Location reference recognition from texts: A survey and comparison

XUKE HU*, Institute of Data Science, German Aerospace Center (DLR), Germany

ZHIYONG ZHOU†, Department of Geography, University of Zurich, Switzerland

HAO LI, Department of Aerospace and Geodesy, Technical University of Munich, Germany

YINGJIE HU, Department of Geography, University at Buffalo, USA

FUQIANG GU, College of Computer Science, Chongqing University, China

JENS KERSTEN, Institute of Data Science, German Aerospace Center (DLR), Germany

HONGCHAO FAN, Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Norway

FRIEDERIKE KLAN, Institute of Data Science, German Aerospace Center (DLR), Germany

A vast amount of location information exists in unstructured texts, such as social media posts, news stories, scientific articles, web pages, travel blogs, and historical archives. Geoparsing refers to recognizing location references from texts and identifying their geospatial representations. While geoparsing can benefit many domains, a summary of the specific applications is still missing. Further, there lacks a comprehensive review and comparison of existing approaches for location reference recognition, which is the first and core step of geoparsing. To fill these research gaps, this review first summarizes seven typical application domains of geoparsing: geographic information retrieval, disaster management, disease surveillance, traffic management, spatial humanities, tourism management, and crime management. We then review existing approaches for location reference recognition by categorizing these approaches into four groups based on their underlying functional principle: rule-based, gazetteer matching-based, statistical learning-based, and hybrid approaches. Next, we thoroughly evaluate the correctness and computational efficiency of the 27 most widely used approaches for location reference recognition based on 26 public datasets with different types of texts (e.g., social media posts and news stories) containing 39,736 location references worldwide. Results from this thorough evaluation can help inform future methodological developments and can help guide the selection of proper approaches based on application needs.

CCS Concepts: • **Computing methodologies** → **Information extraction**; • **Information systems** → **Retrieval effectiveness**; **Geographic information systems**.

Additional Key Words and Phrases: geoparsing, location reference recognition, machine learning, comparative review

*Corresponding author

†Corresponding author

Authors' addresses: Xuke Hu, xuke.hu@dlr.de, Institute of Data Science, German Aerospace Center (DLR), Jena, Germany; Zhiyong Zhou, zhiyong.zhou@geo.uzh.ch, Department of Geography, University of Zurich, Zurich, Switzerland; Hao Li, hao_bgd.li@tum.de, Department of Aerospace and Geodesy, Technical University of Munich, Germany; Yingjie Hu, yhu42@buffalo.edu, Department of Geography, University at Buffalo, Buffalo, USA; Fuqiang Gu, gufq@cqu.edu.cn, College of Computer Science, Chongqing University, Chongqing, China; Jens Kersten, Jens.Kersten@dlr.de, Institute of Data Science, German Aerospace Center (DLR), Jena, Germany; Hongchao Fan, hongchao.fan@ntnu.no, Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Trondheim, Norway; Friederike Klan, Friederike.Klan@dlr.de, Institute of Data Science, German Aerospace Center (DLR), Jena, Germany.

1 INTRODUCTION

Location matters, and not just for real estate [179]. With the rapid development of the Global Navigation Satellite System (GNSS), sensor-rich (e.g., inertial sensors, Wi-Fi module, and cameras) smart devices, and ubiquitous communication infrastructure (e.g., cellular and 4G networks and Wi-Fi access points), our capability of obtaining location information of moving objects and events in both indoor and outdoor spaces has been dramatically improved [160]. This exponential growth in location-based capabilities has significantly enhanced our understanding of geospatial processes [179] and fueled the development of location-based services (LBS) with wide-ranging applications in various domains, such as business, entertainment, and crisis management [87]. Apart from sensor equipment, natural language texts, such as social media posts, web pages, and news stories, serve as a significant source of geospatial information through location references. These location references encompass both simple place names, also known as toponyms, as well as more complex location descriptions that incorporate additional spatial modifiers like direction, distance, and spatial relationships [177]. Geoparsing, an ongoing research problem studied extensively over the past two decades [9, 15, 82, 92, 163], refers to the process of extracting location information from texts. It involves two crucial steps: (1) recognizing location references from texts, also known as toponym recognition or location reference recognition, and (2) identifying the geospatial representations of the recognized location references, commonly referred to as toponym resolution or geocoding. Figure 1 illustrates the workflow of geoparsing.
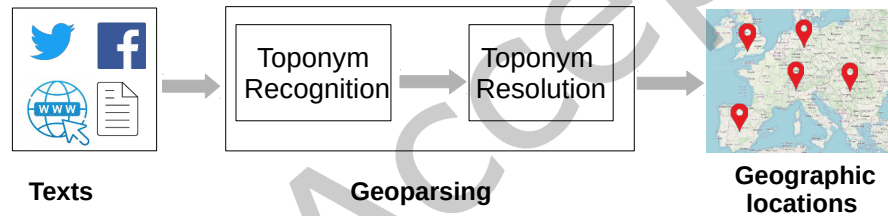

Fig. 1. The general workflow of geoparsing and its two steps.

Geoparsing has traditionally been used in formal texts for location extraction, such as web pages, news, scientific articles, travel blogs, and historical archives [15, 179]. However, the drastically increased importance of social media data (SMD) in various domains such as social science, political science, policy-making, and humanitarian relief [18, 22, 38, 76, 171] has facilitated efforts to extend geoparsing to informal texts [179]. According to Statista [1], the number of worldwide social network users will reach 4.4 billion by 2025. On average, 500 million tweets [2] and 4.75 billion Facebook items [3] are shared each day. Formal texts normally do not have location-related metadata, while informal texts, such as tweets, can be geotagged, i.e., a Twitter user can select a location and attach that location to the posted message. However, geotagged tweets are rare, and according to Cheng et al. [33], Morstatter et al. [135], and Kumar et al. [101], only 0.42%, 3.17%, and 7.90% of the total number of tweets contain geotags, respectively. In addition, Twitter removed the precise geotagging feature in 2019, showing only a rough location, e.g., the bounding box of a tagged place, rather than a pair of latitude and longitude coordinates. This

---

[1]https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2]https://www.dsayce.com/social-media/tweets-day/

[3]https://blog.wishpond.com/post/115675435109/40-up-to-date-facebook-facts-and-stats

change could further decrease the number of geotagged tweets [86]. In a nutshell, extracting location information from unstructured texts is often necessary. Notably, informal texts, such as tweets, are short, have few or no formatting or grammatical requirements, and can have uncommon abbreviations, slang, and misspellings, which pose additional challenges for geoparsing [180].

While there exist quite some studies on geoparsing [68, 146], we identify two gaps in the literature that motivate this current review paper. First, the many possible applications of geoparsing are scattered in individual papers [1, 15, 58, 63] or are only partially reviewed [67, 83], and there lacks a systematic and more comprehensive summary of these applications. Consequently, it is difficult for researchers who are new to geoparsing to have a quick view of these many possible applications. Second, existing review papers on geoparsing, such as [68, 125, 134, 181], focused on the entire workflow of geoparsing (i.e., both of the two steps) rather than location reference recognition alone (i.e., the first step only). While providing more comprehensive coverage on the topic of geoparsing, existing efforts reviewed only some approaches for the step of location reference recognition. In recent years, many new approaches for location reference recognition have been developed, such as Flair NER [4], NeuroTPR [182], nLORE [53], and GazPNE2 [82]. Given the high importance of location reference recognition in geoparsing (i.e., only those references that are correctly recognized can be geo-located), it is necessary to have a review that specifically focuses on the possible and recent approaches for location reference recognition.

This work aims at filling the two research gaps discussed above. First, we summarize seven typical application domains of geoparsing, which are geographical information retrieval (GIR) [57, 146], disaster management [111, 162], disease surveillance [64, 159, 172], traffic management [77, 115, 129], spatial humanities [63, 154], tourism management [27, 36, 37], and crime management [17, 42, 178]. Second, we review existing approaches for location reference recognition by categorizing the approaches into four groups: rule-based, gazetteer matching-based, statistical learning-based, and hybrid approaches. Noticing that many existing approaches were not cross-compared on the same datasets, we also conduct experiments to compare and evaluate the reviewed 27 existing approaches on 26 public datasets. We thoroughly analyze various aspects of the existing approaches, encompassing their performance on both formal and informal texts, their effectiveness across different types of places such as administrative units and traffic ways, and their computational efficiency.

The remainder of this paper is structured as follows: In Section 2, we summarize seven typical application domains of geoparsing. In Section 3, we review existing approaches for location reference recognition. We evaluate existing approaches on the same public datasets in Section 4. Finally, we conclude the paper in Section 5 and discuss some potential future directions.

2 SEVEN APPLICATION DOMAINS OF GEOPARSING

Geoparsing offers numerous potential applications. In this section, we provide a concise overview of seven prominent application domains frequently explored in the literature. Figure 2 provides an illustration of these domains.

GIR: One of the primary applications of geoparsing is geographic information retrieval. Historically, documents have been indexed by subject, author, title, and type. However, a diverse and large group of information system users (e.g., readers, natural resources managers, scientists, historians, journalists, and tourists) desire geographically-oriented access to document collections, such as by retrieving interesting contents about specific geographic locations [26, 57, 108, 130, 145, 173, 191]. For instance, resources in digital libraries can be indexed by locations contained in descriptive metadata records associated with the resources, thereby improving users' experience in searching for their needed resources [57]. People are looking for web pages containing useful information about everyday tasks, such as local merchants, services, and news [26]. The public can consume up-to-date information related to COVID-19 (e.g., disease prevention, disease transmission, and death reports) on Twitter by locations [130].
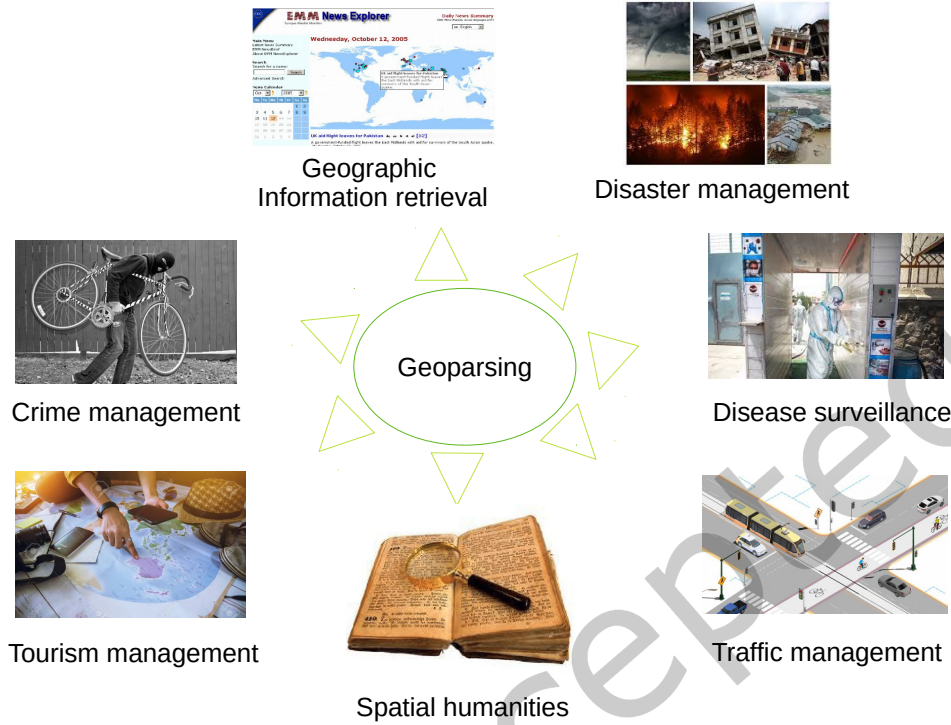
Fig. 2. Seven application domains of geoparsing.

Disaster management: News stories and SMD contain enormous historical and real-time disaster information. Location-enabled SMD can be very helpful to timely map the situational information, such as rescue requests [164, 198], resource (e.g., food, clothing, water, medical treatment, and shelter) needs and availability [21, 50], and facility status (e.g., building collapse, road closure, pipe broken, and power outage) [23, 52, 121, 157] in the aftermath of disasters. With a crisis map, first responders can track the unfolding situation and identify stricken locations that require prioritized intervention [19] and realize optimized real-time resource allocation [164], government agencies can conduct the damage assessment of the disasters in a faster manner [192], and the public can search for the locations where they can obtain needed resources. By extracting spatiotemporal, environmental, and other information about disaster events from news stories, flood-prone areas can be identified [194], the responsibility of atmospheric phenomena for floods can be understood [20], the spatial and temporal distributions of natural disasters during a long period can be analyzed [114], and the evolution of disasters (e.g., the phases of preparedness, impact, response, and recovery) can be tracked [88, 183, 184].

Disease surveillance: Scientific articles, historical archives, news reports, and social media contain detailed information about disease events, such as where the disease was first reported and how it spread spatiotemporally. Mining geographic locations and other related information of disease events can help track diseases [34, 64, 136, 140, 159, 172], perform early warning and quick response [97], and understand the mechanisms underlying the emergence of diseases [12, 93]. For example, geoparsing historical archives (e.g., The annual US Patent Office Reports 1840-1850 and Registrar General's Reports) can help track the spread of potato disease 'late blight' in the 19th-century in the United States [172] and understand the relationship between cholera related disease

and place names during Victorian times [136]. Scientific articles were geoparsed to analyze the demographic, environmental, and biological correlation of the occurrence of emerging infectious diseases at a global scale [12, 93]. Social media can also reflect the movement of the public and their feelings during pandemics through geotags or mentioned locations in texts. Location-enabled tweets were applied to analyze the mental health status of the public after the occurrence of COVID-19 [80, 197], to track and visualize the spread and diffusion of COVID-19 [16], and to reveal human mobility patterns [89, 91].

Traffic management: Twitter users report near-real-time information about traffic events (e.g., crashes and congestion). Detecting traffic events, their precise locations, and other related information from tweets is important for an effective transportation management system [3, 13, 61, 71, 161, 168]. The detected traffic events can also support urban policy-making [40], such as helping drivers to avoid risk zones and choose the fastest and safest routes [10], to help the transportation management sector reduce fatalities and restore traffic flow as quickly as possible [10], to predict future traffic jams [11], and to improve road safety by recognizing high-risk areas [129]. By doing so, Twitter users acting as social sensors can complement existing physical transport infrastructure (e.g., video cameras and loop detectors) cost-effectively, which is especially important for developing countries where resources are limited.

Spatial humanities: 'Spatial turn' was used to describe a general movement observed since the end of the 1990s, emphasizing the reinsertion of place and space in the humanities [185]. Digitizing and geoparsing large historical textual collections, such as books, reports, and novels, create new ways for research in the humanities (e.g., Archaeology, History, and Literature) [49, 62, 63, 69, 78, 131, 136, 172], such as to understand the historical geographies of nineteenth-century Britain and its relationships with the wider world [62], to identify the significance of specific commodities in relation to particular places and time [78], to analyze a correspondence between eighteenth-century aesthetic theory and the use of the terms 'beautiful', 'picturesque', 'sublime', and 'majestic' in contemporaneous and later accounts of the Lakes region [49], and to reveal the spatial structure of a narrative in fictional novels [131].

Tourism management: According to Statista, among all the active blogs, travel is rated as the top 5 topics shared by bloggers [4]. Travel blogs contain a wealth of information about visited places organized as bloggers' experiences and insights as well as their perceptions of these places [75]. These narratives reflect the blogger's behavior and interaction with places and also the relationships among the places. Geoparsing travel blogs is helpful for understanding places [74], such as to find their features and related activities, and can help describe a place with tourism attributes to support tour planning [74, 75, 99, 196]. Applications include helping travelers choose preferred places and visit them in an appropriate order at a proper time and supporting wayfinding given the spatial relation of places [75].

Crime management: Many countries do not make crime data available to their citizens [17] or provide only coarse-grained details [5], such as the total number of thefts in a district or a province. According to the Crime Information Need Survey [17], around 78.3% of respondents in Indonesia agreed that crime information should be available to the public. The needed information includes crime type, perpetrator, victim, time, and, very importantly, location. Meanwhile, crime-related information is often scattered across news and social media. Mining and gathering crime-related information from these text-based sources can be useful for informing the public and may even help predict and prevent some crimes [14, 41, 42, 150, 156, 166]. In particular, geoparsing can help extract location information of crimes, which can help residents to choose places to live and help travelers to avoid certain unsafe places [17].

Different applications have distinct requirements for the approaches for location reference recognition. For example, emergency response applications primarily rely on analyzing informal texts like tweets, while scientific

---

[4]https://www.statista.com/statistics/187267/number-of-bloggers-in-usa/
[5]http://liiofindia.org/

articles serve as the main source for understanding the mechanisms underlying disease emergence. GIR just needs coarse-grained geospatial information, such as a city, while traffic management requires the fine-grained location (e.g., a street) of traffic events; Geoparsing historical documents that contain billions of words requires a fast processing workflow. Therefore, to guide the selection of proper approaches for location reference recognition based on application needs, examining the characteristics of existing approaches is necessary, which will be introduced in Section 4.

3 A SURVEY OF EXISTING APPROACHES

In this section, we review existing approaches for location reference recognition. In subsection 3.1, we review individual approaches by categorizing them into four groups, and in subsection 3.2, we review existing comparative studies and differentiate our current review from the existing studies.

3.1 Approaches for location reference recognition

In the existing literature, Leidner and Lieberman [105], Monteiro et al. [134], Purves et al. [146] identified three types of approaches for location reference recognition, which are rule-based, gazetteer matching-based, and statistical learning-based. However, many studies, such as [58, 81, 106, 124], used a combination of different approaches to compensate for the shortcomings of each other. Therefore, in this review, we add a fourth type, *hybrid approaches*, which combines two or all three types of approaches, and we use these four types to organize our review on location reference recognition. We show this classification schema in Figure 3.
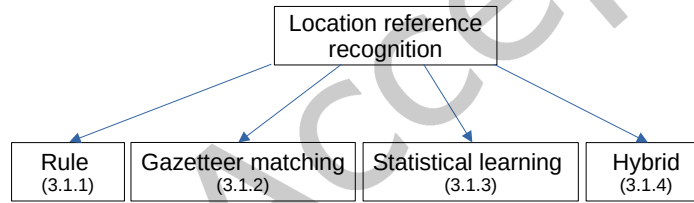

Fig. 3. Classification of existing approaches for location reference recognition.

3.1.1 Rule-based approaches. Location references in texts often have certain lexical, syntactic, and orthographic features. A set of rules, such as regular expressions (REs) and context-free grammars (CFGs), can be defined to decide if an n-gram of texts is a location reference or not [105]. n-grams are linear sequences of n words in texts. For example, given a text $T = \{w_0\ w_1\ w_2...w_n\}$, its unigrams or n-grams of size $n = 1$ include $\{w_0\},\{w_1\},\{w_2\}...\{w_n\}$. Its bigrams or n-grams of size $n = 2$ include $\{w_0\ w_1\},\{w_1\ w_2\}...\{w_{n-1}\ w_n\}$.

Table 1 lists some RE and grammar rules used in previous studies [60, 61, 105, 124]. Each row in the table indicates a rule. The former twelve rules are REs, using Part-of-speech (POS) tags and/or keywords. We use the standard meta characters (i.e., '?', '+', and '∗') of REs. The '?' sign indicates the presence of a tag zero or one time, the '+' sign indicates the presence of a tag at least one time, and the '∗' sign indicates the presence of a tag any times (zero or more). Numbers indicate different types of words. 1 represents street indicators, such as *'street'*, *'highway'*, *'road'*, *'sh'*, and *'beltway'*. 2 represents words that specify direction or a distance in measurable terms, such as *'10'*, *'away'*, *'from'*, *'miles'*, *'km'*, *'south'*, and *'northbound'*. 3 represents place category words, such as *'city'*, *'str'*, *'avenue'*, *'rd'*, and *'village'*. Used POS tags include Nouns (NN), Proper Nouns (NNP), Determiners (DT), Adjectives (JJ), Cardinal Numbers (CD), and Conjunctions (CC). The last two rules are grammar rules. X denotes candidate n-grams. 4 represents place category words that are often used with *'of'*, such as *'city'*, *'town'*,

'gulf', and *'river'*. 5 represents spatial prepositions that normally appear before a location, such as *'in'*, *'around'*, *'on'*, *'near'*, and *'between'*.

Table 1. Examples of RE and grammar rules for location reference recognition.

Rules	Examples
<NN>+	tiburon blvd; san manteo
<NNP>+	Heidelberg; San Francisco
<DT>?<JJ>?<NN>+	the golden gate bridge; long island
<CD>?<DT>?<JJ>?<NN>+	third street; 11th avenue
<DT>?<JJ>?<NN>+<CD>?	freeway 91; highway 12
<DT>?<JJ><NN>(1)	the high cotton lane; high star drive
(1)<CD>	beltway 10; sh 73
(2)+<NNP>+(3)*	south Northumbria bridge road; northeast Munich
<NNP>+(3)*(2)+	Camanche Avenue east; Heidelberg North
(2)+(3)*(of)?<NNP>+	25 miles sw of San Francisco; 25min away from New York State
(3)*(of)?<NNP>+(2)+	town of San Francisco; district of Columbia
<A-Z><a-z>*berg	Heidelberg; Freiberg
(4) (of) X -><LOC>	city of beaumont; Gulf of Mexico;
(5) X -><LOC>	this overturned tanker in **marin** has created a huge jam on **wb 580**

Several studies used only rules to extract location references. For instance, Giridhar et al. [61] used road-traffic-related tweets to detect and locate point events, such as car accidents. Specifically, a set of REs were defined according to the composition of nouns, determiners, adjectives, cardinal numbers, conjunctions, and possessive endings. Furthermore, to decrease false positives, grammar-based rules were implemented based on spatial prepositions, such as *'in'*, *'at'*, *'between'*, and *'near'*. Zou et al. [198] analyzed the rescue request on Twitter during Hurricane Harvey. They assumed that the formal description of an address in the United States is in the form of [Street Number, Street Name, Apartment Number (optional), City, State, Zip Code]. Since all rescue request tweets in their study contain zip codes, the full address in each tweet can be extracted by locating the zip code as the ending point and searching for the starting point based on several conditional criteria.

Although many studies classify rule-based approaches as one category [6, 105, 134], pure rules-based approaches are rare. All the rule-based approaches discussed in [134] are, in fact, hybrid approaches. This is likely because the approaches that rely on linguistic patterns only are ineffective [163]. Defining comprehensive and resilient rules to account for all potential instances of location references in texts, particularly in microblogs characterized by diverse writing styles and loose grammar [152], remains a challenging task. Nevertheless, rules can significantly bolster gazetteer matching and statistical learning-based approaches, as elucidated in the subsequent sections.

3.1.2 Gazetteer matching-based approaches. A gazetteer is a dictionary of place names associated with geospatial information (e.g., place types and geographic coordinates) and some additional information such as population size, administrative level, and alternative names. Gazetteers play important roles in location reference recognition in many studies. GeoNames [6] is a most widely used gazetteer, and OpenStreetMap (OSM) [7], in a broad sense, can be considered as a gazetteer as well. There are 12,255,028 [8] and 23,876,956 [9] places in GeoNames and OSM, respectively. Figure 4 illustrates the point density map of the places in OSM and GeoNames.

In gazetteer matching-based approaches, the n-grams of a text are first matched against a gazetteer, which are then filtered or disambiguated with a couple of heuristics. Gazetteer matching-based approaches are still
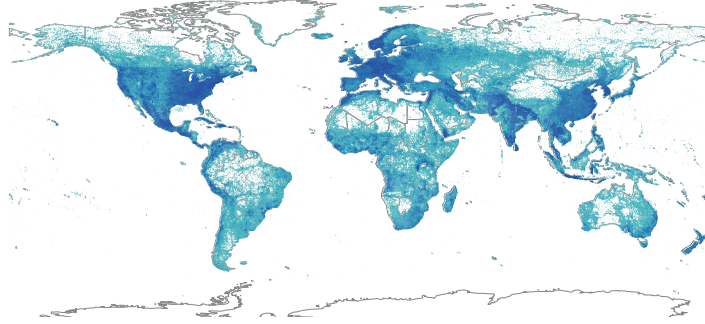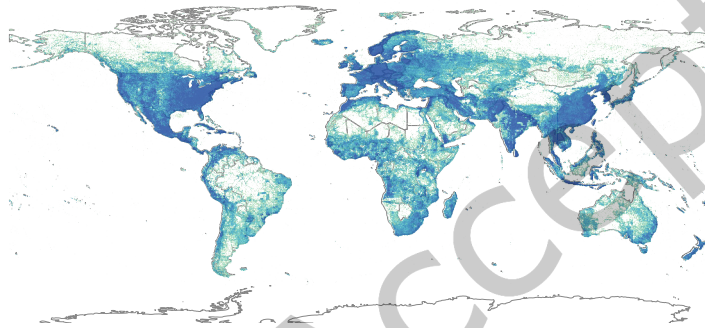
---

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

[7]https://www.openstreetmap.org/

[8]Retrieved from the official web of GeoNames on 2022.02.25

[9]Retrieved from OSMNames on 2022.02.25

OpenStreetMap

GeoNames

Feature per square
(0.1*0.1 degree):
- 0.0001
- 3.9981
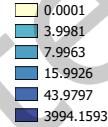- 7.9963
- 15.9926
- 43.9797
- 3994.1593

Fig. 4. Point density maps of the places contained in OpenStreetMap and GeoNames.

faced with two main challenges. The first is that many location references appearing in texts are missing from gazetteers due to various reasons, such as name variation (e.g., *'South rd'* for *'South road'* and *'Frankfurt airport'* for *'Frankfurt international airport'*) and data incompleteness (e.g., the missing of *'Hidden Valley Church of Christ'* from a gazetteer) [58]. Second, gazetteer matching-based approaches often run into ambiguity issues. For instance, the names *'Washington'*, *'MO'*, *'South Wind'*, and *'1 ft'* all exist in gazetteers, but can also refer to other types of entities. This is called geo/non-geo ambiguities, while geo/geo ambiguities refer to the situation in that different spatial locations use the same name, such as Manchester, NH, USA versus Manchester, UK. For simplicity, we use *ambiguities* and *ambiguous* to refer to geo/non-geo ambiguities by default. We will use the full name *geo/geo ambiguities* to refer to the second situation. The main focus of gazetteer-based approaches is often to overcome the two mentioned challenges by using heuristics to perform disambiguation (to increase precision) and by including place name variants to expand the used gazetteer (to increase recall).

Table 2 summarizes the commonly used heuristics for disambiguation. The first four heuristics are used to reduce the number of candidate place names matched in gazetteers, thereby decreasing the number of ambiguous place names. The 5th heuristic uses common words. The 6th and 7th heuristics leverage the external and internal cues of candidate n-grams. The 8th heuristic leverages the POS tags of candidate n-grams. The 9th heuristic leverages the dictionary of other entity types (e.g., *Person*), such as to judge if a candidate n-gram (*'Houston'*)

with its preceding or succeeding word (*'Alexander'*) in texts appears in the dictionary of person names. The 10th heuristic leverages other related place names to judge if an n-gram is valid or not. For example, *'IN'* is ambiguous. However, when it precedes *'Chennai'* in texts (e.g., *'Chennai, IN'*), which is likely to be a location and related to *'IN'*, then *'IN'* is treated as a valid location. If an n-gram can be determined as a valid location by some heuristics, the other n-grams with the same name in the text are also treated as a valid location, such as *'stay safe Houston, flood in Houston is serious'*, where both *'Houston'* are judged as valid locations since the preceding word (*'in'*) of the second *'Houston'* is a spatial indicator.

Table 2. Common heuristics used for disambiguation in gazetteer matching-based approaches

ID	Heuristics	Examples
1	Limit the length of place names in gazetteers	Keep only 1- and 2-grams
2	Limit the type of places in gazetteers	Keep only continent, country, state, and city
3	Limit the scale of places in gazetteers	Keep only places with a population over 1000
4	Limit the spatial range of gazetteers	Use the gazetteers in the area of Florence
5	Filter place names of common (stop) words	'today', 'long', 'that building' and 'the street'
6	Use spatial indicators in texts	'in', 'near', and 'at' that appear before a place
7	Use orthographic cues	Capitalization of words, such as 'Houston' and 'Germany'
8	Filter candidates by POS tags	Keep only noun phrases in texts
9	Use a dictionary of other entity types	Person names, such as 'Washington Irving' and 'Houston Alexander'
10	Use other related place names	'Chennai, IN' and 'stay safe Houston, flood in Houston is serious'

Many studies used gazetteer matching-based approaches to recognize location references from texts [3, 6, 12, 15, 35, 43, 54, 127, 129, 141, 141, 143, 163, 169, 170, 191]. One of the earliest geoparsing approaches was proposed by Woodruff and Plaunt [191] to support georeferenced document indexing and retrieval. A gazetteer containing around 120,000 places in California was first built on the US Geological Survey's Geographic Names Information System (USGS 1985) and the land use data from the US Geological Survey's Geographic Information Retrieval and Analysis System (GIRAS). The place names in a document were identified by matching texts' n-grams containing non-stop words against the gazetteer. If a token had no matches in the gazetteer, it was deplurialized (e.g., *'valleys'* to *'valley'*) and rematched with the gazetteer. Amitay et al. [15] developed Web-a-Where for recognizing and geocoding continents, countries, states, and cities as well as their abbreviations in web pages. A gazetteer was created by collecting about 75,000 place names across the world from different data sources: USGS, World-gazetteer.com [10], UNSD [11], and ISO 3166-1 [12]. The system first extracted candidate place names in a given page by matching against the gazetteer. Then, four heuristics were sequentially used to disambiguate and geocode the candidate place name, such as the vicinity of two candidate places (e.g., *'Chicago, IL'*) and the population of places. Clough [35] proposed identifying candidate place names by matching against gazetteers, which were then filtered using stop words and context cues, such as to filter person names with simple heuristic $< title >< loc >$ (e.g., *'Mr. Sheffield'*), where $< loc >$ is a candidate place name and also in the dictionary of person names. Used gazetteers include the Ordnance Survey 1:50,000 Scale Gazetteer for the UK (OS [13]), Seamless Administrative Boundaries of Europe dataset (SABE [14]), and Getty Thesaurus of Geographic Names (TGN [15]). Pouliquen et al. [143] proposed geoparsing approaches for multilingual texts. Candidate place names were first identified by matching with a multilingual gazetteer, which were then disambiguated through a dictionary of

person names (e.g., *'George Bush'* and *'Tony Blair'*) and stop words (e.g., *'And'*, *'Du'*, *'Auch'*) in a multilingual context. The multilingual gazetteer was created from three sources: Global Discovery database of place names (Global Discovery 2006), the multilingual KNAB database (KNAB 2006), and a European Commission internal document.

Gazetteer matching-based approaches were also used to extract locations from tweets. For instance, Paradesi [141] proposed a Twitter geoparser, TwitterTagger. It matched the noun phrases of a tweet text with the entries in gazetteers (i.e., USGS database), which was followed by disambiguating the matched entry with two heuristics. The first was to check if spatial indicators (e.g., *'in'* and *'near'*) were used before a noun phrase. The second was to check whether other users used a spatial indicator before the same noun phrase in their tweets. Middleton et al. [127] proposed a multilingual geoparser for tweets named Geoparserpy. To overcome the place name variation issues, a set of heuristics were applied to expand OSM place names. To deal with abbreviations, a multilingual corpus of the street and building types from OSM was used to compute obvious variants for common location types (e.g., *'Southampton Uni'* for *'Southampton University'*). To overcome the ambiguity issue, uni-gram location names that are non-nouns were filtered using a multilingual WordNet corpus lookup, such as *'ok'* and *'us'*, which can refer to locations or other types depending on their POS tag. Location phrases were then filtered using a multilingual stop-word corpus. de Bruijn et al. [43] introduced TAGGS, a method that leveraged metadata and contextual spatial information from groups of related tweets. TAGGS matched uni- and bi-grams from tweet texts with GeoNames and then filtered the candidates using various heuristics, such as excluding candidates associated with the 1000 most frequently occurring words.

Studies, such as [3, 6, 23, 129, 168, 193, 194], focused only on local events whose geographical scope is known, such as floods or traffic accidents happened in a certain city. Therefore, they would normally use a local gazetteer that contains only the places in a certain region, which can dramatically mitigate the issues of geo/non-geo ambiguities and geo/geo ambiguities. Although the proposed geoparsing approaches are not globally applicable, they are effective in dealing with local events. For instance, Al-Olimat et al. [6] proposed a Location Name Extraction tool (LNEx), which used n-gram statistics and location-related dictionaries to handle the abbreviations and automatically filter and augment the place names in the OSM gazetteer (handling name contractions and auxiliary contents). Ahmed et al. [3] utilized tweets to monitor real-time traffic congestion. They extracted location references from the tweets by matching n-grams with a list of road names in Chennai. To handle place name variants, they employed the Jaro-Winkler metric to calculate the similarity between the n-grams and the road names in gazetteers. Milusheva et al. [129] used traffic-related tweets to derive the locations of road traffic crashes in Nairobi, Kenya. Specifically, they developed a gazetteer matching-based geoparsing method to identify the location of car crashes. A gazetteer of landmarks (e.g., roads, schools, and bus stops) for five counties that constitute the Nairobi metro area was created from OSM, GeoNames, and Google Places. The location of car crashes was then determined by matching the n-grams of the tweets with the entries in the gazetteer. Gazetteer matching-based approaches are straightforward to implement and can readily adapt to multilingual contexts. They prove particularly effective in specific applications, such as those with a limited geographic scope (e.g., a city) or those that primarily require coarse-grained location information, such as countries. However, proposing a generally applicable approach for location reference recognition using gazetteer matching and simple heuristics remains challenging due to the prevalence of name variants and geo/non-geo ambiguity in natural language texts. To address this challenge, numerous studies have sought to combine gazetteer matching with rules and/or statistical learning methods to overcome the limitations of each approach. These combined approaches will be discussed in the subsequent sections.

3.1.3 Statistical learning-based approaches. Statistical learning-based approaches are built on annotated training corpora containing texts associated with the expected location references. The annotated corpora are used to train a model via manually defined features, such as infrequent strings, length, capitalization, contextual features,

and/or features automatically learned by deep learning methods. The trained model is then applied to unlabeled texts, and the same features are computed to decide on the association of texts and location references. The basic architecture of statistical learning-based approaches is illustrated in Figure 5, which use either traditional machine learning techniques, such as Random Forest (RF) [57], or deep learning techniques, such as Long Short-Term Memory (LSTM) [182]. Statistical learning-based approaches can be further divided into two groups: *learning-based named entity recognition (NER)* and *learning-based place name extraction (PNE)*. In the following, we discuss these two groups of approaches respectively.

bad accident on Waiyaki Way next to Kianda

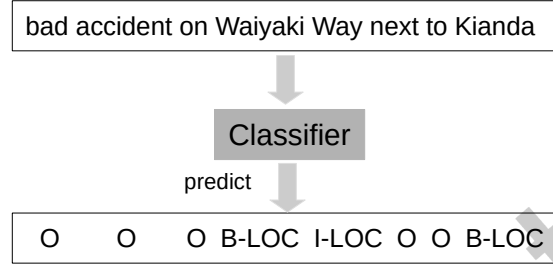Classifier

predict

O   O   O B-LOC I-LOC O O B-LOC

Fig. 5. Basic architecture of statistical learning-based approaches. *O* denotes non-type. *B-LOC* and *I-LOC* denote the beginning and inner part of a location reference, respectively.

Learning-based NER: Location reference recognition can be considered as a subtask of NER, which has been extensively studied. Therefore, many studies [59, 71, 73, 95, 111, 121, 175] used existing statistical learning-based NER models or retrain them to extract location references from texts. For instance, Lingad et al. [111] used OpenNLP [16], TwitterNLP [152], Yahoo!Placemaker, and Stanford NER to extract place names from 2878 disaster-related tweets. Stanford NER and OpenNLP were also retrained and evaluated by using 10-fold cross-validation in their study. The results show that retrained models achieved a much higher F1 score than pretrained models. Karimzadeh et al. [96] proposed a geoparsing system for tweets, named GeoTxt. It integrated six publicly available NERs for location reference recognition, which are Stanford NER, Illinois CogComp [151], GATE ANNIE [25], MIT IE [17], Apache OpenNLP, and LingPipe [18]. Belcastro et al. [23] utilized tweets to discover sub-events after a disaster, such as collapsed buildings, broken gas pipes, and flooded roads. CoreNLP [120] was adopted to recognize street and district names, which were then geocoded by matching with a local gazetteer that covers the disaster area. Fan et al. [52] proposed uncovering the unfolding of disaster events based on tweets. Place names were extracted using Stanford NER, which was then filtered and geocoded by keeping only the matched places in Google Geocoding API and excluding the places outside affected areas. Tateosian et al. [172] used CLAVIN [19] to geoparse two historical collections: US Patent Office Reports 1841-1850 and Google Books Corpus. CLAVIN is an open-sourced geoparser that utilizes Apache OpenNLP for place name extraction. Mircea [130] implemented a prototype dashboard for real-time classification, geolocation, and interactive visualization of COVID-19 tweets. spaCy [20] was used to extract city and country names from tweet texts and user profiles. Mao et al. [121] proposed mapping near-real-time power outages from tweets using a retrained NeuroNER model [46]. Suat-Rojas et al. [168] retrained spaCy NER for the detection of location references from Spanish tweets pertaining to traffic accidents in a Colombian city.

---

[16]https://opennlp.apache.org/
[17]https://github.com/mit-nip/MITIE
[18]http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html
[19]https://github.com/Novetta/CLAVIN
[20]https://spacy.io/

Recently, many deep learning-based NERs have also been proposed. For example, Limsopatham and Collier [110] proposed recognizing name entities from tweets by enabling BiLSTM to automatically learn orthographic features using both character embedding and word embedding. Akbik et al. [5] proposed Flair, an NLP tool that used contextual string embeddings for sequence labeling tasks, such as part-of-speech (POS) tagging and NER. Qi et al. [148] proposed a deep learning-based NLP toolkit, named Stanza, which adopted a contextualized string representation-based tagger. Recently, the fully-connected self-attention architecture (a.k.a. Transformer) attracts a lot of attention due to its parallelism and advantage in modeling long-range contexts. For instance, Ushio and Camacho-Collados [176] presented a Python library for NER model fine-tuning, named T-NER. It facilities the training and testing of a Transformer-based NER model. Nine public NER datasets from different domains are compiled as part of the T-NER library, such as CoNLL 2003, OntoNoted 5.0, and WNUT 2017 datasets.

Learning-based PNE: In addition to utilizing or retraining existing NER models, numerous studies have developed their own models for location reference recognition employing various machine learning [138, 155, 165] and deep learning techniques [9, 24, 30, 32, 100, 117, 123, 176, 193]. For instance, Nissim et al. [138] trained the Curran and Clark (C&C) maximum entropy tagger [39] for recognizing location references from Scottish historical documents, using the built-in C&C features, including morphological and orthographical features, information about the word itself, POS tags, named entity tag history, and contextual features. The model was evaluated on 648 Scottish historical documents containing 10,868 sentences and 5,682 places. Kumar and Singh [100] adopted a multi-channel convolutional neural network (CNN) architecture to extract location references from tweets. The model was evaluated on 5,107 earthquake-related tweets with 6,690 place names using 10-fold cross-validation. Xu et al. [193] proposed DLocRL, a deep-learning pipeline for fine-grained location recognition and linking in tweets. Specifically, they first used BiLSTM-CRF to train a POI recognizer. Then, given an input pair ⟨POI, Profile⟩, a linking module was trained to judge whether the location profile corresponds to the POI. The profile is an entry in a POI dictionary. The approach was evaluated on the Singaporean national Twitter dataset that was first used in [106], containing 3,611 tweets and 1,542 POIs. Cadorel et al. [30] proposed to extract a property's location and neighborhood from French housing advertisements by recognizing place names and retrieving relationships between them. Specifically, a BiLSTM-CRF network with a concatenation of several text representations, including CamenBERT [122], Flair, and Word2Vec [128] was used to extract place names.

To mitigate the effort of manually annotating a large training dataset, semi-supervised approaches have been developed. For instance, Wang et al. [182] proposed generating training data from Wikipedia articles, which was then used to train a BiLSTM model called NeuroTPR. Their model contains several layers to account for the linguistic irregularities in Twitter texts, such as using character embeddings to capture the morphological features of words, and contextual embeddings to capture the semantics of tokens in tweets. The approach was evaluated on 1,000 tweets related to the 2017 Hurricane in Texas and Louisiana. Qiu et al. [149] proposed ChineseTR, a weakly supervised Chinese toponym recognizer. It first generated training examples based on word collections and associated word frequencies from various texts. Based on the training examples, a BiLSTM-CRF network built on the BERT word embedding was explored to train a toponym recognizer. The approach was evaluated on three Chinese NLP datasets, WeiboNER, Boson, and MSRA [21]. Khanal and Caragea [98] used a multi-task learning setting to augment the learning of fine-grained location identification. The three tasks related to crisis events are key-phrase identification, eyewitness-account classification, and humanitarian category classification. The learning was conducted on one of the three popular transformer-based models: BERT [47], Albert [104], and RoBERTa [116]. Several public datasets for the training of the three tasks were utilized in multi-task learning. The proposed approach was evaluated on two disaster-related Twitter datasets that were used in Middleton et al. [127].

---

[21]https://github.com/InsaneLife/ChineseNLPCorpus

Given abundant annotated data, statistical learning-based approaches can automatically recognize location references according to the contextual cues and the intrinsic features of location references without requiring additional expert knowledge and gazetteers. However, a large number of labeled training sentences are often not available, making it difficult to use these approaches in many situations [70]. Furthermore, deep learning-based models normally take much more time to recognize place names from texts than rule and gazetteer matching-based approaches.

3.1.4 Hybrid approaches. Every single technique has its own drawbacks. Thus, researchers have proposed fusing different techniques to achieve the best of all [25, 51, 79, 106, 119, 188, 195]. Hybrid approaches can be further divided into four types based on the way they combine the previous three approaches: fusing rule and gazetteer, fusing rule and statistical learning, fusing gazetteer and statistical learning, and fusing rule, gazetteer, and statistical learning.

Fusing rule and gazetteer: Many studies [119, 123, 132, 133, 144, 184, 188] fused rules and gazetteers to overcome the shortcomings of each other. Manually defined rules are fragile and the detected location references can be thus further verified by gazetteers. Inversely, rules can help remove the ambiguities of the location references detected by gazetteer matching and by recognizing those references that are not included in gazetteers. For instance, Pouliquen et al. [144] proposed identifying cities and countries from newspapers in multiple languages. Location references were recognized by matching texts' n-grams written in upper case with a multi-language gazetteer, named Global Discovery gazetteer. The matches were then filtered by stop words and person names. To recognize the morphological variants of places, regular expressions were used to list all possible suffixes and suffix combinations of location references. By doing so, some unseen places in gazetteers can be recognized, such as *'Lontoolaisen'*, because it consists of *'Lontoo'* that is in the gazetteer and the suffix *'laisen'*. The approach was evaluated on 28 texts with 1,650 places in 8 languages, such as English, Spanish, and Russian. Weissenbacher et al. [188] presented a geoparsing system for scientific articles related to phylogeography. GeoNames was first searched to detect location references in articles, and then a black-list (e.g., *'How'*, *'Although'*, *'Gene'*, and *'Body'*) and a set of rules were created to remove noisy entities found in GeoNames. Malmasi and Dras [119] first used a POS rule-based tree-splitting method to extract noun phrases from tweets, and then matched the n-grams of the noun phrases with the entries of GeoNames. Dutt et al. [51] presented SAVITR, a system that geo-visualizes tweets during emergencies. They used a POS tagger to find proper nouns and then used REs to mitigate the ambiguity of proper nouns with the prefix and suffix words (e.g., *'road'*, *'south'*, and *'city'*) of place names. Last, the phrases extracted by the above methods are verified and geocoded using a gazetteer (i.e., GeoNames or OSM) in India. Martínez and Pascual [123] presented LORE, a knowledge-based model that captures location references from English and Spanish tweets. First, bi-grams and uni-grams in the tweets were matched with entries in the GeoNames gazetteer and then filtered by heuristics. Second, linguistic patterns involving location-indicative words (e.g., *'city'* and *'street'*), location markers (e.g., *'north'* and *'10km'*), and POS tags were derived to recognize location expressions, such as *'25 miles NW of London City'*. They derived the linguistic patterns from 500 English tweets and 100 Spanish tweets and then used 900 English tweets and 500 Spanish tweets to test LORE.

Fusing rule and statistical learning: Statistical learning models might not generalize well due to limited training samples, and manually defined rules can be added to boost the performance of the trained models, e.g., by correcting evident errors. For instance, Acheson and Purves [2] introduced a geoparsing approach for scientific articles in PDF format. They initially employed Stanford NER to recognize potential location references and subsequently applied rules to filter these candidates, such as to include candidates with terms like *'University'* or *'Institute'* while excluding candidates with terms like *'Inc'* and *'GmbH'*. Google Geocoding API was then used to determine the spatial representation of the detected location references. The approach was evaluated on two article corpora in the domain of Orchards and Cancer, containing 150 and 200 articles, respectively. Das and Purves [40] proposed detecting traffic events (e.g., traffic accidents and congestion) in India using tweets. Specifically,

they combined the detected location references by Stanford NER, retrained OpenNLP, and a rule-based system involving spatial indicators (e.g., *'in'*, *'at'*, and *'near'*), POS tags, and 85 words representing place categories (e.g., *'hospital'*, *'road'*, and *'clinic'*).

Fusing gazetteer and statistical learning: Gazetteers are utilized in two primary ways: (1) combining the detection outcomes of statistical learning models with gazetteer matching, and (2) incorporating the gazetteer matching results (e.g., presence or absence of an n-gram in the gazetteers) as input features for statistical learning models. Examples of the first way are [57, 72, 79, 106]. For instance, Freire et al. [57] proposed geoparsing descriptive metadata records associated with digital resources. Initial location references were recognized by matching tokens of records with candidate entries in GeoNames. A Random Forest classifier was then trained to disambiguate and link the initial location references. Li and Sun [106] proposed recognizing POIs in tweets. Candidate POIs in tweets were first extracted by matching with a POI inventory, which was constructed from check-in data in Foursquare. A trained time-aware POI tagger based on CRF was then utilized to remove the ambiguity of the candidates based on the context cues in the text. Hoang and Mothe [79] combined the detection results of multiple publicly available approaches, such as Ritter's tool [152], Gate NLP framework [25], and Stanford NER, and then filtered the results using DBPedia. Different configurations of the NER approaches and DBPedia were tested on the Ritter's dataset [152] and MSM2013 dataset [31]. Examples of the second way include [53, 90, 118, 142, 187]. For instance, Inkpen et al. [90] trained three CRF models for recognizing city, province/state, and country mentions based on manually defined features, including gazetteer features. The models were intended to detect location references in tweets and categorize them into three types. The models were evaluated using 10-fold cross-validation on 6,000 tweets with 1,270 country mentions, 772 state/province mentions, and 2,327 city mentions. Weissenbacher et al. [187] introduced a novel approach for recognizing location references within research articles. The method used a CRF model, incorporating various features, such as lexical (i.e., POS tags), semantic, and gazetteer features. Fernández-Martínez and Periñán Pascual [53] proposed nLORE, a BiLSTM-CRF architecture for location reference recognition, exploiting linguistic and gazetteer features from LORE [124]. The model was trained on 7,000 tweets and tested on 1,063 tweets.

Fusing rule, gazetteer, and statistical learning: Some studies combined all the three techniques for location reference recognition [50, 60, 81, 82, 107, 118]. For instance, Gelernter and Zhang [60] proposed a cross-lingual location reference recognizer, combining the results of a named location parser based on gazetteer matching, a rule-based building parser, a rule-based street parser, and a trained CRF-based named entity parser. The rules of the street and building parsers were created based on POS tags and indicator words, such as adjective plus noun and street and building indicators (e.g., *'street'* and *'highway'* in English and *'calle'* and *'carreterra'* in Spanish). They used a dataset of 4,488 Spanish crisis-related tweets to evaluate the approach. Of these, 3,182 tweets were used for training, and the remaining tweets served as the test set. Additionally, to evaluate the English extractor, the Spanish dataset was translated into English using Google Translate.

Magge et al. [118] employed a deep feedforward neural network to determine whether a given phrase in biomedical articles represents a toponym. They utilized rules to generate approximately 8 million training samples from unannotated datasets. These generated samples, along with manually annotated training samples, were used to train the deep learning model. The input vector for the model was constructed by concatenating various features, including the context of the phrase represented by word embeddings, properties of the phrase (e.g., presence in GeoNames), and properties of the document (e.g., abstract, introduction, body, or table). Dutt et al. [50] focused on understanding crucial aspects of need-tweets and availability-tweets during disasters. They aimed to extract information about the required resources (e.g., water, food, shelter, medicines), the quantity of the resources needed or available, the geographical location of the need or availability, and the individuals or organizations involved in providing or needing them. Regarding geoparsing, the authors enhanced their previously proposed system, Savitr [51], by combining the location references detected by spaCy and a rule-based system. They further filtered the location references using a gazetteer to improve the accuracy of location

extraction. More recently, a place name extractor named GazPNE was proposed by Hu et al. [81]. GazPNE utilized a neural classifier trained on place names from OSM in the US and India, along with synthesized non-place names generated by rules. However, due to its limited use of context information, GazPNE still faced ambiguity issues. To address this, a more robust approach called GazPNE2 was developed [82]. GazPNE2 utilized two pretrained transformer models, BERT and BERTweet [137], to disambiguate the detected location references.

3.2 Comparative studies

In addition to individual studies that focused on developing new methods, researchers also conducted experiments to compare existing methods based on the same datasets. Liu et al. [113] created a medium-scale corpus of locative expressions from multiple social media sources which include the TellUsWhere corpus [189], two sets of micro-blog posts from Twitter, comments from YouTube, forums, blog posts from tier one of the ICWSM-2011 Spinn3r dataset [22], Wikipedia, and documents from the British National Corpus [29]. They then compared the performance of a couple of location reference recognition models over these seven corpora, which include Locative Expression Recogniser (LER) [112], retrained Stanford NER, pretrained Stanford NER, GeoLocator [58], UnLockText, and Twitter NLP. Gritta et al. [68] evaluated the performance of five geoparsers (GeoTxt, Edinburgh Geoparser [69], Yahoo! PlaceSpotter, CLAVIN, and Topocluster [44]) on two datasets, Local-Global Lexicon (LGL) [109], and WikToR that was programmatically created by the author. For location reference recognition, GeoTxt used Stanford NER, Edinburgh Geoparser used LT-TTT2, TopoCluster used Stanford NER, and CLAVIN used Apache OpenNLP. The evaluation results showed that Stanford NER performed the best in location reference recognition, and Edinburgh Geoparser and CLAVIN performed the best in geocoding. Wang and Hu [181] developed an extensible and unified platform for evaluating geoparsers, named EUPEG, which enabled direct comparison of nine geoparsers on eight public corpora, which are LGL, GeoVirus [65], TR-News [95], GeoWebNews [66], WikToR [68], GeoCorpora [179], Hu2014 [84], and Ju2016 [94]. The compared geoparsers include GeoTxt, Edinburgh Geoparser, TopoCluster, CLAVIN, Yahoo! PlaceSpotter, CamCoder [65] that used spaCy NER for location reference recognition, DBpedia Spotlight [126], and two systems that used Stanford NER and spaCy NER for location reference recognition, respectively. Won et al. [190] evaluated the performance of five NERs and voting systems that combined the NERs in extracting place names from two historical correspondence collections, Mary Hamilton Papers and the Samuel Hartlib collection. The NERs include NER-Tagger [102], Stanford NER, spaCy, Edinburgh Geoparser, and Polyglot-NER [7]. The results showed that although the individual performance of each NER system was corpus dependent, the ensemble combination can achieve consistent measures of precision and recall, outperforming the individual NER systems. At the International Workshop on Semantic Evaluation 2019 [23], a task for toponym resolution in scientific articles was launched. The evaluation results were presented in [186]. Several systems were evaluated on a corpus of 150 full PubMed articles as 105 articles for training and 45 articles for testing, containing in total 8,360 toponyms. In the subtask of toponym recognition, all systems except one adopted Deep Recurrent Neural Networks. The system proposed by a team from Alibaba Group achieved the highest F1 score by adopting BiLSTM-CRF and training it on various datasets, including OntoNote5.0, CoNLL13, and weakly labeled training corpora.

There are two major differences between this study and the aforementioned comparative studies. First, these existing comparative studies focused on the entire workflow of geoparsing, while we focus on a narrower topic, i.e., location reference recognition, and provide a deeper review and comparison of methods on this topic. Second, our comparative experiments (presented in the following section) are more comprehensive than existing studies. We used more datasets (26 datasets, containing 39,736 places worldwide) and compared 27 different approaches. In the following, we present the results from the comparative experiments.

---

[22]https://www.icwsm.org/2011/data.php

[23]https://alt.qcri.org/semeval2019/

Table 3. Main features of approaches evaluated in this study

Approach and Version	Category	Recognized Entity Type	Target Text	Development Language	Publishing Time
Stanford NER 4.3.1	statistical learning	4 classes	formal texts	Java	2021
spaCy 3.2.1	statistical learning	18 classes	formal texts	Python	2021
Stanza 1.2	statistical learning	18 classes	formal texts	Python	2021
OpenNLP 1.9.4	statistical learning	4 classes	formal texts	Java	2021
DBpedia Spotlight	statistical learning	N/A	formal texts	Python	2021
NER-Tagger	statistical learning	4 classes	tweets	Python	2016
Polyglot 16.07.04	statistical learning	3 classes	formal texts	Python	2016
NeuroNER	statistical learning	4 classes	formal texts	Python	2017
CogComp 4.0	statistical learning	4 classes	formal texts	Java	2018
OSU TwitterNLP	hybrid	10 classes	tweets	Java	2011
TwitIE-Gate 9.0.1	hybrid	4 classes	tweets	Java	2013
TNER	statistical learning	28 classes	formal texts	Python	2021
Flair NER	statistical learning	4 classes	formal texts	Python	2021
Flair NER (Ont)	statistical learning	18 classes	formal texts	Python	2021
BERT-base-NER	statistical learning	4 classes	formal texts	Python	2020
CLIFF 2.6.1	statistical learning	LOC	formal texts	Python	2020
Edinburgh 1.2	hybrid	LOC	formal texts	c	2021
GazPNE2	hybrid	LOC	tweets	Python	2022
LORE	hybrid	LOC	tweets	c++	2020
nLORE	hybrid	LOC	tweets	c++	2021
SPENS	hybrid	LOC	formal texts	N/A	2018
RSD	hybrid	LOC	tweets	N/A	2018
RGD	hybrid	LOC	tweets	N/A	2018
RS	hybrid	LOC	tweets	N/A	2018
BaseSemEval12	hybrid	LOC	formal texts	Python	2018
NeuroTPR	statistical learning	LOC	tweets	Python	2020
Geoparserpy 2.1.4	gazetteer matching	LOC	tweets	Python	2020

4 COMPARISON OF EXISTING APPROACHES

4.1 Methods

To inform future methodological developments for location reference recognition and help guide the selection of proper approaches based on application needs, we examine numerous characteristics of existing approaches for location reference recognition. We use or implement the 27 most widely used approaches including both general NERs and location-specific approaches. Note that we do not include several approaches, such as LNEx [6], GazPNE [81], and SAVITR [51], as they are limited to local regions and cannot be applied globally, whereas our test datasets consist of place names from around the world.

Table 3 summarizes the features of the compared approaches. The version number is indicated alongside each approach's name. The second column represents the approach's category based on its underlying principle. NERs not only recognize locations but also other entity types, as denoted in the third column. The notation of 3 classes, 4 classes, 10 classes, and 18 classes corresponds to {LOC, PER, ORG}, {LOC, PER, ORG, MISC}, {PERSON, GEO-LOCATION, COMPANY, PRODUCT, FACILITY, TV-SHOW, MOVIE, SPORTSTEAM, BAND, OTHER}, and {LOC, PERSON, ORG, FAC, GPE, CARDINAL, DATE, EVENT, LANGUAGE, LAW, MONEY, NORP, ORDINAL, PERCENT, PRODUCT, QUANTITY, TIME, WORK_OF_ART}, respectively. The 28 classes include entities from the 18 classes mentioned earlier, as well as {CELL TYPE, CELL LINE, CHEMICAL, CORPORATION, DISEASE, DNA, GROUP, PROTEIN, RNA, OTHER}. The fourth column indicates the type of texts on which each approach

was developed, while the fifth column specifies the development language used. The last column denotes the proposal or update date of a particular version of an approach.

- **Stanford NER (4.3.1)** [55] [24]: It is a Java-based NER system that utilizes CRF, which was developed and maintained by the Stanford Natural Language Processing Group. We keep the entities of LOC (location) detected by Stanford NER as locations.
- **spaCy (3.2.1)**: It is a general NLP tool. We use its retrained model (*en_core_web_lg*) and keep the entities of LOC, FAC (facility), and GPE (geopolitical entity) detected by spaCy as locations.
- **Stanza (1.2)** [148] [25]: It is a general NLP toolkit and includes a NER tool, which was built on BiLSTM and CRF. We keep the entities of LOC, FAC, and GPE as locations.
- **OpenNLP (1.9.4)** [126]: The Apache OpenNLP library is an open-sourced and machine learning-based toolkit for processing natural language text. We keep the entities of LOCATION detected by OpenNLP as locations.
- **DBpedia Spotlight** [126] [26]: It is for recognizing and linking entities to DBpedia. We treat the place mentions detected by this approach as locations.
- **NER-Tagger** [102] [27]: It is a NER tool for tweets, built on BiLSTM and CRF. We treat the entities tagged with B-LOC and I-LOC as locations.
- **Polyglot (16.07.04)** [7] [28]: It is a natural language pipeline and includes a multi-language NER tool. The entities tagged with I-LOC are regarded as locations.
- **NeuroNER** [46] [29]: It is a BiLSTM-CRF-based NER system developed by MIT. We use the pretrained model and keep the entities of LOC, FAC, and GPE detected by NeuroNER as locations.
- **CogComp (4.0)** [151] [30]: It is a NER tagger, developed by the University of Illinois. The entities tagged with LOC are taken as locations.
- **OSU Twitter NLP** [152] [31]: It is a NER tool for tweets. The entities tagged with GEO-LOCATION and FACILITY are treated as locations.
- **TwitIE-Gate (9.0.1)** [25] [32]: It is a Twitter-specific NER tool, providing an executable pipeline on an open-source software toolkit GATE [33] (General Architecture for Text Engineering). The entities tagged with LOCATION by this approach are treated as locations.
- **TNER** [176] [34]: It is an All-Round Python Library for Transformer-based NER. We keep the entities of LOC, FAC, and GPE as locations.
- **Flair NER** [4] [35]: Flair is an NLP framework designed to facilitate training and distribution of sequence labeling and text classification. Flair-NER is the standard 4-class NER model trained on CoNLL-03. We keep the entities of LOC as locations.

---

[24] https://nlp.stanford.edu/software/CRF-NER.html

[25] https://stanfordnlp.github.io/stanza/

[26] https://www.dbpedia-spotlight.org/

[27] https://github.com/glample/tagger

[28] https://polyglot.readthedocs.io/en/latest/index.html

[29] https://github.com/Franck-Dernoncourt/NeuroNER

[30] https://github.com/CogComp/cogcomp-nlp/tree/master/ner

[31] https://github.com/aritter/twitter_nlp

[32] https://gate.ac.uk/wiki/twitie.html

[33] https://gate.ac.uk/

[34] https://github.com/asahi417/tner

[35] https://huggingface.co/flair/ner-english

- **Flair NER (Ontonotes)** [158] [36]: This is the large 18-class NER model trained on Ontonotes that ships with Flair. It is named *Flair NER (Ont)* for short in this review. We include entities tagged with LOC, GPE, and FAC as locations.
- **BERT-base-NER** [37]: It is a fine-tuned BERT model that is ready to use for NER. We include entities tagged with B-LOC and I-LOC as locations.
- **GazPNE2** [82] [38]: It fuses global gazetteers and two pretrained transformer models. The latest version utilizes Stanza to enhance GazPNE2.
- **CLIFF (2.6.1)** [48] [39]: It integrates the results of Stanford NER and a modified CLAVIN (Cartographic Location and Vicinity Indexer) geoparser.
- **LORE** [124]: It is a rule-based location extractor for tweets.
- **nLORE** [123]: It is a deep learning model, an advanced version of **LORE**. We use the trained model provided by the author to extract locations.
- **Edinburgh Geoparser (1.2)** [69] [40]: It is a geoparsing approach developed by Edinburgh University, which combines rules and gazetteers to extract place names from texts.
- **BaseSemEval12** [118] [41]: It is a baseline system for SemEval-2019 Task 12 (i.e., Toponym Resolution in Scientific Papers) that uses a 2-layer feedforward neural network.
- **NeuroTPR** [182] [42]: It is a neuro-net toponym recognition approach trained on recurrent neural networks. We use their trained model and implementation to detect location mentions in texts.
- **Geoparserpy (2.1.4)** [127]: It is a gazetteer matching-based geoparser. We use its implementation and deploy the required OSM gazetteer to extract place names from texts.
- **SPENS** [190]: This approach combines the result of five different systems in a voting mechanism, including Stanford NER, Polyglot NER, Edinburgh Geoparser, NER-Tagger, and spaCy. It is thus named SPENS for short. We reimplement the approach using the code or API of the five modules.
- **Ritter+Stanford NER+DBpedia** [79]: It uses DBpedia to filter the merged detection by Ritter's tool (also named OSU Twitter NLP) and Stanford NER. We name this approach RSD for short and reimplement the approach using the code or API of the three modules.
- **Ritter+GATE+DBpedia** [79]: It uses DBpedia to filter the merged detection by Ritter's tool and GATE. We name this approach RGD for short and reimplement the approach using the code or API of the three modules.
- **Ritter+Stanford NER** [79]: It merges detection by Ritter's tool and Stanford NER. We name this approach RS for short and reimplement the approach using the code or API of the two modules.

All methods are configured based on thorough experimental results to ensure the selection of optimal parameter settings. For example, we consider not only *Location* and *GPE* but also *Facility* detected by Stanza as locations since this can achieve the best F1 score on the entire datasets.

4.2 Test data

We collect 26 commonly used datasets, which serve as our test data. The datasets comprise 3 formal datasets (i.e., news) and 23 informal datasets (i.e., tweets), containing 39,736 place names in total, as shown in Table 4. They can be categorized into two groups based on the purpose of the datasets: Location Extraction (LE) and

---

[36]https://huggingface.co/flair/ner-english-ontonotes-large
[37]https://huggingface.co/dslim/bert-base-NER
[38]https://github.com/uhuohuy/GazPNE2
[39]https://cliff.mediacloud.org/
[40]http://www.ltg.ed.ac.uk/software/geoparser/
[41]https://github.com/amagge/semeval-ffnn-baseline
[42]https://github.com/geoai-lab/NeuroTPR

NER. The former only annotates *Location* while the latter annotates not only *Location*, but also the other types, such as *Person*, *Organization*, and *Facility*. Note that we do not use some available geoparsing datasets that were used to evaluate the geoparsing approaches by Wang and Hu [181], such as WikToR [68] due to their limited coverage of toponyms. For instance, in the WikToR dataset, each text or article corresponds to a Wikipedia page titled with a specific toponym, along with specified coordinates. However, only that particular toponym was automatically annotated, while other toponyms mentioned in the text were disregarded. Although suitable for evaluating toponym resolution approaches, this dataset does not adequately address toponym recognition approaches. The description of the used datasets is as follows:

- **LaFlood2016, HouFlood2015, CheFlood2015** [43]: They are three flood-related datasets, which were created by Al-Olimat et al. [6]. The locations in the three datasets were annotated as one of the three types: *inLOC*, *outLOC*, and *ambLOC*, denoting the locations inside the area (e.g., *'Houston'*) of interest, outside the area, and ambiguous locations (e.g., *'my house'*), respectively. We only evaluate the approaches on the *inLOC* and *outLOC* locations, ignoring the *ambLOC* locations. *'Louisiana'*, *'Houston'*, *'Texas'*, and *'Chennai'*, as well as their abbreviations, such as *'La'*, *'Hou'*, and *'Tx'* appear frequently in the datasets. Moreover, many location mentions appear in hashtags, such as *'#laflood'*, *'#txwx'*, and *'#ChennaiRain'*.

- **Harvey2017** [44]: The dataset is related to 2017 Hurricane Harvey and was created by Wang et al. [182]. The dataset contains many fine-grained locations, such as *'398 Garden Oaks Blvd'* and *'26206 longenbaugh rd'*. No places appear in hashtags since they have been removed from the dataset.

- **NzEq2013, NyHurcn2012** [45]: The two Twitter datasets correspond to the New Zealand earthquake in 2013 and New York Hurricane in 2012, respectively. They were created by Middleton et al. [127]. We found several missing place names (e.g., *'Christchurch'*) which, however, appear frequently in the two datasets. To mitigate this issue, we manually create two missing place name lists (i.e. [(*'new'*,*'zealand'*), (*'nz'*), (*'uk'*), (*'christchurch'*), (*'chch'*), (*'lyttleton'*), (*'southland'*), (*'wellington'*), (*'south', 'island'*)] and [(*'new'*,*'york'*), (*'nyc'*), (*'new'*,*'york'*,*'city'*), (*'ny'*)]) for the two datasets, respectively. We define that the detection of an entity which is not annotated in the dataset but in the corresponding missing list is a true positive. Moreover, sub-place names exist in dataset NzEq2013. For example, in the text *'Christchurch hospital is now back in operation'*, both *'Christchurch hospital'* and *'Christchurch'* were annotated as *Location*. To tackle this issue, we remove sub-place names from the dataset.

- **Martinez_I, Martinez_II, Martinez_III:** The three Twitter datasets correspond to multiple crises and emergency events (e.g., earthquakes, floods, car accidents, bombings, shootings, terrorists, and incidents) that happened across the world. They were initially utilized in Fernández-Martínez and Periñán Pascual [53], Martínez and Periñán-Pascual [124]. One of the features of the datasets is that many fine-grained locations, such as *'13219 S penrose Ave'* and *'Exit 34'* as well as complex location expressions, such as *'50 miles SW of Liverpool'* and *'25mins away from Northumbria Street'* were annotated.

- **GeoCorpora** [46]: It was created by Wallgrün et al. [179]. In the dataset, location references in tweets were not only annotated but also linked to GeoNames. The dataset corresponds to multiple worldwide events (e.g., earthquake, ebola, fire, flood, protest, and rebel) in 2014 and 2015. Most annotated places are admin units, such as continent, country, state, and city.

- **CrisisBench-1000, HumAID-1000, COVID19-1000** [47]: These three datasets, developed by [82], consist of 1000 randomly selected tweets from CrisisBench [8], HumAID [56], and a COVID19 dataset [103],

---

[43]The datasets can be obtained by filling out the Dataset Registration form https://docs.google.com/forms/d/e/1FAIpQLScf6-DNwkgJXPS5e28Mj18hIW3Ap_Ym7Kna-SO7oSmiC72qGw/viewform

[44]https://github.com/geoai-lab/NeuroTPR/tree/master/Data/TestData/HarveyTweet2017

[45]https://revealproject.eu/geoparse-benchmark-open-dataset/

[46]https://github.com/geovista/GeoCorpora

[47]https://github.com/uhuohuy/GazPNE2/tree/main/data/test_data

respectively. For each dataset, place names were manually annotated, encompassing various types such as admin units (e.g., countries and villages), traffic ways (e.g., streets and highways), natural features (e.g., hills and rivers), and POIs (e.g., parks and schools).

- **BTC-A, BTC-B, BTC-E, BTC-F, BTC-G, BTC-H** [48]: The Board Twitter Corpus (BTC) was created by Derczynski et al. [45]. The datasets were sampled across different regions, temporal periods, and types of Twitter users. Apart from *Location*, *Organization* and *Person* were also annotated. Several annotated place names are in mentions (e.g., *'@HoustonFlood'*). However, they are usually ignored by existing location extractors. Thus, we remove the place name in the mentions from the six datasets.
- **NEEL2016** [49]: It is the gold dataset of 2016 Named Entity rEcognition and Linking (NEEL) Challenge. The dataset includes tweets covering multiple noteworthy events from 2011 to 2013, such as the death of Amy Winehouse, the London Riots, the Oslo bombing, and the Westgate Shopping Mall shootout. Entities of different types, such as *Location*, *Person*, *Organization*, *Event*, and *Product* were annotated and linked to DBPedia. We use its training set, which contains 2,135 tweets and 602 places.
- **Ritte's dataset** [50]: It was initially used by Ritter et al. [152]. *Location*, *Facility*, *Person*, and *Organization* were annotated in the dataset. We use its training set, which contains 2,394 tweets and 276 places.
- **MSM2013** [51]: It is the gold dataset of Concept Extraction Challenge held at the Making Sense of Microposts Workshop in 2013 (#MSM2013). Entities of *Person*, *Organization*, *Location*, and *MISC* were annotated. We use its training set, which contains 2,815 tweets and 619 places.
- **WNUT2016** [52]: It is the gold data of the shared task on named entity recognition in Twitter. The task is part of the 2nd Workshop on Noisy User-generated Text (W-NUT 2016). Ten types of entities were annotated, such as *Location*, *Facility*, *Person*, and *Movie*. We use its training set, which contains 3,850 tweets and 791 places.
- **LGL** [53]: Local-Global Lexicon (LGL) corpus was created by Lieberman et al. [109]. Toponyms were manually annotated and geocoded from 588 human-annotated news articles published by 78 local newspapers.
- **GeoVirus** [54]: The GeoVirus dataset, introduced by Gritta et al. [65], serves as an evaluation resource for geoparsing methodologies within the context of news articles pertaining to disease outbreaks and epidemics, such as Ebola, Bird Flu, and Swine Flu. Only admit units were annotated in the dataset. Buildings, POIs, streets, and rivers were disregarded.
- **TR-News** [55]: TR-News was created by Kamalloo and Rafiei [95]. Toponyms were manually annotated and geocoded from 118 news articles from various news sources.

We adopted the standard comparison metrics: precision, recall, and F1-score. In the case of overlapping or partial matches, we penalize an approach by adding 1/2 FP (False Positive) and 1/2 FN (False Negative) (e.g., if the approach marks *'The Houston'* instead of *'Houston'*), following Al-Olimat et al. [6].

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

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

---

[48]https://github.com/GateNLP/broad_twitter_corpus
[49]http://microposts2016.seas.upenn.edu/challenge.html
[50]https://github.com/aritter/twitter_nlp/blob/master/data/annotated/ner.txt
[51]https://www.researchgate.net/profile/Andrea-Varga-4/publication/256682215_MSM2013_Concept_Extraction_Challenge_dataset
[52]https://metatext.io/datasets/wnut-2016
[53]https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/lgl.xml
[54]https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/GeoVirus.xml
[55]https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/TR-News.xml

Table 4. Summary of 26 datasets. There are in total 39,736 places.

Name	Source	Type	Tweet (Article) Count	Place Count	Resolved	Description
LaFlood2016 [6]	tweet	LE	1,500	2,295	No	Louisiana flood in 2016
HouFlood2015 [6]	tweet	LE	1,500	3,060	No	Houston flood in 2015
CheFlood2015 [6]	tweet	LE	1,500	3,671	No	Chennai flood in 2015
Harvey2017 [182]	tweet	LE	1,000	2,107	No	2017 Hurricane Harvey in Texas and Louisiana
NzEq2013 [127]	tweet	LE	1,994	1,252	No	New York hurricane in 2012
NyHurcn2012 [127]	tweet	LE	1,997	764	No	New Zealand earthquake in 2013
Martinez_I [124]	tweet	LE	800	539	No	Multiple emergency events across the world
Martinez_II [124]	tweet	LE	1,371	642	No	Multiple emergency events across the world
Martinez_III [53]	tweet	LE	8,063	5,122	No	Multiple emergency events across the world
CrisisBench-1000 [82]	tweet	LE	1,000	861	No	1,000 tweets from CrisisBench [8]
HumAID-1000 [82]	tweet	LE	1,000	1,422	No	1,000 tweets from HumAid [56]
COVID19-1000 [82]	tweet	LE	1,000	1,245	No	1,000 tweets from COVID-19 [103]
GeoCorpora [179]	tweet	LE	6,634	3,083	Yes	Multiple events across the world
BTC-A [45]	tweet	NER	2,000	229	No	Section A of Broad Twitter Corpus
BTC-B [45]	tweet	NER	200	148	No	Section B of Broad Twitter Corpus
BTC-E [45]	tweet	NER	2,000	572	No	Section E of Broad Twitter Corpus
BTC-F [45]	tweet	NER	2,113	1,330	No	Section F of Broad Twitter Corpus
BTC-G [45]	tweet	NER	1,999	287	No	Section G of Broad Twitter Corpus
BTC-H [45]	tweet	NER	1,000	119	No	Section H of Broad Twitter Corpus
NEEL2016 [153]	tweet	NER	2,135	602	Yes	Dataset of Named Entity rEcognition and Linking Challenge in 2016
Ritte's dataset [152]	tweet	NER	2,394	276	No	A general-purpose NER dataset
MSM2013 [31]	tweet	NER	2,815	619	No	Dataset of Concept Extraction Challenge at the Making Sense of Microposts Workshop in 2013
WNUT2016 [167]	tweet	NER	3,850	791	No	Dataset of shared task on NER in Twitter at the Workshop on Noisy User-generated Text in 2016
LGL [109]	news	LE	588	5,057	Yes	Local-Global Lexicon corpus
GeoVirus [65]	news	LE	229	2,167	Yes	WikiNews related to global disease and epidemics
TR-News [95]	news	LE	118	1,300	Yes	Annotated news articles from various news sources

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

4.3 Results of location reference recognition

We execute the 27 approaches on the complete set of test datasets, and their overall precision, recall, and F1 scores are presented in Figure 6. To obtain these scores, we compute the sums of FP, FN, and TP across all datasets, instead of calculating precision, recall, and F1 scores individually for each dataset and subsequently averaging them across the entire dataset. This approach was adopted due to the imbalanced distribution of place name counts (ranging from 119 to 5122) across different datasets. Detailed results of these approaches on each dataset can be accessed from the provided online file [56].

Most of the employed approaches demonstrate high precision, with 21 out of the total 27 achieving a precision exceeding 0.7. In contrast, most of these approaches exhibit significantly low recall, with only two approaches, GazPNE2 and LORE, achieving a recall surpassing 0.7. This indicates that most approaches missed a considerable

---

[56]https://docs.google.com/spreadsheets/d/16cyuyDhty04hQE1gBfP4zq23Lr3OWJ5EY_bsVMoYxfQ/edit#gid=1536784545

number of location references. To further investigate the underlying reasons for this, we conduct a detailed analysis in Section 4.4, where we examine the detection accuracy of the approaches across various types of texts and location references. Furthermore, it is noteworthy that the top five best-performing approaches, namely GazPNE2, Flair NER (Ont), nLORE, Flair NER, and Stanza, are all based on deep learning and have been introduced within the last four years. This observation underscores the superior performance and notable progress achieved by deep learning in addressing this task. Moreover, two voting-based systems, namely SPENS and RS, achieve remarkable results by combining the detection outcomes of multiple classic approaches, thereby enhancing the performance of each individual approach. These two voting systems demonstrate comparable performance to Stanza and Flair NER, suggesting a great potential for leveraging voting mechanisms in location reference recognition. Such potential holds significant promise, considering the availability of numerous place name extractors.
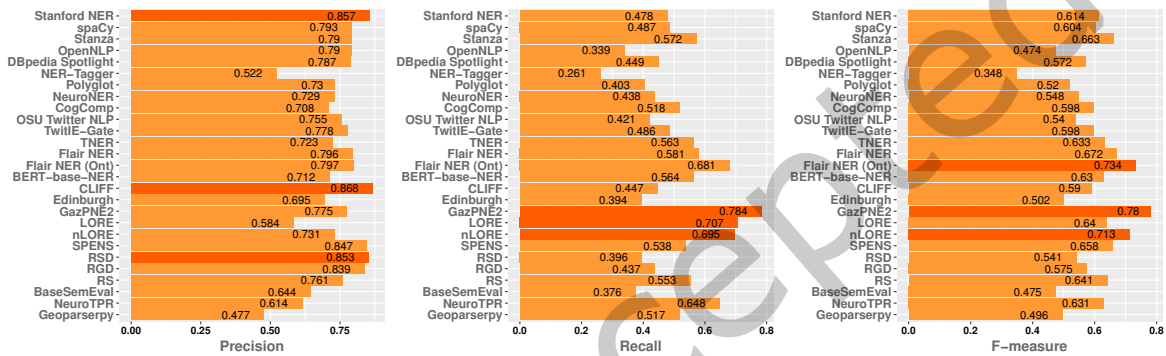

Fig. 6. Overall precision, recall, and F1 score of the 27 approaches on the entire set of datasets with 39,736 location references.

4.4 Error analysis

We conducted an error analysis to gain insight into these approaches' mistakes. The analysis focused on their performance regarding formal and informal texts and detecting location references across various categories and forms.

4.4.1 Text type. Among the test datasets, three comprise formal texts, while twenty-three consist of informal texts, encompassing 8,517 and 31,219 places, respectively. Figure 7 and Figure 8 depict the performance of the approaches on the formal and informal text datasets, respectively. Notably, Flair NER (Ont), Flair NER, SPENS, Stanford NER, and Stanza exhibit superior performance on formal texts, whereas GazPNE2 and nLORE outperform others on informal texts. This discrepancy arises primarily because the former five approaches were trained using formal texts, while the latter two approaches were specifically designed to handle tweets. The dissimilarity in place definitions across these datasets also contributes to the observed variations. Formal datasets predominantly focus on coarse-grained toponyms such as countries and cities, including adjectival forms (e.g., *'American'*, *'British'*, *'Chinese'*) but disregard fine-grained places like highways, POIs, streets, and buildings. For instance, the LGL dataset contains 337 adjectival toponyms, the GeoVirus dataset has 14, and the TR-News dataset contains 124. Conversely, informal datasets such as *Martinez_III* and *HouFlood2015* encompass fine-grained places and places mentioned in hashtags while omitting adjectival toponyms. Furthermore, different approaches adopt distinct place definitions. For instance, some NER approaches, like Stanford NER, disregard fine-grained places, whereas others, like GazPNE2 and nLORE, include fine-grained places. Owing to these disparities in place definitions, accurately

evaluating the performance of these approaches on formal and informal texts poses challenges. Therefore, we conduct more comprehensive evaluations to assess their performance with regard to different types and forms of places in the subsequent sections.
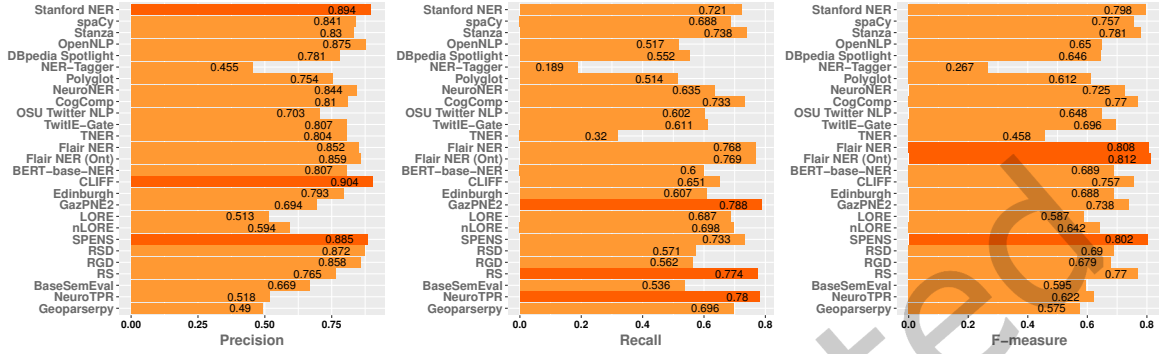

Fig. 7. Precision, recall, and F1 score of the tested approaches on the datasets of formal texts, containing 8,517 places.
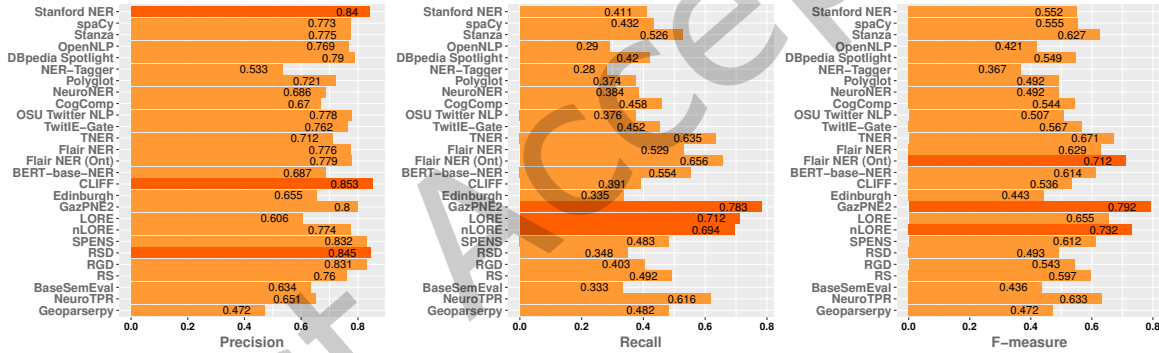

Fig. 8. Precision, recall, and F1 score of the tested approaches on the datasets of informal texts, containing 31,219 places.

4.4.2 Place category. The location references within the datasets are classified into four distinct categories: admin units (e.g., country, state, town, and suburb), traffic ways (e.g., street, road, highway, and bridge), natural features (e.g., river, creek, beach, and hill), and POIs (e.g., park, church, school, and library). For this particular experiment, we select four datasets, namely Harvey2017, GeoCorpora, LGL, and TR-News, as they provide information regarding the category of the places mentioned. In the Harvey2017 dataset, the places are categorized into ten different types [85]. To ensure consistency in our classification, we assign the types of house number addresses, street names, highways, exits of highways, and intersections of roads as traffic ways, the type of natural features as natural features, the types of other human-made features and local organizations as POIs, and the types of admin units and multiple areas as admin units. In the remaining three datasets, place names are linked to the entries in GeoNames. Our classification approach is based on the feature codes provided by GeoNames[57]. Places

---

with feature codes **A** (e.g., country, state, and region) and **P** (e.g., city and village) are classified as admin units. Places with feature codes **R** (e.g., road and railroad) are categorized as traffic ways. Natural features include places with feature codes **H** (e.g., stream and lake), **T** (e.g., mountain, hill, and rock), **U** (e.g., undersea and valley), and **V** (e.g., forest and grove). POIs consist of places with feature codes **L** (e.g., park and port) and **S** (e.g., sport, building, and farm). Across the four datasets, there are 9,790 admin units, 773 traffic ways, 263 natural features, and 754 POIs.

The detection rate, defined as the proportion of correctly detected places out of the total places within a specific category, serves as a measure of performance. In this study, we consider only exact matches as correct detections. Figure 9 presents the detection rates of the approaches in the four categories. We can observe that many approaches show superior performance in recognizing coarse-grained places, with 13 of 27 achieving a detection rate of over 60% for admin units. However, most approaches struggle to recognize fine-grained places. Only 2 of 27 successfully identify over 60% of traffic ways, while 6 of 27 and 4 of 27 achieve the same for natural features and POIs, respectively. These three categories represent geographical scopes that are significantly more precise than admin units and hold great value in various critical applications, such as emergency rescue and traffic event detection. It is worth mentioning that GazPNE2 stands out by achieving a recognition rate of over 70% in all four categories.

Fig. 9. Detection rate of approaches on the four categories with 9,790 admin units, 773 traffic ways, 263 natural features, and 754 POIs.

4.4.3 Form of location references. We consider three forms of location references: the place names with numbers (e.g., *'500 Neches Ave'* and *'Highway 25'*), abbreviation of place names (e.g., *'us'* and *'tx'*), and place names in hashtags (e.g., *'#**Houston**Flood'* and *'#**Chennai**'*). Place names with numbers typically refer to fine-grained locations, such as highways, roads, and home addresses. For the abbreviation of place names, we define it as a

single word with a character length of no more than 3. There are a total of 1,621 place names in the number form, 3,697 place names in the abbreviation form, and 6,560 place names in the hashtag form. The detection rate of the approaches for each form of place names is presented in Figure 10. We can observe that recognizing place names with numbers poses a challenge, as only 4 of the 27 approaches achieve a detection rate of over 0.3. On the other hand, recognizing abbreviations is comparatively more straightforward, with more than half (16 of the approaches) achieving a recognition rate of over 30% for abbreviations. However, it is worth noting that none of the approaches achieve a detection rate exceeding 0.6 for either of these two forms. Recognizing place names in hashtags also presents a challenge, with only 5 approaches achieving a detection rate of over 0.3. Nonetheless, it is encouraging to see that GazPNE2 and LORE can recognize over 70% of the place names in hashtags.
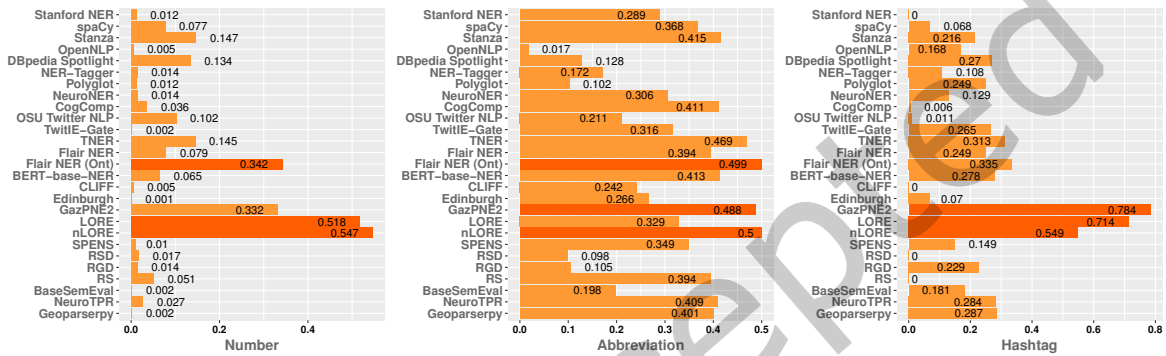

Fig. 10. Detection rate of approaches on location references in three forms: place names with numbers (1,621), abbreviations (3,697), and place names in hashtags (6,560).

4.5 Computational efficiency

In this section, we delve into the computational efficiency, specifically the speed, of different approaches. Many applications involve processing large volumes of text, such as major historical books and reports (e.g., the Old Bailey Online) consisting of millions or even billions of words [63], as well as millions of crisis-related tweets [147]. Therefore, a rapid geoparsing procedure is crucial, making speed a critical factor. To assess the computational efficiency, we execute each approach on the entire dataset and measure the time consumed by each approach. We exclude Edinburgh Geoparser and DBpedia Spotlight from the comparison since they are online services, and it is not possible to measure their processing time directly on the server. Most of the approaches are executed on a MacBook Pro laptop equipped with an Intel Core i7 (2.2 GHz 6-Core) processor and 16 GB of RAM. However, three approaches, namely OSU TwitterNLP, LORE, and nLORE, are executed on a Lenovo laptop with an Intel Core i5 (2.5 GHz 4-Core) processor and 3.8 GB of RAM, as they require a Linux or Windows environment. Figure 11 provides an overview of the speed of the approaches.

We can observe a significant variation in the processing speed among the different approaches. The time required for these approaches to process the entire datasets, which contain 1,092,093 words, ranges from 6 minutes to 33 hours. Interestingly, OSU Twitter NLP exhibits an unexpectedly long processing time of nearly 9.6 hours. As a result, the approaches RSD, RGD, and RS, which rely on OSU Twitter NLP, also take approximately 10 hours. The other approaches that take over 5 hours are all deep learning-based. On the other hand, approaches such as spaCy, Cliff, LORE, Polyglot, and OpenNLP are approximately 20 times faster compared to these deep learning-based approaches. However, it is worth noting that the deep learning-based approaches achieve significantly higher
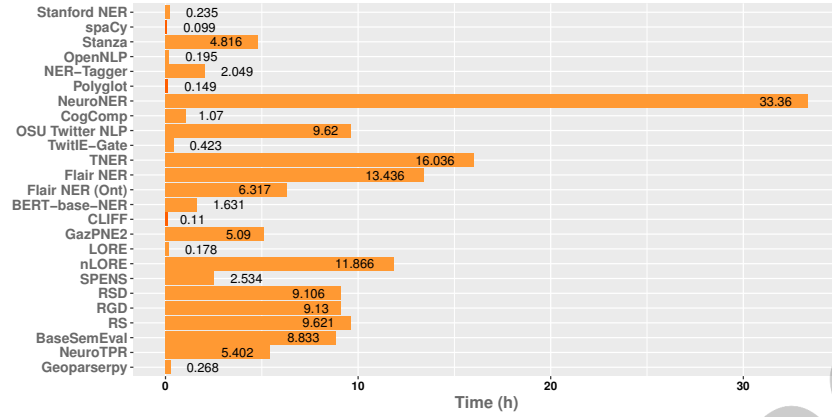
Fig. 11. Time consumption of the approaches running on the entire datasets.

Table 5. Summary of best-performing approaches in various contexts.

	Top 1	Top 2	Top 3	Top 4	Top 5
Informal	GazPNE2	nLORE	Flair NER (Ont)	TNER	LORE
Formal	Flair NER (Ont)	Flair NER	SPENS	Stanford NER	Stanza
Admin Unit (R)	GazPNE2	Flair NER (Ont)	NeuroTPR	Flair NER	Geoparserpy
Traffic Way (R)	GazPNE2	Flair NER (Ont)	BERT-base-NER	nLORE	NeuroTPR
Natural Feature (R)	GazPNE2	NeuroTPR	Flair NER	RS	LORE
POI (R)	GazPNE2	NeuroTPR	LORE	nLORE	Flair NER (Ont)
Number (R)	nLORE	LORE	Flair NER (Ont)	GazPNE2	Stanza
Abbreviation (R)	nLORE	Flair NER (Ont)	GazPNE2	TNER	Stanza
Hashtag (R)	GazPNE2	LORE	nLORE	Flair NER (Ont)	TNER

F1 scores than the other approaches. Therefore, there exists a trade-off between correctness and computational efficiency.

4.6 Summary

This section summarizes the top-performing approaches in various contexts, including formal and informal texts and different types and forms of place names, presented in Table 5. Additionally, we analyze the weaknesses of these approaches, which are documented in Table 6. An approach is deemed weak (indicated by a cross mark) within a given context if it ranks within the lower 50%. In the tables, *R* denotes the recall or detection rate of the approaches concerning specific place categories and forms.

5 CONCLUSIONS AND OUTLOOK

This paper summarizes seven typical applications of geoparsing and surveys existing approaches for location reference recognition, categorizing them as rule-based, gazetteer matching-based, statistical learning-based, and hybrid approaches. We then thoroughly evaluate 27 approaches across 26 datasets, considering overall accuracy, performance on formal and informal texts and various place categories and forms, and computational efficiency. From the results, we can conclude that: (1) deep learning is so far the most promising technique in location reference recognition; (2) the integration of existing approaches through a voting mechanism surpasses individual

Table 6. Summary of best-performing approaches' weakness (indicated by cross marks). 'Abb' denotes 'Abbreviation'.

	Speed	Formal	Informal	Admin Unit (R)	Traffic Way (R)	Natural Feature (R)	POI (R)	Number (R)	Hashtag (R)	Abb (R)
GazPNE2	✗									
Flair NER (Ont)	✗									
nLORE	✗		✗			✗				
TNER	✗		✗	✗						
BERT-base-NER				✗						
LORE			✗							
Flair NER	✗									
SPENS	✗				✗		✗	✗	✗	✗
Stanford NER					✗			✗	✗	✗
Stanza	✗					✗			✗	
NeuroTPR	✗		✗					✗		
RS	✗							✗	✗	
Geoparserpy		✗	✗		✗	✗	✗	✗		

limitations and offers enhanced robustness; (3) the performance of different approaches varies on the type of texts and location references, and their computational efficiency also varies drastically. Users should select the most suitable approach based on specific application demands.

Several research directions can be further explored in the future.

- **Location reference recognition**: The emergence of Large Language Models (LLMs) like GPT3 (175B) [28] and LLaMA (65B) [174] has garnered significant attention in recent years due to their exceptional language generation and processing capabilities. Notably, models like ChatGPT have further propelled advancements in various fields, including translation, sentiment analysis, text summarization, and information retrieval. Location reference recognition is no exception to the impact of these models. However, one significant challenge is the large size and high computational requirements of these LLMs, making them impractical for local deployment on personal computers. Therefore, a crucial future research direction is developing robust location reference recognizers that balance accurate performance, manageable memory, and computing costs.

- **Geocoding**: Existing geocoding or toponym resolution approaches primarily focus on formal texts, leaving a gap in handling informal texts like tweets. While some studies have proposed geocoding approaches for tweets, their applicability is often limited to specific known geographic regions, such as a city affected by a flood [3, 6]. In such cases, simply searching for a local gazetteer suffice for geocoding. Only a few studies attempted to tackle the challenge of geocoding tweets at a global scale [96, 139]. This task presents two main challenges: geo/geo ambiguities caused by limited contexts in short texts of tweets and unseen place names caused by place name variants and the informal features of tweets that often contain abbreviations, slang, and misspellings. Three main ways might be explored to overcome the challenges: (1) to leverage clustering techniques [43] that can group tweets of the same topic to expand the context of tweets; (2) to combine multiple state-of-the-art geocoders in a voting mechanism; (3) to integrate small or middle-scaled LLMs with global gazetteers to enhance geocoding capabilities.

- **Datasets for geoparsing research**: Current datasets for geoparsing research primarily consists of formal text, with limited availability of Twitter datasets designed for geoparsing. Existing Twitter datasets for general NER research often lack geographic coordinates for labeled entities, rendering them insufficient for comprehensive geoparsing research. Additional datasets comprising informal texts with labeled location references and corresponding geographic coordinates are necessary. Furthermore, most location references

in the existing datasets are admin units, such as countries and cities, while finer-grained location references, such as traffic ways and POIs, are scarce. However, they are essential in many applications, such as determining the precise locations where rescue is needed during disasters. A comprehensive Twitter dataset encompassing diverse fine-grained locations worldwide would significantly contribute to the progress of methods for recognizing and geocoding fine-grained location references in texts.

REFERENCES

[1] Ahmed Abdelkader, Emily Hand, and Hanan Samet. 2015. Brands in newsstand: Spatio-temporal browsing of business news. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–4.

[2] Elise Acheson and Ross S Purves. 2021. Extracting and modeling geographic information from scientific articles. *PloS one* 16, 1 (2021), e0244918.

[3] Mohammed Faisal Ahmed, Lelitha Vanajakshi, and Ramasubramanian Suriyanarayanan. 2019. Real-Time Traffic Congestion Information from Tweets Using Supervised and Unsupervised Machine Learning Techniques. *Transportation in Developing Economies* 5, 2 (2019), 1–11.

[4] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[5] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*. 1638–1649.

[6] Hussein Al-Olimat, Krishnaprasad Thirunarayan, Valerie Shalin, and Amit Sheth. 2018. Location Name Extraction from Targeted Text Streams using Gazetteer-based Statistical Language Models. *Proceedings of the 27th International Conference on Computational Linguistics* (Aug. 2018), 1986–1997. https://www.aclweb.org/anthology/C18-1169

[7] Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-NER: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 586–594.

[8] Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2020. CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing. *arXiv preprint arXiv:2004.06774* (2020).

[9] Edwin Aldana-Bobadilla, Alejandro Molina-Villegas, Ivan Lopez-Arevalo, Shanel Reyes-Palacios, Victor Muñiz-Sanchez, and Jean Arreola-Trapala. 2020. Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text. *Remote Sensing* 12, 18 (2020), 3041.

[10] Farman Ali, Amjad Ali, Muhammad Imran, Rizwan Ali Naqvi, Muhammad Hameed Siddiqi, and Kyung-Sup Kwak. 2021. Traffic accident detection and condition analysis based on social networking data. *Accident Analysis & Prevention* 151 (2021), 105973.

[11] Balsam Alkouz and Zaher Al Aghbari. 2020. SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks. *Information Processing & Management* 57, 1 (2020), 102139.

[12] Toph Allen, Kris A Murray, Carlos Zambrana-Torrelio, Stephen S Morse, Carlo Rondinini, Moreno Di Marco, Nathan Breit, Kevin J Olival, and Peter Daszak. 2017. Global hotspots and correlates of emerging zoonotic diseases. *Nature communications* 8, 1 (2017), 1–10.

[13] Ebtesam Alomari, Iyad Katib, Aiiad Albeshri, Tan Yigitcanlar, and Rashid Mehmood. 2021. Iktishaf+: A big data tool with automatic labeling for road traffic social sensing and event detection using distributed machine learning. *Sensors* 21, 9 (2021), 2993.

[14] Meshrif Alruily, Aladdin Ayesh, and Hussein Zedan. 2014. Crime profiling for the Arabic language using computational linguistic techniques. *Information processing & management* 50, 2 (2014), 315–341.

[15] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 273–280.

[16] Stelios Andreadis, Gerasimos Antzoulatos, Thanassis Mavropoulos, Panagiotis Giannakeris, Grigoris Tzionis, Nick Pantelidis, Konstantinos Ioannidis, Anastasios Karakostas, Ilias Gialampoukidis, Stefanos Vrochidis, et al. 2021. A social media analytics platform visualising the spread of COVID-19 in Italy via exploitation of automatically geotagged tweets. *Online Social Networks and Media* 23 (2021), 100134.

[17] Rexy Arulanandam, Bastin Tony Roy Savarimuthu, and Maryam A Purvis. 2014. Extracting crime information from online newspaper articles. In *Proceedings of the second australasian web conference-volume 155*. 31–38.

[18] Marco Avvenuti, Stefano Cresci, Fabio Del Vigna, Tiziano Fagni, and Maurizio Tesconi. 2018. CrisMap: a big data crisis mapping system based on damage detection and geoparsing. *Information Systems Frontiers* 20, 5 (2018), 993–1011.

[19] Marco Avvenuti, Stefano Cresci, Leonardo Nizzoli, and Maurizio Tesconi. 2018. GSP (Geo-Semantic-Parsing): geoparsing and geotagging with machine learning on top of linked data. In *European Semantic Web Conference*. Springer, 17–32.

[20] Dariusz B Baranowski, Maria K Flatau, Piotr J Flatau, Dwikorita Karnawati, Katarzyna Barabasz, Michal Labuz, Beata Latos, Jerome M Schmidt, Jaka AI Paski, et al. 2020. Social-media and newspaper reports reveal large-scale meteorological drivers of floods on Sumatra.

Nature communications 11, 1 (2020), 1–10.

[21] Moumita Basu, Sipra Das Bit, and Saptarshi Ghosh. 2022. Utilizing microblogs for optimized real-time resource allocation in post-disaster scenarios. *Social Network Analysis and Mining* 12, 1 (2022), 1–20.

[22] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. 2020. Learning political polarization on social media using neural networks. *IEEE Access* 8 (2020), 47177–47187.

[23] Loris Belcastro, Fabrizio Marozzo, Domenico Talia, Paolo Trunfio, Francesco Branda, Themis Palpanas, and Muhammad Imran. 2021. Using social media for sub-event detection during disasters. *Journal of Big Data* 8, 1 (2021), 1–22.

[24] Cillian Berragan, Alex Singleton, Alessia Calafiore, and Jeremy Morley. 2022. Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science* (2022), 1–20.

[25] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A Greenwood, Diana Maynard, and Niraj Aswani. 2013. Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. 83–90.

[26] Karla AV Borges, Alberto HF Laender, Claudia B Medeiros, and Clodoveu A Davis Jr. 2007. Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*. 31–36.

[27] Igo Ramalho Brilhante, Jose Antonio Macedo, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. 2015. On planning sightseeing tours with TripBuilder. *Information Processing & Management* 51, 2 (2015), 1–15.

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[29] Lou Burnard. 1995. *Users reference guide for the British National Corpus*. Oxford University Computing Services.

[30] Lucie Cadorel, Alicia Blanchi, and Andrea GB Tettamanzi. 2021. Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text. In *Proceedings of the 11th on Knowledge Capture Conference*. 41–48.

[31] Amparo Elizabeth Cano Basave, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. 2013. Making sense of microposts (# msm2013) concept extraction challenge. (2013).

[32] Zi Chen, Badal Pokharel, Bingnan Li, and Samsung Lim. 2020. Location Extraction from Twitter Messages using Bidirectional Long Short-Term Memory Model.. In *GISTAM*. 45–50.

[33] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 759–768.

[34] Rumi Chunara, Jason R Andrews, and John S Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene* 86, 1 (2012), 39.

[35] Paul Clough. 2005. Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval*. 25–30.

[36] Andrea Fronzetti Colladon, Barbara Guardabascio, and Rosy Innarella. 2019. Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. *Decision Support Systems* 123 (2019), 113075.

[37] Stefano Cresci, Andrea D'Errico, Davide Gazzé, Angelica Lo Duca, Andrea Marchetti, and Maurizio Tesconi. 2014. Towards a DBpedia of Tourism: the case of Tourpedia.. In *International Semantic Web Conference (Posters & Demos)*. 129–132.

[38] Rafael Prieto Curiel, Stefano Cresci, Cristina Ioana Muntean, and Steven Richard Bishop. 2020. Crime and its fear in social media. *Palgrave Communications* 6, 1 (2020), 1–12.

[39] James R Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. 164–167.

[40] Rahul Deb Das and Ross S Purves. 2019. Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India. *IEEE Transactions on Intelligent Transportation Systems* 21, 12 (2019), 5213–5222.

[41] Tirthankar Dasgupta, Lipika Dey, Rupsa Saha, and Abir Naskar. 2018. Automatic Curation and Visualization of Crime Related Information from Incrementally Crawled Multi-source News Reports. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. 103–107.

[42] Tirthankar Dasgupta, Abir Naskar, Rupsa Saha, and Lipika Dey. 2017. Crimeprofiler: Crime information extraction and visualization from news media. In *Proceedings of the International Conference on Web Intelligence*. 541–549.

[43] Jens A de Bruijn, Hans de Moel, Brenden Jongman, Jurjen Wagemaker, and Jeroen CJH Aerts. 2018. TAGGS: Grouping tweets to improve global geoparsing for disaster response. *Journal of Geovisualization and Spatial Analysis* 2, 1 (2018), 2.

[44] Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

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

[46] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487* (2017).

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

[48] Catherine D'Ignazio, Rahul Bhargava, Ethan Zuckerman, and Luisa Beck. 2014. Cliff-clavin: Determining geographic focus for news articles. *NewsKDD: Data Science for News Publishing, at KDD 2014* (2014).

[49] Christopher Donaldson, Ian N Gregory, and Joanna E Taylor. 2017. Locating the beautiful, picturesque, sublime and majestic: spatially analysing the application of aesthetic terminology in descriptions of the English Lake District. *Journal of Historical Geography* 56 (2017), 43–60.

[50] Ritam Dutt, Moumita Basu, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. Utilizing microblogs for assisting post-disaster relief operations via matching resource needs and availabilities. *Information Processing & Management* 56, 5 (2019), 1680–1697.

[51] Ritam Dutt, Kaustubh Hiware, Avijit Ghosh, and Rameshwar Bhaskaran. 2018. Savitr: A system for real-time location extraction from microblogs during emergencies. In *Companion Proceedings of the The Web Conference 2018*. 1643–1649.

[52] Chao Fan, Fangsheng Wu, and Ali Mostafavi. 2020. A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access* 8 (2020), 10478–10490.

[53] Nicolás José Fernández-Martínez and Carlos Periñán Pascual. 2021. *nLORE: A Linguistically Rich Deep-Learning System for Locative-Reference Extraction in Tweets.* https://doi.org/10.3233/AISE210103

[54] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1625–1628.

[55] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 363–370.

[56] Muhammad Imran Ferda Ofli Firoj Alam, Umair Qazi. 2021. HumAID: Human-Annotated Disaster Incidents Data from Twitter. In *15th International Conference on Web and Social Media (ICWSM)*.

[57] Nuno Freire, José Borbinha, Pável Calado, and Bruno Martins. 2011. A metadata geoparsing system for place name recognition and resolution in metadata records. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. 339–348.

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

[59] Judith Gelernter and Nikolai Mushegian. 2011. Geo-parsing messages from microtext. *Transactions in GIS* 15, 6 (2011), 753–773.

[60] Judith Gelernter and Wei Zhang. 2013. Cross-lingual geo-parsing for non-structured data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*. 64–71.

[61] Prasanna Giridhar, Tarek Abdelzaher, Jemin George, and Lance Kaplan. 2015. On quality of event localization from social network feeds. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 75–80.

[62] Ian Gregory, Paul Atkinson, Andrew Hardie, Amelia Joulain-Jay, Daniel Kershaw, Catherine Porter, Paul Rayson, and Christopher John Rupp. 2016. From digital resources to historical scholarship with the British library 19th century newspaper collection. (2016).

[63] Ian Gregory, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. 2015. Geoparsing, GIS, and textual analysis: current developments in spatial humanities research. *International Journal of Humanities and Arts Computing* 9, 1 (2015), 1–14.

[64] Milan Gritta. 2019. *Where are you talking about? advances and challenges of geographic analysis of text with application to disease monitoring.* Ph. D. Dissertation. University of Cambridge.

[65] Milan Gritta, Mohammad Pilehvar, and Nigel Collier. 2018. Which melbourne? augmenting geocoding with maps. (2018).

[66] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. A pragmatic guide to geoparsing evaluation. *arXiv preprint arXiv:1810.12368* (2018).

[67] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A pragmatic guide to geoparsing evaluation. *Language resources and evaluation* 54, 3 (2020), 683–712.

[68] Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018. What's missing in geographical parsing? *Language Resources and Evaluation* 52, 2 (2018), 603–623.

[69] Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, 1925 (2010), 3875–3889.

[70] Marco Guerini, Simone Magnolini, Vevake Balaraman, and Bernardo Magnini. 2018. Toward zero-shot entity recognition in task-oriented conversational agents. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. 317–326.

[71] Carlos Gutierrez, Paulo Figuerias, Pedro Oliveira, Ruben Costa, and Ricardo Jardim-Goncalves. 2015. Twitter mining for traffic events detection. In *2015 Science and information conference (SAI)*. IEEE, 371–378.

[72] Mena B Habib and Maurice van Keulen. 2013. A hybrid approach for robust multilingual toponym extraction and disambiguation. In *Intelligent Information Systems Symposium*. Springer, 1–15.

[73] Andrew Halterman. 2017. Mordecai: Full text geoparsing and event geocoding. *Journal of Open Source Software* 2, 9 (2017), 91.

[74] Erum Haris and Keng Hoon Gan. 2017. Mining graphs from travel blogs: a review in the context of tour planning. *Information Technology & Tourism* 17, 4 (2017), 429–453.

[75] Erum Haris, Keng Hoon Gan, and Tien-Ping Tan. 2020. Spatial information extraction from travel narratives: Analysing the notion of co-occurrence indicating closeness of tourist places. *Journal of Information Science* 46, 5 (2020), 581–599.

[76] Jenine K Harris, Sarah Moreland-Russell, Rachel G Tabak, Lindsay R Ruhr, and Ryan C Maier. 2014. Communication about childhood obesity on Twitter. *American journal of public health* 104, 7 (2014), e62–e69.

[77] Jingrui He, Wei Shen, Phani Divakaruni, Laura Wynter, and Rick Lawrence. 2013. Improving traffic prediction with tweet semantics. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[78] Uta Hinrichs, Beatrice Alex, Jim Clifford, Andrew Watson, Aaron Quigley, Ewan Klein, and Colin M Coates. 2015. Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. *Digital Scholarship in the Humanities* 30, suppl_1 (2015), i50–i75.

[79] Thi Bich Ngoc Hoang and Josiane Mothe. 2018. Location extraction from tweets. *Information Processing & Management* 54, 2 (2018), 129–144.

[80] Alexander Hohl, Moongi Choi, Richard Medina, Neng Wan, and Ming Wen. 2021. Understanding Adverse Population Sentiment Towards the Spread of COVID-19 in the United States. *medRxiv* (2021).

[81] Xuke Hu, Hussein Al-Olimat, Jens Kersten, Matti Wiegmann, Friederike Klan, Yeran Sun, and Hongchao Fan. 2021. GazPNE: Annotation-free Deep Learning for Place Name Extraction from Microblogs Leveraging Gazetteer and Synthetic Data by Rules. *International Journal of Geographical Information Science* (2021), 1–28. https://doi.org/10.1080/13658816.2021.1947507

[82] Xuke Hu, Zhiyong Zhou, Yeran Sun, Jens Kersten, Friederike Klan, Hongchao Fan, and Matti Wiegmann. 2022. GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models. *IEEE Internet of Things Journal* (2022), 1–1. https://doi.org/10.1109/JIOT.2022.3150967

[83] Yingjie Hu and Benjamin Adams. 2021. Harvesting big geospatial data from natural language texts. In *Handbook of Big Geospatial Data*. Springer, 487–507.

[84] Yingjie Hu, Krzysztof Janowicz, and Sathya Prasad. 2014. Improving wikipedia-based place name disambiguation in short texts using structured data from dbpedia. In *Proceedings of the 8th workshop on geographic information retrieval*. 1–8.

[85] Yingjie Hu and Jimin Wang. 2020. How do people describe locations during a natural disaster: an analysis of tweets from Hurricane Harvey. *arXiv preprint arXiv:2009.12914* (2020).

[86] Yingjie Hu and Ruo-Qian Wang. 2020. Understanding the removal of precise geotagging in tweets. *Nature Human Behaviour* 4, 12 (2020), 1219–1221.

[87] Haosheng Huang, Georg Gartner, Jukka M Krisp, Martin Raubal, and Nico Van de Weghe. 2018. Location based services: ongoing evolution and research agenda. *Journal of Location Based Services* 12, 2 (2018), 63–93.

[88] Qunying Huang and Yu Xiao. 2015. Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information* 4, 3 (2015), 1549–1568.

[89] Xiao Huang, Zhenlong Li, Yuqin Jiang, Xiaoming Li, and Dwayne Porter. 2020. Twitter reveals human mobility dynamics during the COVID-19 pandemic. *PloS one* 15, 11 (2020), e0241957.

[90] Diana Inkpen, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi, and Diman Ghazi. 2017. Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems* 49, 2 (2017), 237–253.

[91] Aminreza Iranmanesh and Resmiye Alpar Atun. 2021. Reading the changing dynamic of urban social distances during the COVID-19 pandemic via Twitter. *European Societies* 23, sup1 (2021), S872–S886.

[92] Christopher B Jones, Ross Purves, Anne Ruas, Mark Sanderson, Monika Sester, M Van Kreveld, and Robert Weibel. 2002. Spatial information retrieval and geographical ontologies an overview of the SPIRIT project. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 387–388.

[93] Kate E Jones, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, and Peter Daszak. 2008. Global trends in emerging infectious diseases. *Nature* 451, 7181 (2008), 990–993.

[94] Yiting Ju, Benjamin Adams, Krzysztof Janowicz, Yingjie Hu, Bo Yan, and Grant McKenzie. 2016. Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *European Knowledge Acquisition Workshop*. Springer, 353–367.

[95] Ehsan Kamalloo and Davood Rafiei. 2018. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*. 1287–1296.

[96] Morteza Karimzadeh, Scott Pezanowski, Alan M MacEachren, and Jan O Wallgrün. 2019. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS* 23, 1 (2019), 118–136.

[97] Mikaela Keller, Clark C Freifeld, and John S Brownstein. 2009. Automated vocabulary discovery for geo-parsing online epidemic intelligence. *BMC bioinformatics* 10, 1 (2009), 1–9.

[98] Sarthak Khanal and Doina Caragea. 2021. Multi-task Learning to Enable Location Mention Identification in the Early Hours of a Crisis Event. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4051–4056.

[99] Hiroshi Kori, Shun Hattori, Taro Tezuka, and Katsumi Tanaka. 2007. Automatic generation of multimedia tour guide from local blogs. In *International conference on multimedia modeling*. Springer, 690–699.

[100] Abhinav Kumar and Jyoti Prakash Singh. 2019. Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction* 33 (2019), 365–375.

[101] Abhinav Kumar, Jyoti Prakash Singh, and Nripendra P Rana. 2017. Authenticity of geo-location and place name in tweets. (2017).

[102] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

[103] Rabindra Lamsal. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence* 51, 5 (2021), 2790–2804.

[104] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[105] Jochen L Leidner and Michael D Lieberman. 2011. Detecting geographical references in the form of place names and associated spatial natural language. *Sigspatial Special* 3, 2 (2011), 5–11.

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

[107] Michael D Lieberman and Hanan Samet. 2011. Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 843–852.

[108] Michael D Lieberman and Hanan Samet. 2012. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 731–740.

[109] Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)*. IEEE, 201–212.

[110] Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for named entity recognition in Twitter messages. *COLING 2016* (2016).

[111] John Lingad, Sarvnaz Karimi, and Jie Yin. 2013. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on world wide web*. 1017–1020.

[112] Fei Liu. 2013. Automatic identification of locative expressions from informal text. (2013).

[113] Fei Liu, Maria Vasardani, and Timothy Baldwin. 2014. Automatic identification of locative expressions from social media text: A comparative analysis. In *Proceedings of the 4th International Workshop on Location and the Web*. 9–16.

[114] Xiao Liu, Haixiang Guo, Yu-ru Lin, Yijing Li, and Jundong Hou. 2018. Analyzing spatial-temporal distribution of natural hazards in China by mining news sources. *Natural Hazards Review* 19, 3 (2018), 04018006.

[115] Xinyue Liu, Xiangnan Kong, and Yanhua Li. 2016. Collective traffic prediction with partially observed traffic history using location-based social media. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 2179–2184.

[116] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[117] Kai Ma, YongJian Tan, Zhong Xie, Qinjun Qiu, and Siqiong Chen. 2022. Chinese toponym recognition with variant neural structures from social media messages based on BERT methods. *Journal of Geographical Systems* (2022), 1–27.

[118] Arjun Magge, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2018. Deep neural networks and distant supervision for geographic location mention extraction. *Bioinformatics* 34, 13 (2018), i565–i573.

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

[120] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[121] Huina Mao, Gautam Thakur, Kevin Sparks, Jibonananda Sanyal, and Budhendra Bhaduri. 2018. Mapping near-real-time power outages from social media. *International Journal of Digital Earth* (2018).

[122] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894* (2019).

[123] Nicolás José Fernández Martínez and Carlos Periñán Pascual. 2020. Reglas basadas en conocimiento para la extraccion de referencias locativas complejas en tweets. *RaeL Revista Electronica de Linguistica Aplicada* 19 (2020), 136–164.

[124] Nicolás José Fernández Martínez and Carlos Periñán-Pascual. 2020. Knowledge-based rules for the extraction of complex, fine-grained locative references from tweets. *RAEL: revista electrónica de lingüística aplicada* 19, 1 (2020), 136–163.

[125] Fernando Melo and Bruno Martins. 2017. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS* 21, 1 (2017), 3–38.

[126] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. 1–8.

[127] Stuart E Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 1–27.

[128] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[129] Sveta Milusheva, Robert Marty, Guadalupe Bedoya, Sarah Williams, Elizabeth Resor, and Arianna Legovini. 2021. Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning. *PloS one* 16, 2 (2021), e0244317.

[130] Andrei Mircea. 2020. Real-time classification, geolocation and interactive visualization of COVID-19 information shared on social media to better understand global developments. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

[131] Ludovic Moncla, Mauro Gaio, Thierry Joliveau, and Yves-François Le Lay. 2017. Automated geoparsing of paris street names in 19th century novels. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. 1–8.

[132] Ludovic Moncla, Mauro Gaio, and Sébastien Mustiere. 2014. Automatic itinerary reconstruction from texts. In *International Conference on Geographic Information Science*. Springer, 253–267.

[133] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd acm sigspatial international conference on advances in geographic information systems*. 183–192.

[134] Bruno R Monteiro, Clodoveu A Davis Jr, and Fred Fonseca. 2016. A survey on the geographic scope of textual documents. *Computers & Geosciences* 96 (2016), 23–34.

[135] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7.

[136] Patricia Murrieta-Flores, Alistair Baron, Ian Gregory, Andrew Hardie, and Paul Rayson. 2015. Automatically analyzing large texts in a GIS environment: The registrar general's reports and cholera in the 19th century. *Transactions in GIS* 19, 2 (2015), 296–320.

[137] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200* (2020).

[138] Malvina Nissim, Colin Matheson, and James Reid. 2004. Recognising geographical entities in Scottish historical documents. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*, Vol. 35. Citeseer.

[139] Leonardo Nizzoli, Marco Avvenuti, Maurizio Tesconi, and Stefano Cresci. 2020. Geo-semantic-parsing: AI-powered geoparsing by traversing semantic knowledge graphs. *Decision Support Systems* 136 (2020), 113346.

[140] Jesse O'Shea. 2017. Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International journal of medical informatics* 101 (2017), 15–22.

[141] Sharon Myrtle Paradesi. 2011. Geotagging tweets using their content. In *Twenty-Fourth International FLAIRS Conference*.

[142] Kelly S Peterson, Julia Lewis, Olga V Patterson, Alec B Chapman, Daniel W Denhalter, Patricia A Lye, Vanessa W Stevens, Shantini D Gamage, Gary A Roselle, Katherine S Wallace, et al. 2021. Automated Travel History Extraction From Clinical Notes for Informing the Detection of Emergent Infectious Disease Events: Algorithm Development and Validation. *JMIR public health and surveillance* 7, 3 (2021), e26719.

[143] Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, et al. 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. *arXiv preprint cs/0609065* (2006).

[144] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tom De Groeve. 2004. Geographical information recognition and visualization in texts written in various languages. In *Proceedings of the 2004 ACM symposium on Applied computing*. 1051–1058.

[145] Ross S Purves, Paul Clough, Christopher B Jones, Avi Arampatzis, Benedicte Bucher, David Finch, Gaihua Fu, Hideo Joho, Awase Khirni Syed, Subodh Vaid, et al. 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International journal of geographical information science* 21, 7 (2007), 717–745.

[146] Ross S Purves, Paul Clough, Christopher B Jones, Mark H Hall, and Vanessa Murdock. 2018. Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval* 12, 2-3 (2018), 164–318.

[147] Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special* 12, 1 (2020), 6–15.

[148] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082* (2020).

[149] Qinjun Qiu, Zhong Xie, Shu Wang, Yunqiang Zhu, Hairong Lv, and Kai Sun. [n. d.]. ChineseTR: A weakly supervised toponym recognition architecture based on automatic training data generator and deep neural network. *Transactions in GIS* ([n. d.]).

[150] F Rahma and A Romadhony. 2021. Rule-Based Crime Information Extraction on Indonesian Digital News. In *2021 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, 10–15.

[151] Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. 147–155.

[152] Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*. 1524–1534.

[153] Giuseppe Rizzo, Amparo Elizabeth Cano Basave, Bianca Pereira, Andrea Varga, Matthew Rowe, Milan Stankovic, and A Dadzie. 2015. Making Sense of Microposts (# Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge.. In *# MSM*. 44–53.

[154] CJ Rupp, Paul Rayson, Alistair Baron, Christopher Donaldson, Ian Gregory, Andrew Hardie, and Patricia Murrieta-Flores. 2013. Customising geoparsing and georeferencing for historical texts. In *2013 IEEE International Conference on Big Data*. IEEE, 59–62.

[155] Meryem Sagcan and Pinar Karagoz. 2015. Toponym recognition in social media for estimating the location of events. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 33–39.

[156] SPCW Sandagiri, BTGS Kumara, and Banujan Kuhaneswaran. 2020. Detecting Crimes Related Twitter Posts Using SVM based Two Stages Filtering. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 506–510.

[157] Christopher Scheele, Manzhu Yu, and Qunying Huang. 2021. Geographic context-aware text mining: enhance social media message classification for situational awareness by integrating spatial and temporal features. *International Journal of Digital Earth* 14, 11 (2021), 1721–1743.

[158] Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. arXiv:2011.06993 [cs.CL]

[159] Peter Scott, Martin K-F Bader, Treena Burgess, Giles Hardy, and Nari Williams. 2019. Global biogeography and invasion risk of the plant pathogen genus Phytophthora. *Environmental Science & Policy* 101 (2019), 175–182.

[160] Jianga Shang, Xuke Hu, Fuqiang Gu, Di Wang, and Shengsheng Yu. 2015. Improvement schemes for indoor mobile location estimation: A survey. *Mathematical Problems in Engineering* 2015 (2015).

[161] Lanyu Shang, Yang Zhang, Christina Youn, and Dong Wang. 2022. SAT-Geo: A social sensing based content-only approach to geolocating abnormal traffic events using syntax-based probabilistic learning. *Information Processing & Management* 59, 2 (2022), 102807.

[162] Eric Shook and Victoria K Turner. 2016. The socio-environmental data explorer (SEDE): a social media–enhanced decision support system to explore risk perception to hazard events. *Cartography and Geographic Information Science* 43, 5 (2016), 427–441.

[163] Mário J Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. 2006. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems* 30, 4 (2006), 378–399.

[164] Jyoti Prakash Singh, Yogesh K Dwivedi, Nripendra P Rana, Abhinav Kumar, and Kawaljeet Kaur Kapoor. 2019. Event classification and location prediction from tweets during disasters. *Annals of Operations Research* 283, 1 (2019), 737–757.

[165] N Sobhana, Pabitra Mitra, and SK Ghosh. 2010. Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications* 1, 3 (2010), 143–147.

[166] K Srinivasa and P Santhi Thilagam. 2019. Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers. *Information Processing & Management* 56, 6 (2019), 102059.

[167] Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. 138–144.

[168] Nestor Suat-Rojas, Camilo Gutierrez-Osorio, and Cesar Pedraza. 2022. Extraction and Analysis of Social Networks Data to Detect Traffic Accidents. *Information* 13, 1 (2022), 26.

[169] Evan A Sultanik and Clayton Fink. 2012. Rapid geotagging and disambiguation of social media text via an indexed gazetteer.. In *ISCRAM*.

[170] Javier Tamames and Victor de Lorenzo. 2010. EnvMine: A text-mining system for the automatic extraction of contextual information. *BMC bioinformatics* 11, 1 (2010), 1–10.

[171] Andrea H Tapia, Kathleen A Moore, and Nichloas J Johnson. 2013. Beyond the trustworthy tweet: A deeper understanding of microblogged data use by disaster response and humanitarian relief organizations.. In *ISCRAM*.

[172] Laura Tateosian, Rachael Guenter, Yi-Peng Yang, and Jean Ristaino. 2017. Tracking 19th century late blight from archival documents using text analytics and geoparsing. In *Free and open source software for geospatial (FOSS4G) conference proceedings*, Vol. 17. 17.

[173] Benjamin E Teitler, Michael D Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. 2008. NewsStand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. 1–10.

[174] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[175] Sayan Unankard, Xue Li, and Mohamed A Sharaf. 2015. Emerging event detection in social networks with location sensitivity. *World Wide Web* 18, 5 (2015), 1393–1417.

[176] Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An All-Round Python Library for Transformer-based Named Entity Recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 53–62.

[177] Maria Vasardani, Stephan Winter, and Kai-Florian Richter. 2013. Locating place names from place descriptions. *International Journal of Geographical Information Science* 27, 12 (2013), 2509–2532.

[178] Lara Vomfell, Wolfgang Karl Härdle, and Stefan Lessmann. 2018. Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems* 113 (2018), 73–85.

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

[180] Jimin Wang and Yingjie Hu. 2019. Are we there yet? Evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*. 1–6.

[181] Jimin Wang and Yingjie Hu. 2019. Enhancing spatial and textual analysis with EUPEG: an extensible and unified platform for evaluating geoparsers. *Transactions in GIS* 23, 6 (2019), 1393–1419.

[182] Jimin Wang, Yingjie Hu, and Kenneth Joseph. 2020. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS* (2020).

[183] Ruo-Qian Wang, Yingjie Hu, Zikai Zhou, and Kevin Yang. 2020. Tracking flooding phase transitions and establishing a passive hotline with ai-enabled social media data. *IEEE Access* 8 (2020), 103395–103404.

[184] Wei Wang and Kathleen Stewart. 2015. Spatiotemporal and semantic information extraction from Web news reports about natural hazards. *Computers, environment and urban systems* 50 (2015), 30–40.

[185] Barney Warf and Santa Arias. 2008. *The spatial turn: Interdisciplinary perspectives.* Routledge.

[186] Davy Weissenbacher, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez. 2019. Semeval-2019 task 12: Toponym resolution in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 907–916.

[187] Davy Weissenbacher, Abeed Sarker, Tasnia Tahsin, Matthew Scotch, and Graciela Gonzalez. 2017. Extracting geographic locations from the literature for virus phylogeography using supervised and distant supervision methods. *AMIA Summits on Translational Science Proceedings* 2017 (2017), 114.

[188] Davy Weissenbacher, Tasnia Tahsin, Rachel Beard, Mari Figaro, Robert Rivera, Matthew Scotch, and Graciela Gonzalez. 2015. Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. *Bioinformatics* 31, 12 (2015), i348–i356.

[189] Stephan Winter, Kai-Florian Richter, Tim Baldwin, Lawrence Cavedon, Lesley Stirling, Matt Duckham, Allison Kealy, and Abbas Rajabifard. 2011. Location-based mobile games for spatial knowledge acquisition. *Cognitive Engineering for Mobile GIS* 780 (2011), 1–8.

[190] Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. ensemble named entity recognition (ner): evaluating ner Tools in the identification of Place names in historical corpora. *Frontiers in Digital Humanities* 5 (2018), 2.

[191] Allison Gyle Woodruff and Christian Plaunt. 1994. GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science* 45, 9 (1994), 645–655.

[192] Desheng Wu and Yiwen Cui. 2018. Disaster early warning and damage assessment analysis using social media data and geo-location information. *Decision support systems* 111 (2018), 48–59.

[193] Canwen Xu, Jing Li, Xiangyang Luo, Jiaxin Pei, Chenliang Li, and Donghong Ji. 2019. DLocRL: A deep learning pipeline for fine-grained location recognition and linking in tweets. In *The World Wide Web Conference*. 3391–3397.

[194] MM Yagoub, Aishah A Alsereidi, Elfadil A Mohamed, Punitha Periyasamy, Reem Alameri, Salama Aldarmaki, and Yaqein Alhashmi. 2020. Newspapers as a validation proxy for GIS modeling in Fujairah, United Arab Emirates: identifying flood-prone areas. *Natural Hazards* 104, 1 (2020), 111–141.

[195] Pranali Yenkar and SD Sawarkar. 2021. Gazetteer based unsupervised learning approach for location extraction from complaint tweets. In *IOP Conference Series: Materials Science and Engineering*, Vol. 1049. IOP Publishing, 012009.

[196] Hua Yuan, Hualin Xu, Yu Qian, and Yan Li. 2016. Make your travel smarter: Summarizing urban tourism information from massive blog data. *International Journal of Information Management* 36, 6 (2016), 1306–1319.

[197] Yongjian Zhu, Liqing Cao, Jingui Xie, Yugang Yu, Anfan Chen, and Fengming Huang. 2021. Using social media data to assess the impact of COVID-19 on mental health in China. *Psychological medicine* (2021), 1–8.

[198] Lei Zou, Danqing Liao, Nina SN Lam, Michelle A Meyer, Nasir G Gharaibeh, Heng Cai, Bing Zhou, and Dongying Li. 2023. Social media for emergency rescue: An analysis of rescue requests on Twitter during Hurricane Harvey. *International Journal of Disaster Risk Reduction* 85 (2023), 103513.