: one forward layer capturing information before the target word and one backward layer capturing information after the target word. *Layer 6* combines the outputs of the two LSTM layers and feeds the combined output into a fully connected layer. *Layer 7* is a CRF layer which takes the output from the fully connected layer and performs sequence labeling. The CRF layer uses the standard IOB model from NER research to label each word but focuses on locations. Thus, a word is annotated as either "B-LOC" (the beginning of a location phrase), "I-LOC" (inside a location phrase), or "O" (outside a location phrase).

NeuroTPR has several design features that enhance its performance on the task of toponym recognition from social media messages. First, NeuroTPR integrates both case-sensitive and caseless character embeddings. Previous research often used case-sensitive character embeddings only. While using different representations for upper- and lower-case characters helps the model make use of case information (which can be especially helpful for processing well-formatted text, such as news articles), this design makes the model overly sensitive to the irregular capitalization in some user-generated text. An alternative is to use caseless character embeddings only. However, this alternative can miss the useful information passed by many correct uses of letter case. Thus, NeuroTPR integrates both case-sensitive and caseless character embeddings to overcome this issue. Second, NeuroTPR uses the pre-trained word embeddings that are specifically derived from tweets. We use the GloVe word embeddings that were trained on 2 billion tweets with 27 billion tokens and 1.2 million vocabulary items (Pennington, Socher, & Manning, 2014). These word embeddings, specifically trained on a large tweet corpus, include many vernacular words and abbreviations used by people in tweets. Previous geoparsing and NER models typically use word embeddings trained on well-formatted text, such as news articles, and many vernacular words are not be covered by those embeddings. When that happens, an embedding for a generic *unknown* token is usually used to represent

this vernacular word and, as a result, the actual semantics of the word are lost. Third, compared with the basic BiLSTM–CRF model from Lample et al. (2016), NeuroTPR adds ELMo and POS to capture the dynamic and contextualized semantics of words and their POS types. Compared with the DM_NLP model of Wang et al. (2019) from the SemEval 2019 competition, NeuroTPR removes chunking and NER tags which tend to be erroneous when applied to text with informal sentence structures. Besides, NeuroTPR adds an extra layer of caseless character embedding and integrates tweet-based GloVe word embeddings, both of which were not used in the two previous models.

3.2 | Training data sets

We train NeuroTPR using two data sets. The first is an existing and human-annotated Twitter data set that we obtained from the *WNUT 2017 Shared Task on Novel and Emerging Entity Recognition* (Derczynski, Nichols, van Erp, & Limsopatham, 2017). This is a real Twitter data set which contains toponyms, along with other types of entities, annotated by human annotators. We select 599 tweets from this data set which contain toponyms and we keep only toponyms in the annotations. However, the WNUT data set is small, and training a deep learning model usually requires a large amount of annotated training data. Manually generating such training data is a labor-intensive and time-consuming process.

Within this context, we propose a workflow to automatically generate annotated data which will be used as the second data set for training NeuroTPR. This automatic workflow makes use of the first paragraphs of Wikipedia articles that often contain rich annotations of the mentioned entities in the form of hyperlinks. We generate an annotated training data set by extracting these first paragraphs from a Wikipedia dump and retaining only the phrases whose hyperlinks point to articles about geographic location. In our pilot experiments, we determined whether a hyperlink pointed to a location or not by examining whether the corresponding Wikipedia article was geotagged with a pair of latitude and longitude coordinates (Hecht & Moxley, 2009). However, we soon discovered that Wikipedia articles tagged with coordinates were not necessarily about locations. For example, the Wikipedia article about *Normandy landings* (https://en.wikipedia.org/wiki/Normandy_landings) is tagged with a pair of coordinates which indicates the location of this important military event; we would typically not consider *Normandy landings* as a toponym. Eventually, we found that Infobox templates from Wikipedia on geography and place (https://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes/Geography_and_place) are the ideal tool for determining whether or not a hyperlink points to a location.

We therefore developed a method to check each hyperlink to determine whether or not it was a location using this information. Specifically, if the Infobox of the pointed Wikipedia article was consistent with one of the geography templates, the hyperlink was kept as a toponym annotation; otherwise, the hyperlink was removed. Since the data are generated for training NeuroTPR on the task of processing tweets, we make the data more similar to tweets by splitting the Wikipedia paragraphs into sentences and keeping only those within 140 characters (140 characters are used because one of the test data sets in the later experiments contains tweets from Hurricane Harvey which were collected before Twitter expanded its length limitation to 280 characters in November 2017). One can change this setting to 280 characters in the workflow depending on the application need. We also considered an additional strategy for creating training data of *random flipping* in order to make the generated training data even more similar to tweets. We develop a program that goes into each word of the generated training data and randomly changes or removes one character of the word with a probability of 2%, thus simulating misspelling errors often contained in user-generated text. As will be shown in the experiments below, this *random flipping* strategy failed to improve model performance. However, we gained valuable insight into model performance through the experiments to test this strategy, and thus will discuss it further below.

In sum, we use two data sets to train NeuroTPR. The first, WNUT2017, is a small but real Twitter data set annotated by humans, and the second is a larger data set automatically generated from Wikipedia articles using a proposed workflow. It is worth noting that the second data set can be of arbitrary size since it is automatically

generated. In addition, the data generated from Wikipedia articles are only used for training and are not used for any testing experiments. We share the source code of the developed automatic workflow at https://github.com/geoai-lab/NeuroTPR.

4 | EXPERIMENTS

4.1 | Test data sets, evaluation metrics, and baseline models

In a previous work we developed a benchmarking platform called EUPEG, which is an Extensible and Unified Platform for Evaluating Geoparsers (Wang & Hu, 2019b). EUPEG was developed for evaluating geoparsers as complete pipelines, that is, for both toponym recognition and toponym resolution. In this work we will leverage some resources from EUPEG but will focus on the toponym recognition step. While EUPEG provides eight annotated corpora collected from the literature, only one corpus, GeoCorpora (developed by Wallgrün, Karimzadeh, MacEachren, & Pezanowski, 2018), contains social media messages (tweets). Many toponyms in GeoCorpora refer to large-scale geographic features, such as continents (e.g., Africa), countries (e.g., U.S. and Ukraine), states (e.g., California and Alabama), and major cities (e.g., New York City and London). Fine-grained toponyms (e.g., street names) have only limited coverage in GeoCorpora but are often seen in tweets sent out during a disaster. We will still use GeoCorpora as one of our test data sets, but will create another test data set, called Harvey2017, with 1,000 human annotated tweets derived from a large Twitter data set collected during a major disaster, Hurricane Harvey. Finally, we also use Ju2016 (Ju et al., 2016) as a test data set. Ju2016 is not a social media message data set; it contains sentences automatically extracted from Web pages. One special feature of Ju2016, however, is that all characters are in lower case; it has no capitalization. Our previous experiments showed that many geoparsers completely failed on such a data set without capitalization (Wang & Hu, 2019a). Therefore, although Ju2016 is not a social media data set, it is still interesting to see the performance of different models on it. GeoCorpora can be downloaded from the GitHub site of its authors (https://github.com/geovista/GeoCorpora), and Ju2016 can be downloaded from our GitHub site (https://github.com/geoai-lab/EUPEG). In what follows, we describe the process of creating the Harvey2017 data set.

The original Hurricane Harvey Twitter data set is available from the library repository of North Texas University (https://digital.library.unt.edu/ark:/67531/metadc993940/). It was collected between August 18 and September 22, 2017. It contains 7,041,866 tweets retrieved based on a set of hashtags and keywords, such as #HurricaneHarvey, #Harvey2017, and #HoustonFlood. A manual examination of this data set shows that many tweets contain disaster-related information (e.g., floods) and often describe detailed locations such as street names, road intersections, and door number addresses. The content of these tweets and the fact that they were posted during a major disaster make this data set especially suitable for testing NeuroTPR.

We use the following steps to create a manually annotated data set with 1,000 tweets. First, we create a regular expression with about 70 terms related to location descriptions, such as “street”, “avenue”, “park”, “square”, “bridge”, “rd”, and “ave”. We run it against the entire data set and obtain a subset of 15,834 tweets that are more likely to contain specific locations. We randomly select 1,000 tweets from this subset and manually annotate the locations contained in them. Since some tweets from the first batch of 1,000 do not contain specific locations (e.g., a tweet may say, “My side street is now a rushing tributary.”), we replace those tweets with others randomly selected from the remainder of the subset during the manual annotation process, so that each of the 1,000 tweets in this data set contains at least one specific location. These location descriptions densely contained in the 1,000 tweets provide abundant opportunities for testing the performance of a toponym recognition model. It is worth noting, however, that a model needs not only to determine whether a location exists in a tweet but also to find out how many locations exist (many tweets contain two or more locations) and the positions (i.e., character indices) of these locations. Nevertheless, one could design a model that might achieve a fair performance on this particular

data set by assuming that each tweet has at least one location. NeuroTPR does not make such an assumption. Future work could add tweets that do not contain locations to expand this data set and test performance along these lines.

Annotating locations from text, however, is not a straightforward task. As previously discussed by other researchers (Gritta, Pilehvar, & Collier, 2018a; Wallgrün et al., 2018; Zhang & Gelernter, 2014), the concept of "location" can be elusive and the same phrase can be annotated as a location or not depending on the definition adopted by a particular data set. For example, in the *LGL* data set, Lieberman, Samet, and Sankaranarayanan (2010) considered demonyms, such as "Canadian" and "Australian", as toponyms, and geolocated them to the centers of the corresponding countries. While these demonyms have some geographic meaning, broadly speaking, they are unlikely to be considered as toponyms in some domains, including geography. For this data set of Hurricane Harvey tweets, we annotate the following as locations:

1. administrative place names, such as neighborhoods, towns, cities, states, and countries;
2. names of natural features, such as rivers, mountains, and bayous;
3. names of facilities and landmarks, such as roads, bus stops, buildings, and airports;
4. organizations that have fewer than three instances in the *target region*, such as "Heritage Park Baptist Church" and "Cypress Ridge High School".

Here, the *target region* refers to the geographic area affected by the disaster. The following are not considered as valid locations and therefore are not annotated:

1. demonyms, such as American and Texan;
2. metonymies, such as in "Washington made a decision that …";
3. general location references, such as "this building" and "that road";
4. organizations that have many instances in the target region, such as in "I'm stuck at Walmart".

The last point is probably debatable since "Walmart" or other chain stores mentioned in such a sentence could be considered as a location. From a disaster response perspective, however, we argue that annotating "Walmart" in this case will probably add more noise than signal to the extracted locations, since all Walmarts in the target region may show up on a map if this location is to be geolocated in a next step. This debatable issue illustrates part of the difficulty in annotating locations from text. When using a corpus for testing experiments, we need to consider its location definition which can directly affect the annotated ground truth. With the above guideline for location annotation, we create a data set with 1,000 annotated tweets. We will primarily use this *Harvey2017* data set for our evaluation experiments and discussion, since it better fits our disaster response application. However, GeoCorpora and Ju2016 are used as well to provide a more comprehensive evaluation of NeuroTPR on multiple data sets.

The evaluation metrics used in the experiments are precision, recall, and *F*-score:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$F\text{- score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Precision measures the percentage of correctly identified toponyms (true positives, *TP*) among all the toponyms recognized by a model, which include both true positives and false positives (*FP*). Recall measures the percentage of correctly identified toponyms among all the toponyms that are annotated as ground truth, which include true positives and false negatives (*FN*). The *F*-score is the harmonic mean of precision and recall. It is high when both precision and recall are fairly high, and is low if either of the two is low. These metrics have been widely used in previous studies, such as Inkpen et al. (2017), Karimzadeh (2016), Leidner (2008), and Lieberman et al. (2010).

We compare NeuroTPR to three off-the-shelf tools and two deep learning based models. For the off-the-shelf NER tools, we use the Stanford NER, the caseless Stanford NER, and the spaCy NER, which are frequently used in geoparsing research for the step of toponym recognition. Applying these off-the-shelf NER tools to a data set also involves some design choices. With the typically used three-class Stanford NER and its caseless version, the output contains three classes (i.e., PERSON, ORGANIZATION, LOCATION). One can choose to keep only LOCATION in the output, or keep both LOCATION and ORGANIZATION for a wider coverage to include schools, churches, and other similar entities in the output as well. At first glance, it seems to be a wise decision to include ORGANIZATION in the output, since organizations such as schools and churches are also annotated as locations in the Harvey2017 data set. However, including ORGANIZATION does not necessarily increase the performance of the model compared with using LOCATION alone, since there are also organizations that should not be annotated as locations. For example, in the sentence "Donations to the Red Cross have provided help for people impacted by Hurricane Harvey", "Red Cross" will be mistakenly included in the model output. A similar situation happens to the spaCy NER whose output contains multiple classes related to geography including FACILITY (e.g., buildings, airports, and highways), ORG (e.g., companies, agencies, and institutions), GPE (e.g., countries, cities, and states), and LOC (e.g., non-GPE locations, mountain ranges, and bodies of water). Keeping only LOC in the output will exclude other valid location mentions (e.g., cities), while keeping all related classes will include more phrases that should not be considered as locations. This difficult design choice highlights another problem in directly using a general NER tool for toponym recognition. In our experiments, we test two versions for each of the three off-the-shelf NER tools: one version has a *narrow definition* of location using LOCATION or LOC only, while the other version has a *broad definition* of location by including multiple classes that can be related to locations. Since the Stanford NER offers the option to be retrained, we also test a retrained version of the Stanford NER using the same training data as used by NeuroTPR. In addition, we test two deep learning based toponym recognition models: the basic BiLSTM–CRF model by Lample et al. (2016) based on which our NeuroTPR is developed, and the DM_NLP model (using its toponym recognition part only) by Wang et al. (2019) which achieved the best performance in the 2019 SemEval geoparsing competition. These two models are trained using the same training data as used by NeuroTPR.

4.2 | Experimental procedure and results

As we have two data sets to train NeuroTPR, we begin our experiments by evaluating the effectiveness of different training strategies. Specifically, we experiment with eight different strategies to train NeuroTPR, and their performance based on the Harvey2017 data set is reported in Table 1.

The first strategy (S1) uses the WNUT2017 data set only to train NeuroTPR. While this is a small data set, it can already help NeuroTPR achieve a fair performance, reaching a *F*-score of 0.656. The effectiveness of WNUT2017 can be attributed to its ability to help NeuroTPR learn the informal language structure used in tweets. In strategies S2 to S4, we test the performance of NeuroTPR when it is trained on automatically generated Wikipedia data sets. While our proposed workflow can generate a training data set of arbitrary size, generating a larger data set and training the model on such a data set are more time-consuming. We test the performance of NeuroTPR when it is trained on the data sets generated from 1,000, 3,000, and 5,000 Wikipedia articles randomly selected from a Wikipedia dump. Strategies S5 to S7 use a similar idea but add random flipping to the training data set (i.e., a character of a word is randomly changed or removed with a probability of 2%).

TABLE 1 Performance of NeuroTPR on Harvey2017 using different training strategies

Training strategy	Precision	Recall	*F*-score
S1: WNUT2017 only	.687	.633	.656
S2: 1,000 Wikipedia articles	.551	.392	.458
S3: 3,000 Wikipedia articles	.573	.468	.516
S4: 5,000 Wikipedia articles	.547	.481	.512
S5: 1,000 Wikipedia articles + random flipping	.558	.324	.410
S6: 3,000 Wikipedia articles + random flipping	.566	.359	.439
S7: 5,000 Wikipedia articles + random flipping	.520	.410	.459
S8: 3,000 Wikipedia articles + WNUT2017	**.787**	**.678**	**.728**

Note: The bold values stand for the highest value of precision, recall, and F-score achieved among the eight strategies.

Two observations are obtained from these six experiments (S2 to S7). First, the performance of the model does not necessarily increase with more training data. In fact, the result of S4 is worse than that of S3 which uses fewer Wikipedia articles. The automatically generated training data are not perfect, since some Wikipedia articles do not annotate all the toponyms mentioned in the text. As a result, using more Wikipedia articles may also introduce more noise into the training data. Second, adding random flipping does not improve the performance of the model. This result is surprising, as we expected that adding random flipping would make the training data more similar to user-generated text and therefore improve the performance of the trained model.

To understand why random flipping fails, we carefully examine the training process. We find that when simulated misspellings are present in the training data, they will all be represented with the embedding for the *unknown* token, since such misspelled words do not exist in the vocabulary of the pre-trained word embeddings. Consequently, those randomly flipped words confuse, rather than help, the model during the training process. Meanwhile, when misspellings do exist in the test data, they can be partially handled by the character embeddings included in our model design. Thus, a misspelled word, such as "Californa" in the sentence of "Leaving Texas and heading to Californa", can still be recognized by NeuroTPR even when it is not trained on a data set with simulated misspellings. In the last strategy, we use a combination of 3,000 Wikipedia articles without random flipping and the WNUT2017 data set for training, and obtain the best precision, recall, and *F*-score among all the tested strategies.

With the most effective training strategy identified, we continue our experiments by comparing NeuroTPR with the three off-the-shelf NER tools (each has two versions), the retrained Stanford NER, and two deep learning based models. The performance of these models on the Harvey2017 data set is reported in Table 2. Note that *narrow location* means we only keep LOCATION or LOC in the output of the model, whereas *broad location* means we keep all the entity types that are likely to contain locations (i.e., LOCATION and ORGANIZATION for the Stanford NER, both default and caseless, and LOC, ORG, FACILITY, and GPE for the spaCy NER).

As can be seen, the performance of the four off-the-shelf Stanford NER models and the retrained Stanford NER dominates that of the spaCy NER. In particular, the default Stanford NER with LOCATION only (i.e., narrow location) achieves the highest precision among all the models. This performance is impressive and demonstrates the effectiveness of this classic NER tool. However, this Stanford NER also has a low recall of 0.399, which suggests many correct locations are not recognized. If we put this low recall in the context of disaster response, this result suggests that applying an off-the-shelf Stanford NER to the posted tweets will miss over 60% of valid location mentions. A closer examination of the results of this Stanford NER shows that most of the correctly recognized toponyms are city and state names, such as Houston and Texas, while many fine-grained toponyms, such as street names, school names, and church names, are missed. However, these fine-grained toponyms are critical for locating the people who may need help during and after disasters. Including both LOCATION and

TABLE 2 Performance of different tools and models on the Harvey2017 data set

Model	Precision	Recall	*F*-score
Stanford NER (narrow location)	**.828**	.399	.539
Stanford NER (broad location)	.729	.440	.548
Retrained Stanford NER	.604	.410	.489
Caseless Stanford NER (narrow location)	.803	.320	.458
Caseless Stanford NER (broad location)	.721	.336	.460
spaCy NER (narrow location)	.575	.024	.046
spaCy NER (broad location)	.461	.304	.366
Basic BiLSTM–CRF (Lample et al., 2016)	.703	.600	.649
DM_NLP (toponym recognition) (Wang et al., 2019)	.729	**.680**	.703
NeuroTPR	.787	.678	**.728**

Note: The bold values stand for the highest value of precision, recall, and F-score achieved among the eight strategies.

ORGANIZATION in the output of the Stanford NER (i.e., broad location) increases the recall score but decreases the precision. Interestingly, the retrained Stanford NER performs worse than the default Stanford NER (note that the retrained Stanford NER is still case sensitive). The default Stanford NER was trained on a variety of annotated corpora including the CoNLL 2003 training set, MUC 6 and MUC 7 training data sets, ACE 2002, and their in-house data. This training data difference may have contributed to the better performance of the default Stanford NER. The two versions of the caseless Stanford NER achieve lower performance than the default versions as well. While there is some irregular capitalization in tweets, many of them do use standard upper and lower case. Since the caseless Stanford NER models do not use letter case as an input feature, they miss the useful information contained in much correct capitalization.

The two deep learning models are more challenging baselines, and DM_NLP achieves the best score for recall. However, NeuroTPR shows the best performance overall as demonstrated by its highest *F*-score. Compared with the basic BiLSTM–CRF model, NeuroTPR performs better in all three metrics which demonstrate the value of our improved designs, including the double layers of character embeddings, tweet-based GloVe, and ELMo. Compared with DM_NLP, NeuroTPR shows higher precision and *F*-score and similar recall.

We look further into the output of NeuroTPR to understand the errors. We find three major types of errors. First, NeuroTPR seems to often miss interstate highway names, such as "I-45" in a tweet like "Traffic is fluid on I-45." Interestingly, the Stanford NER seems to make the same mistakes as well. Interstate highway names are an important type of place name in the U.S.A, and are likely to be used by people in future disasters to describe locations.

Second, in many cases, NeuroTPR recognizes part of a complete street name. For example, it can recognize "18th Rd" in "E 18th Rd" while missing the "E". Such a result is considered as both a false positive and a false negative in the scores reported in Table 2, since we require *exact matching* between the extracted road names and the annotation. If we allow *inexact matching*, this example could be considered as correct.

Third, NeuroTPR fails to recognize some toponyms when they show up at positions very different from their typical positions in a regular sentence. For example, some tweets simply append one or multiple city names (e.g, "Port Arthur") at the end of the text body, probably for the purpose of textually tagging the affected geographic regions. These toponyms are sometimes missed.

One simple way to address the first issue might be to use an extra regular expression, such as "I-/d+", to identify interstate highway names and include them in the model output. A similar strategy could be applied to the second issue. For example, NeuroTPR can be first used to identify street names, and then a regular expression is used to check whether an indication of cardinal directions is used as a prefix or suffix to the street names. However, those strategies could also introduce false positives or new errors, and need to be empirically tested via experiments.

It is also worth noting that NeuroTPR is not trained using any of the Hurricane Harvey tweets. While WNUT2017 is also a tweet-based data set, its content is very different from the Harvey2017 data set that focuses on seeking help or sharing location-based disaster information. Training NeuroTPR using some tweets from the large Hurricane Harvey data set is likely to increase the performance of the model on the Harvey2017 data set. We test the possibility of further training NeuroTPR using 50 tweets from the Hurricane Harvey data set (in addition to the existing training data) but outside of the 1,000 test tweets, and obtain 0.832 precision, 0.843 recall, and 0.837 *F*-score. Training the model using the tweets from a specific event, however, could lead to overfitting. Besides, we do not always have the necessary data to retrain a model. Imran, Mitra, and Castillo (2016) discussed a strategy of employing the Standby Task Force volunteers to annotate the purposes of social media messages (e.g., *donation needs* and *caution and advice*) during a crisis event in real time and then using the annotated data to train machine learning models rapidly. A similar idea could be adopted to obtain annotated data for training a toponym recognition model for a particular disaster.

We also test the performance of the baseline models and NeuroTPR on GeoCorpora, and the results are reported in Table 3. GeoCorpora considers administrative places, natural features, facilities, and organizations (such as schools and churches) as locations, and does not include demonyms or metonymies. We see a performance increase for most tested models. As discussed previously, a majority of the toponyms in GeoCorpora are names of countries, states, and cities. Thus, GeoCorpora can be considered as an easier test data set than Harvey2017. However, a similar pattern of model performance is observed, with the default Stanford NER achieving the top precision and NeuroTPR achieving the best performance overall.

Finally, we test the performance of the models on the Ju2016 data set. The text records are from Web pages and do not have capitalization. Due to the data generation process, Ju2016 does not annotate all toponyms contained in the text. Therefore, the performance of the models can only be measured using the metric of *accuracy* which is calculated as:

$$\text{Accuracy} = \frac{|\text{Annotated} \cap \text{Recognized}|}{|\text{Annotated}|}, \tag{4}$$

where *Annotated* represents the set of toponyms in the ground truth, and Recognized represents the set of toponyms recognized by a model from the text. Note that accuracy is a typical metric for evaluating geoparsing models when the test corpus does not have complete annotation of all the toponyms. It has been used in previous research (i.e., Gelernter & Mushegian, 2011; Gritta et al., 2018; Karimzadeh, 2016). The performance of the models on Ju2016 is reported in Table 4. As can be seen, some off-the-shelf NER tools that rely on proper letter case, such as the default

TABLE 3 Performance of different tools and models on the GeoCorpora data set

Model	Precision	Recall	*F*-score
Stanford NER (narrow location)	**.899**	.526	.664
Stanford NER (broad location)	.751	.553	.637
Retrained Stanford NER	.590	.364	.450
Caseless Stanford NER (narrow location)	.898	.487	.631
Caseless Stanford NER (broad location)	.774	.503	.610
spaCy NER (narrow location)	.503	.037	.069
spaCy NER (broad location)	.579	.453	.508
Basic BiLSTM–CRF (Lample et al., 2016)	.631	.527	.574
DM_NLP (toponym recognition) (Wang et al., 2019)	.797	.650	.715
NeuroTPR	.800	**.761**	**.780**

Note: The bold values stand for the highest value of precision, recall, and F-score achieved among the eight strategies.

TABLE 4 Performance of different tools and models on the Ju2016 data set

Model	Accuracy
Stanford NER (narrow location)	.010
Stanford NER (broad location)	.012
Retrained Stanford NER	.078
Caseless Stanford NER (narrow location)	.460
Caseless Stanford NER (broad location)	.514
spaCy NER (narrow location)	.000
spaCy NER (broad location)	.006
Basic BiLSTM–CRF (Lample et al., 2016)	.595
DM_NLP (toponym recognition) (Wang et al., 2019)	.723
NeuroTPR	**.821**

Note: The bold values stand for the highest value of precision, recall, and F-score achieved among the eight strategies.

Stanford NER and the spaCy NER, fail on this data set without capitalization. This result is consistent with our previous research (Wang & Hu, 2019a) and echos the point made by Gritta, Pilehvar, Limsopatham, et al. (2018), that some geoparsers do not work on text without capitalization. Interestingly, DM_NLP, which almost failed on Ju2016 in our previous study (Wang & Hu, 2019a), achieves a fair performance in this experiment. To understand the reason, we examine the experimental design in this and our previous studies. The DM_NLP in our previous experiment was trained on the CoNLL 2003 data set whose annotations contain only well-formatted texts with proper capitalization from news articles. The DM_NLP in this experiment is trained on WNUT2017 and 3,000 Wikipedia articles, and WNUT2017 data contain some training instances without proper capitalization. With components such as character embeddings and contextualized word embeddings, a deep learning model like DM_NLP seems to quickly adapt to the training data used. However, the Stanford NER retrained using the same data is still case sensitive and largely fails on Ju2016. This result suggests that deep learning models may be better able to adapt to training data than traditional machine learning models with handcrafted input features.

5 | CONCLUSIONS AND FUTURE WORK

Social media messages, such as tweets, have frequently been used by people during disasters to share information and seek help. The locations described in these messages are critical for first responders to reach the people in need. This article has presented NeuroTPR, a neuro-net toponym recognition model for extracting locations from social media messages. A major advantage of NeuroTPR is its ability to recognize many (fine-grained) toponyms that are otherwise missed by off-the-shelf NER tools commonly used for toponym recognition. For example, compared with the default Stanford NER with only LOCATION in the output, NeuroTPR can correctly recognize about 70% more toponyms according to our experimental results. NeuroTPR is designed based on a general BiLSTM–CRF architecture, and includes a number of improved designs, such as caseless and case-sensitive character embeddings, tweet-based word embeddings, and contextualized word embeddings, to enhance its performance on toponym recognition from social media messages. We train NeuroTPR using an existing human-annotated Twitter data set and a Wikipedia-based data set automatically generated using a developed workflow. We test different training strategies and find that a combination of the human-annotated tweets and automatically generated data yields the best performance. Evaluation experiments based on three test data sets, namely Harvey2017, GeoCorpora, and Ju2016, demonstrate the improved performance of NeuroTPR in comparison with three off-the-shelf NER tools, one retrained NER tool, and two deep learning models. We have shared the source code of

NeuroTPR, the automatic workflow for generating training data, and the Harvey2017 data set to support future research.

Several directions could be pursued to expand this research. First, toponym recognition is only the first step of geoparsing, and we see great promise in integrating NeuroTPR with a toponym resolution model to develop a complete geoparser. A number of toponym resolution models already exist and are discussed in this article (i.e., DeLozier et al., 2015; Gritta et al., 2018b; Ju et al., 2016; Overell & Rüger, 2008). However, these toponym resolution models mainly focus on cities, states, countries, or other large-scale toponyms rather than fine-grained locations such as street names. Further, some street names are highly ambiguous (e.g., there are thousands of Main Streets in the U.S.), and this high ambiguity can make the problem of toponym resolution more complex. For applications to disaster response, the complexity of this problem can be largely reduced by focusing on the disaster-affected area and using a local gazetteer. Thus, instead of performing place name disambiguation on a street name with thousands of instances throughout the world, the model may only need to differentiate among two or three streets for a name in the local area, and may not even need to perform disambiguation at all. Second, we can further enhance the performance of NeuroTPR by testing other similar model architectures. Given the flexibility of deep neural networks, we can add more layers, change the number of neurons, try new activation functions, and test other hyperparameter combinations. Evolutionary algorithms could be employed in this area to help identify a better model architecture. Third, existing geoparsers often geolocate a toponym to a single point while more detailed spatial footprints, such as lines and polygons, are needed for applications such as disaster response. For a sentence like "major flooding along Clay Rd", a line of the road is probably a better representation than a point at the middle of the street. One factor causing this limited spatial representation is the use of the GeoNames gazetteer in most geoparsers, which contains only point-based locations. Other geographic data sets and methods could be explored to provide more detailed spatial footprints for the toponyms recognized from social media messages.

ORCID

Yingjie Hu http://orcid.org/0000-0002-5515-4125

REFERENCES

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the Edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing*, *9*(1), 15–35.

Ayers, J. W., Caputi, T. L., Nebeker, C., & Dredze, M. (2018). Don't quote me: Reverse identification of research participants in social media studies. *NPJ Digital Medicine*, *1*(1), 30.

Buntain, C., Golbeck, J., Liu, B., & LaFree, G. (2016). Evaluating public response to the Boston Marathon bombing and other acts of terrorism through Twitter. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, Cologne, Germany (pp. 555–558). Palo Alto, CA: AAAI.

Buscaldi, D., & Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, *22*(3), 301–313.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, *17*(1), 124–147.

DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX (pp. 2382–2388). Palo Alto, CA: AAAI.

Derczynski, L., Nichols, E., van Erp, M., & Limsopatham, N. (2017). Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the Third Workshop on Noisy User-Generated Text*, Copenhagen, Denmark (pp. 140–147). Stroudsburg, PA: ACL.

Freire, N., Borbinha, J., Calado, P., & Martins, B. (2011). A metadata geoparsing system for place name recognition and resolution in metadata records. In *Proceedings of the 11th International ACM/IEEE Joint Conference on Digital Libraries*, Ottawa, ON, Canada (pp. 339–348). New York, NY: ACM.

Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *GeoInformatica*, *17*(4), 635–667.

Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from microtext. *Transactions in GIS*, *15*(6), 753–773.

Gritta, M., Pilehvar, M. T., & Collier, N. (2018a). A pragmatic guide to geoparsing evaluation. Preprint, arXiv:1810.12368.

Gritta, M., Pilehvar, M. T., & Collier, N. (2018b). Which Melbourne? Augmenting geocoding with maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia (Vol. 1: Long Papers, pp. 1285–1296). Stroudsburg, PA: ACL.

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing?*Language Resources & Evaluation*, *52*(2), 603–623.

Gu, Y., Qian, Z. S., & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, *67*, 321–342.

Hecht, B., & Moxley, E. (2009). Terabytes of Tobler: Evaluating the first law in a massive, domain-neutral representation of world knowledge. In K. S. Hornsby, C. Claramunt, M. Denis, & G. Ligozat (Eds.), *Spatial information theory* (Lecture notes in computer science, Vol. *5756*, pp. 88–105). Berlin, Germany: Springer.

Hu, Y., Mao, H., & McKenzie, G. (2019). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, *33*(4), 714–738.

Huang, Q., & Xiao, Y. (2015). Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, *4*(3), 1549–1568.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, *47*(4), 67.

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia (pp. 1638–1643). Stroudsburg, PA: ACL.

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., & Ghazi, D. (2017). Location detection and disambiguation from Twitter messages. *Journal of Intelligent Information Systems*, *49*(2), 237–253.

Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, *22*(3), 219–228.

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., & McKenzie, G. (2016). Things and strings: Improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge engineering and knowledge management: EKAW 2016* (Lecture notes in computer science, Vol. *10024*, pp. 353–367). Cham, Switzerland: Springer.

Karimzadeh, M. (2016). Performance evaluation measures for toponym resolution. In *Proceedings of the 10th Workshop on Geographic Information Retrieval*, San Francisco, CA. New York, NY: ACM.

Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J. O., Hardisty, F., Pezanowski, S., … MacEachren, A. M. (2013). GeoTxt: A web API to leverage place references in text. In *Proceedings of the Seventh Workshop on Geographic Information Retrieval*, Orlando, FL (pp. 72–73). New York, NY: ACM.

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., & Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, *23* (1), 118–136.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, CA (pp. 260–270). Stroudsburg, PA: ACL.

Leidner, J. L. (2008). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Irvine, CA: Universal Publishers.

Li, H., Wang, M., Baldwin, T., Tomko, M., & Vasardani, M. (2019). UniMelb at SemEval-2019 Task 12: Multi-model combination for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, MN (pp. 1313–1318). Stroudsburg, PA: ACL.

Lieberman, M. D., Samet, H., & Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 26th IEEE International Conference on Data Engineering*, Long Beach, CA (pp. 201–212). Piscataway, NJ: IEEE.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., … Blanford, J. (2011). SensePlace2: GeoTwitter analytics support for situational awareness. In *Proceedings of the 2011 IEEE Conference on Visual Analytics Science and Technology*, Providence, RI (pp. 181–190). Piscataway, NJ: IEEE.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD (pp. 55–60). Stroudsburg, PA: ACL.

Meier, P. (2010). The unprecedented role of SMS in disaster response: Learning from Haiti. *SAIS Review of International Affairs*, *30*(2), 91–103.

Murthy, D., & Longwell, S. A. (2013). Twitter and disasters: The uses of Twitter during the 2010 Pakistan floods. *Information, Communication & Society*, *16*(6), 837–855.

Overell, S., & Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, *22*(3), 265–287.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar (pp. 1532–1543). Stroudsburg, PA: ACL.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, New Orleans, LA (Vol. 1: Long Papers, pp. 2227–2237). Stroudsburg, PA: ACL.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations & Trends in Information Retrieval*, *12*(2–3), 164–318.

Seetharaman, D., & Wells, G. (2017). *Hurricane Harvey victims turn to social media for assistance*. Retrieved from https://www.wsj.com/articles/hurricane-harvey-victims-turn-to-social-media-for-assistance-1503999001

Silverman, L. (2017). *Facebook, Twitter replace 911 calls for stranded in Houston*. Retrieved from https://www.npr.org/sections/alltechconsidered/2017/08/28/546831780/texas-police-and-residents-turn-to-social-media-to-communicate-amid-harvey

Speriosu, M., & Baldridge, J. (2013). Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria (pp. 1466–1476). Stroudsburg, PA: ACL.

Starbird, K., & Stamberger, J. (2010). Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. In *Proceedings of the 7th International ISCRAM Conference* (Vol. 1, pp. 1–5). Seattle, WA: ISCRAM.

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, *32*(1), 1–29.

Wang, J., & Hu, Y. (2019a). Are we there yet? Evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on Geospatial Humanities*, Chicago, IL (pp. 1–6). New York, NY: ACM.

Wang, J., & Hu, Y. (2019b). Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, *23*(6), 1393–1419.

Wang, R.-Q., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences*, *111*, 139–147.

Wang, X., Ma, C., Zheng, H., Liu, C., Xie, P., Li, L., & Si, L. (2019). DM NLP at SemEval 2018 Task 12: A pipeline system for toponym resolution. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, MN (pp. 917–923). Stroudsburg, PA: ACL.

Weissenbacher, D., Magge, A., O'Connor, K., Scotch, M., & Gonzalez, G. (2019). SemEval-2019 Task 12: Toponym resolution in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, MN (pp. 907–916). Stroudsburg, PA: ACL.

Yadav, V., Laparra, E., Wang, T.-T., Surdeanu, M., & Bethard, S. (2019). University of Arizona at SemEval-2019 Task 12: Deep-affix named entity recognition of geolocation entities. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, MN (pp. 1319–1323). Stroudsburg, PA: ACL.

Yu, M., Huang, Q., Qin, H., Scheele, C., & Yang, C. (2019). Deep learning for real-time social media text classification for situation awareness—Using hurricanes Sandy, Harvey, and Irma as case studies. *International Journal of Digital Earth*, *12*(11), 1230–1247.

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

How to cite this article: Wang J, Hu Y, Joseph K. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS*. 2020;24:719–735. https://doi.org/10.1111/tgis.12627