Chapter 0
Harvesting Big Geospatial Data from Natural Language Texts

Yingjie Hu[1] and Benjamin Adams[2]
[1]Department of Geography, University at Buffalo, Buffalo, USA
[2]Department of Computer Science and Software Engineering, University of Canterbury, Christchurch, New Zealand

Abstract A vast amount of geospatial data exists in natural language texts, such as newspapers, Wikipedia articles, social media posts, travel blogs, online reviews, and historical archives. Compared with more traditional and structured geospatial data, such as those collected by the US Geological Survey and the national statistics offices, geospatial data harvested from these unstructured texts have unique merits. They capture valuable human experiences toward places, reflect near real-time situations in different geographic areas, or record important historical information that is otherwise not available. In addition, geospatial data from these unstructured texts are often *big*, in terms of their *volume*, *velocity*, and *variety*. This chapter presents the motivations of harvesting big geospatial data from natural language texts, describes typical methods and tools for doing so, summarizes a number of existing applications, and discusses challenges and future directions.

0.1 Introduction and Motivation

Geospatial information is produced by a wide variety of data sources. In addition to commonly used datasets from agencies such as the US Geological Survey (USGS) and the US Census, geospatial information is contained in news articles (Lieberman and Samet, 2011; Liu et al, 2014), encyclopedia entries (Hecht and Raubal, 2008; Salvini and Fabrikant, 2016), social media posts (Keßler et al, 2009b; Zhang and Gelernter, 2014), historical archives (Southall, 2014; DeLozier et al, 2016), housing advertisements (Madden, 2017; McKenzie et al, 2018), online reviews (Cataldi et al, 2013; Wang and Zhou, 2016), travel blog entries (Adams and McKenzie, 2013; Ballatore and Adams, 2015), and other sources. From these sources, geospatial data

is embedded in natural language texts and is often presented in the form of place name mentions and place descriptions. For example, a social media post or a news article might mention multiple places through their names, or a travel blog might describe the experience of the writer at a particular place. In today's Big Data era, the *volume* and *variety* of the data from these sources are increasing at an unprecedented *velocity*, and it has become feasible to harvest big geospatial data from texts.

Why do we want to harvest geospatial data from texts? Asking this question is important, since collections of natural language text, e.g., those from social media or news articles, are often not representative of the entire population (Hecht and Stephens, 2014; Malik et al, 2015; Jiang et al, 2018). There are at least three aspects in which the geospatial data harvested from texts is valuable. First, they can provide valuable human experience information, which is not available in other datasets. Travel blog entries, for example, do not simply describe where people have been but also what their *feelings* are toward these places. Such information about human experience is critical for building computational models of *places* (Goodchild, 2011; Merschdorf and Blaschke, 2018). Second, geospatial data harvested from some natural language texts, such as social media posts, reflect near real-time situations and are valuable for applications such as disaster response (MacEachren et al, 2011; Crooks et al, 2013; Huang and Xiao, 2015). This is an important advantage compared with data from questionnaire-based surveys or face-to-face interviews which can take often months or even a few years to produce. While the geospatial data harvested from social media may not be representative, disaster response and other situation awareness applications often focus on identifying incidents, rather than, for example, whether the three people trapped in a collapsed building represent the entire population in the study area. Third, some geospatial data is only available in unstructured texts. Examples include events reported in newspapers, historical battles recorded in old archives, or business addresses contained in Web pages (Nesi et al, 2016; Hu et al, 2017; Barbaresi, 2017). In these cases, harvesting geospatial data from texts is necessary for enabling advanced spatial analysis.

Harvesting geospatial data from unstructured texts has been frequently studied in geographic information retrieval (GIR) under the topic of *geoparsing* (Jones and Purves, 2008; Purves et al, 2018). The goal of geoparsing is to recognize the place names, or *toponyms*, mentioned in texts, and identify the corresponding instances and the location coordinates of the recognized place names (Freire et al, 2011; Gritta et al, 2018). A software tool developed for geoparsing is called a *geoparser*, which takes unstructured natural language texts as the input, and outputs structured geographic data with the recognized place names and their location coordinates. Some geoparsers, e.g., GeoTxt (Karimzadeh et al, 2013), are published as Web services which provide easy access for general users through the Internet.

Geoparsing is typically performed in two consecutive steps: toponym recognition and toponym resolution. For the first step, the goal is to recognize place names from natural language texts without identifying the particular place instance referred by a name. For example, in the sentence, "Washington was an important stop on the rugged Southwest Trail.", the term "Washington" will be recognized as a toponym, but this step will not attempt to understand which Washington this term specifically

refers to (there are more than 50 places named "Washington" in the United States). The second step, toponym resolution, aims to address the place name ambiguity and resolve the place name to its correct instance and geographic location. The toponym resolution step will (ideally) find out that the name "Washington" refers to "Washington, Arkansas" in the sentence, and will locate the place name to its corresponding spatial footprint, such as the geometric center of the city boundary. Figure 3.1 provides an overview of the two steps of geoparsing. The geospatial data harvested from natural language texts usually contain the recognized place names and their spatial footprints, such as points, lines, and polygons.

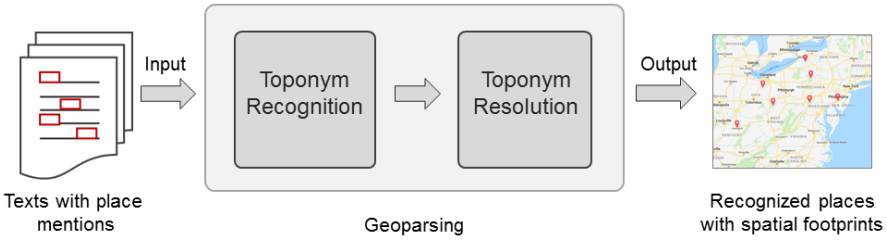


Fig. 0.1 An overview of the input, output, and the two steps of geoparsing.

Geospatial data can also be harvested from texts that do not explicitly mention place names (Wing and Baldridge, 2014). Non-spatial words, such as *beach* and *sunshine*, can be *geo-indicative* (Adams and Janowicz, 2012). That is, in the context of a textual corpus containing documents which are associated with locations on the Earth, certain words and phrases can be more or less likely to be associated with specific locations. Words with non-random spatial distributions will be most apparent in texts that describe physical environments and/or local cultural practices. Texts that are geo-referenced enable us to discover useful knowledge about places. This can be done subsequent to geoparsing as well as on texts that are already geo-referenced by the source. Examples of the latter include tweets with GPS location and travel blog entries tagged with named places (Hahmann et al, 2014; Adams and McKenzie, 2013). For shorter documents it is often the case that the entire text content can be associated with one or a few toponyms. However, for longer texts the task of associating toponyms with the correct selections from the text is still an open research problem and may require more sophisticated semantic entity linking and relation extraction, reflecting a lack of easy-to-use tools in this space.

The remainder of this chapter is organized as follows. Section 2 reviews methods on recognizing and resolving place names from texts, and lists existing geoparsers and human-annotated corpora. Section 3 discusses a number of studies that have harvested big geospatial data from natural langauge texts for various applications. Particularly, these studies are organized into three topics: place-related studies, time-sensitive applications, and special information extraction. Finally, Section 4 presents the challenges and possible directions for the near future.

0.2 Methods and Tools

Various methods have been proposed for harvesting big geospatial data from natural language texts. In this section, we first review the existing methods for toponym recognition and resolution respectively, and then describe the existing tools for completing these two steps. We also discuss location inference from texts using language models, and such approaches are especially useful when texts do not explicitly contain toponyms.

0.2.1 Toponym recognition

The goal of toponym recognition is to recognize the toponyms mentioned in natural language texts. One typical approach is to use a *gazetteer* which is a geographic dictionary that contains organized collections of place names, place types, and spatial footprints (Hill, 2000; Janowicz and Keßler, 2008). Since humans refer to places via their names while machines represent places by their coordinates, gazetteers fill the critical gap between informal human discourses and formal computer representations (Goodchild and Hill, 2008; Keßler et al, 2009a). Accordingly, we can compare natural langauge texts with the entries in a gazetteer to identify the contained place names. For example, Woodruff and Plaunt (1994) used a subset of the Geographic Names Information System (GNIS) gazetteer to identify place names from textual documents related to the region of California. Amitay et al (2004) proposed a system called *Web-a-Where* which can recognize place names from Web pages based on a gazetteer containing continents, countries, states, and cities throughout the world. While straightforward, a main disadvantage of this direct matching approach is that some place names or their vernacular versions may not be contained in a gazetteer and therefore cannot be recognized. To address this issue, methods have been proposed to enrich existing gazetteers with vernacular or vague place names. For example, Twaroch and Jones (2010) proposed a platform, called "People's Place Names" (`http://www.yourplacenames.com`), which encourages local people to contribute vernacular place names. Gelernter et al (2013) developed an automatic algorithm which can add place names from OpenStreetMap and Wikimapia into a gazetteer. Jones et al (2008) developed an approach that leverages a Web search engine to harvest entities related to a vague place name in order to construct its boundary. Geotagged photos and the associated textual tags were also used by many researchers for adding vague places into gazetteers (Grothe and Schaab, 2009; Keßler et al, 2009b; Intagorn and Lerman, 2011; Li and Goodchild, 2012). More recently, geotagged housing posts, in which vernacular place names are often mentioned, were examined for their potential in providing local place names and enriching gazetteers (McKenzie et al, 2018; Hu et al, 2018).

Another approach for recognizing place names from texts is to use natural language processing (NLP) techniques. A key advantage of this approach is that it can be used to identify place names without relying on a gazetteer: it makes use of the

words within the local context of a target word (e.g., the previous and next five words surrounding the target word) to infer whether the target word is part of a place name. One simple way to implement this idea is to define a set of grammartical rules for recognizing toponyms. For example, names in the patterns of "City of ⟨name⟩" and "⟨name⟩ Boulevard" are often place names, while those in the patterns of "Firstname ⟨name⟩" are typically not (Purves et al, 2018). Since these grammatical rules need to be defined manually, machine learning based approaches were proposed to recognize toponyms based on contextual evidence in the text. From this perspective, toponym recognition can be considered as a sub-task of named entity recognition (NER). One frequently used NER tool is Stanford NER which is based on a Conditional Random Field (CRF) sequence model (Finkel et al, 2005) and can recognize multiple types of named entities from texts, such as locations, persons, and organizations. To recognize toponyms, one can limit the identified entities to locations only. Many existing studies have included Stanford NER as part of their workflows. For example, Karimzadeh et al (2013) developed GeoTxt in which the Stanford NER is employed for the named entity recognition step. Gelernter and Mushegian (2011) also used Stanford NER to identify location names from the tweets after the 2011 earthquake in Christchurch, Canterbury. Lieberman et al (2010) leveraged Stanford NER to find location entities from local news articles in order to build spatial indices for textual data. In addition to Stanford NER, researchers also made use of other NER models. For example, Gelernter et al (2013) employed OpenCalais to find building names from texts, and Hu et al (2018) used spaCy NER as one of their four NER models to recognize place names from geotagged housing posts. Many studies also trained their own NER models for toponym recognition by leveraging a variety of evidence from the data, such as part of speech (POS) tags, left words, right words, entity relations, and other possible cues (Lieberman and Samet, 2011; Inkpen et al, 2015).

0.2.2 Toponym resolution

Once place names are recognized from texts in the first step, the second step aims to resolve these names to their corresponding geographic instances. This step is necessary because of the ambiguity existing in the semantics of place names (Leidner, 2008). Amitay et al (2004) discussed two types of ambiguities: geo/geo ambiguity, i.e., the same name, such as *London*, can refer to different geographic instances in the world; and geo/non-geo ambiguity, i.e., the same name, such as *Washington*, can refer to not only places but also persons and other types of entities. Besides, there is the issue of metonymy. For example, we may have a sentence "London voted to pass an act", in which "London" may not represent the place but the government entity, although it is not entirely unreasonable to recognize and resolve "London" to the capital of the UK in this sentence. Perhaps due to this debatable issue, many geoparsers do not directly handle metonymies. In addition, the toponyms recognized in the first step may contain false positives and false negatives. The false positives, i.e.,

the non-place phrases that are mistakenly recognized as toponyms, can be handled by toponym resolution methods in the process of resolving geo/non-geo ambiguity. The false negatives, i.e., the place names that are missed by the toponym recognition step, are more difficult to deal with, since most toponym resolution methods start with only the recognized toponyms rather than trying to expand the set. How to recover these false negatives could be an interesting future research topic.

A variety of methods have been developed for toponym resolution. Early approaches often make use of certain domain knowledge about places (e.g., total population) to define heuristic rules for disambiguation. A simple approach is to resolve a place name to its most prominent or *default* place instance, such as the one that has the highest population or the largest total area (these types of information are often available in gazetteers). Li et al (2002) proposed a method for identifying the default sense of a place name based on the results returned by a search engine (Yahoo!), and their experiments showed that using the obtained default senses alone can already achieve a fair performance (i.e., resolving 78% of their ambiguous place names). Ladra et al (2008) developed a toponym resolution Web service which combined administrative hierarchies, the populations of different places, whether a place is a capital or a main city, and some other information to perform place name disambiguation. Some other rules, such as *one referent per document* (i.e., a toponym that appears in different parts of the same document will most likely refer to the same place instance), were also developed (Leidner, 2008). While hand-crafted rules can already resolve many toponyms, they can be incomplete or arbitrary: Which rules should be included and which should not? How to define the threshold for a city to be considered as a *main city*? And which rules should have higher priorities over other rules? Besides, much manual effort is needed to develop these rules.

Due to the limitations of hand-crafted rules, automatic or semi-automatic approaches are proposed for toponym resolution. Overell and Rüger (2008) proposed a co-occurrence model based on how place names occur together in Wikipedia, and then applied the co-occurrence model to disambiguate place names from texts. Buscaldi and Rosso (2008) developed a conceptual density based approach which disambiguates toponyms using an external reference corpus GeoSemCor. Lieberman and Samet (2011) proposed a multifaceted toponym recognition and resolution approach by leveraging a wide range of methods and information resources including a dictionary of entity names and cue words, statistical methods such as POS tagging and NER, and rule-based toponym refactoring. Speriosu and Baldridge (2013) trained a toponym resolver using geotagged Wikipedia articles which associates geo- and non-geo-words with toponyms, and used the trained resolver to disambiguate place names based on the words in their surrounding contexts. Santos et al (2015) proposed a machine learning approach for place name disambiguation which combined multiple learning features such as the geospatial distances between candidates and other locations in a document and the textual context where the place references occur. Ju et al (2016) combined entity co-occurrence and topic modeling to identify various contextual clues (i.e., related entities and topical words) to enhance place name disambiguation. There are also many other place name disambiguation studies that focused on social media data (e.g., tweets) and leveraged social me-

dia specific features, such as social interactions, location consistency of users, and metadata fields associated with tweets (Zhang and Gelernter, 2014; Awamura et al, 2015; Di Rocco et al, 2016).

0.2.3 Developed geoparsers and tools

A number of software tools have been developed that can recognize and resolve toponyms from texts. This section provides a discussion on these tools and their advantages and limitations, with the goal of helping potential users choose the right tools for their applications. Our discussion is organized into two parts: general NER tools that can be used for identifying toponyms and specifically designed geoparsers.

General NER tools. Toponym recognition and resolution could be considered as a subtask of named entity recognition or word sense disambiguation. As a result, one way to extract place names from texts is to use existing NER tools developed from the computer science community and to keep only *locations* in the extracted entities. As discussed previously, Stanford NER is a tool that has been widely used for recognizing place names. It is based on CRF and implemented using Java (Finkel et al, 2005). While possessing the capability of recognizing toponyms not contained in gazetteers, Stanford NER does not geo-locate the identified place names to its corresponding geographic coordinates, since it is designed as a general NER tool. spaCy NER (`https://spacy.io/`) is an open source tool implemented in Python. Similar to Stanford NER, it can only recognize toponyms without being able to link toponyms with their coordinates. DBpedia Spotlight (Mendes et al, 2011; Daiber et al, 2013) and Open Calais (`http://www.opencalais.com`) are two general NER tools based on external knowledge bases (e.g., Wikipedia). A major disadvantage of them is that they can identify only those place names that are recorded in a knowledge base such as Wikipedia or a gazetteer. An advantage of DBpedia Spotlight, compared with Stanford NER, is that it links the recognized place names to the corresponding entities on DBpedia, which enables the geo-locating of these place names based on their geographic coordinates in DBpedia. Open Calais, however, does not provide such direct links for the recognized place names.

Geoparsers. There exist geoparsers specifically designed for the task of recognizing and resolving place names. Since Stanford NER already provides a strong tool for toponym recognition, many geoparsers were developed by integrating Stanford NER with a toponym resolution component. For example, Karimzadeh et al (2013) developed *GeoTxt*, a Web-based geoparsing tool, that leverages Stanford NER for toponym recognition, and used GeoNames and a set of heuristic rules for toponym resolution. DeLozier et al (2015) designed TopoCluster which is a geoparser that can perform geoparsing without using a gazetteer. They used Stanford NER to recognize toponyms from texts and then resolve toponyms based on the *geographic profiles* of words in the surrounding context. The geographic profile of a word is the spatial distribution of the word characterized by local spatial statistics, and DeLozier et al (2015) derived geographic profiles of words using a set of geotagged Wikipedia ar-

ticles. Cartographic Location And Vicinity INdexer (CLAVIN) is an open-source geoparser that employs both Stanford NER and Apache OpenNLP in its different implementations for toponym recognition, and utilizes a gazetteer and fuzzy search for toponym resolution. Some geoparsers were developed using their own approaches for toponym recognition. For example, the Edinburgh Geoparser is a geoparsing system developed by the Language Technology Group at Edinburgh University (Alex et al, 2015), which used a software package developed by the same group for toponym recognition. The toponym resolution step of the Edinburgh Geoparser can be based on different gazetteers, such as GeoNames and Unlock. There are also commercial geoparsers, such as Yahoo PlaceSpotter (`https://developer.yahoo.com/boss/geo/docs/PM_KeyConcepts.html`) and Geoparser.io (`https://geoparser.io/`), which often put constrains on the number of free API calls that can be requested.

Comparing the performances of geoparsers is often challenging, largely because of a lack of openly available and human annotated corpora (Monteiro et al, 2016; Gritta et al, 2018). Some researchers have made great efforts to alleviate this dearth of open data for testing and training geoparsers. Leidner (2008) contributed TR-CoNLL which is a human annotated news corpus consisting of about 1,000 international news articles from Reuters and about 6,000 toponyms. Lieberman et al (2010) shared a human annotated dataset called Local-Global Lexicon (LGL) corpus, which contains 588 news articles published by 78 local newspapers from highly ambiguous places, such as *Paris News* (Texas) and *Paris Beacon-News* (Illinois). Hu et al (2014) contributed a semi-automatically annotated corpus containing textual descriptions from city websites with two highly ambiguous place names in the U.S., namely *Washington* and *Greenville*. Gritta et al (2018) contributed WikToR which is a corpus of Wikipedia articles with ambiguous names, such as *Lima, Peru*, *Lima, Ohio*, and *Lima, Oklahoma*, automatically annotated by a Python script. Wallgrün et al (2018) published GeoCopora, a dataset of tweets manually annotated using a crowdsourcing approach based on Amazon's Mechanical Turk and further verified by experts. In addition to contemporary corpora, some historical datasets are also made available, such as *War Of The Rebellion* by DeLozier et al (2016). Finally, the ACE 2005 English SpatialML is an annotated news corpus shared on the Linguistic Data Consortium (Mani et al, 2008), but it charges a fee ($1,000) for non-members.

0.2.4 Location inference from language modeling

While geoparsers are effective in recognizing and geo-locating toponyms mentioned in texts, there are situations when place names are not explicitly mentioned in texts. A variety of language models have been developed for geo-referencing texts using all the terms present in a document rather than toponyms only (see Purves et al (2018), Ch. 4.6 for a comprehensive survey). Approaches vary from developing machine learning classifiers of document-level location based on word features (Wing and Baldridge, 2011; Adams and Janowicz, 2012) to creating more tailored lin-

guistic models that analyze spatial language (e.g., spatial prepositions, adjectives, and reference frames) in text in order to identify locations above and beyond place names (Tenbrink and Kuhn, 2011; Stock and Yousaf, 2018). The former often utilize simplistic spatial models, such as regions and geodesic grids, which allows us to train predictive classifiers relatively easily on large amounts of data (Roller et al, 2012; Wing and Baldridge, 2014; Han et al, 2014). When these classifiers are trained on words as features, they are usually single-language models; however, a Unicode character level classifier has been developed that is language independent (Adams and McKenzie, 2018). Linguistic models, in contrast, involve formalisms of spatial language that attempt to capture the semantics of spatial relations in natural language discourse. The developed linguistic models can potentially extract spatial information that is opaque to the other methods, but also make for a more onerous task when applied to big data. For example, one can differentiate between a *locatum* (an object in space) and a *relatum* (another object that the locatum is related to), which can be used by a reader in a (geo)spatial scene to orient and locate the elements described in texts (Bateman et al, 2007). Doing so in an automated manner requires a full NLP pipeline that can identify parts-of-speech and dependencies within the texts prior to the spatial analysis (Chen and Manning, 2014; Avvenuti et al, 2018). In addition, corpus linguistics research is also relevant to location inference. Lexical dialectology (the study of dialects through computational means) can be used to associate specific language features with places on the Earth, which in turn can be used to improve the models for geo-locating texts (Rahimi et al, 2017; Dunn, 2018).

Unlike the geoparsing tools based on toponym resolution that were described in the previous section, location inference from language modeling is still largely done on a bespoke basis in the context of individual research projects. Among the geoparsers listed in the previous section, only TopoCluster (DeLozier et al, 2015) utilizes language modeling as a significant component in the pipeline.

0.2.5 Summary

This section discusses the main methods and tools developed for harvesting big geospatial data from natural language texts. We started from geoparsing, one major approach that collects geospatial data by recognizing and resolving toponyms mentioned in texts. The geo-located toponyms can be used as a basis for geo-locating a whole document (Monteiro et al, 2016; Melo and Martins, 2017). It is necessary to differentiate geoparsing, i.e., the task of recognizing and resolving (potentially colloquial) toponyms from natural language texts, from geocoding in conventional GIS, i.e., the task of locating formatted addresses (e.g., door number with a street name) (Goldberg et al, 2008). Both are important in geographic information science. In addition to geoparsing, we also discussed the harvesting of geospatial data when toponyms are not explicitly mentioned in texts, through the use of language modeling via machine learning and linguistic approaches.

0.3 Applications of Geospatial Data Harvested from Texts

This section discusses some applications that leverage geospatial data harvested from natural language texts. We will start from understanding human experiences toward places, move to using near real-time data for situation awareness, and finally discuss extracting information about place relations in virtual or cognitive spaces.

0.3.1 Understanding places and human experiences

Space and place are two related, but differently conceived concepts in academic geography. Until recently, quantitative statistical analysis of geographic information focused almost exclusively on spatial analysis, while *place* has been a rich subject of academic study in human geography. Recently with the advent of more geographic user-generated content being posted online (a.k.a. volunteered geographic information or VGI), especially on social media, *place* has become a subject of increasing interest for those doing quantitative data-driven research (Elwood et al, 2012; Sui and DeLyser, 2012). In a phenomenological sense, *place* has often been described as *space* engendered with meaning through human experience (either direct or indirect) (Tuan, 1977). Large amounts of unstructured observations of people's experiences in text thus provide a new window to investigate this phenomenological perspective on place, in ways that were previously restricted to smaller scaled humanisitic inquiries. Multiple kinds of textual analysis have been used on this data to provide these sorts of insights. Keyword-based, topical, sentiment, and emotion analyses all provide different ways to generalize about multiple human experiences (cf. Mei et al (2006); Hollenstein and Purves (2010); Chon et al (2012); Adams and McKenzie (2013); Adams (2015); Ballatore and Adams (2015); Doytsher et al (2017)). Apart from providing better understanding of place in a generic sense, analysis of big-geo data to understand place has been used for a variety of applications, including tourism (Hao et al, 2010; Xiang et al, 2015; Rahmani et al, 2017; McKenzie and Adams, 2018), urban research (Cranshaw and Yano, 2010; Campagna, 2014; van Weerdenburg et al, 2019), political science (Bastos et al, 2014), public health (Ghosh and Guha, 2013), marketing (Caverlee et al, 2013), and sociolinguistic research (Eisenstein et al, 2010).

Another domain where place-based geospatial data harvested from texts is increasingly being used is the digital (geospatial) humanities (Bodenhamer et al, 2010). Geospatial information that is buried in massive collections in libraries and online has been seen as a goldmine for spatial historical and literary analysis (Gregory et al, 2015). Historical datasets pose unique challenges, however, as many geoparsing tools are built on gazetteers of modern place names, and therefore custom solutions are often required to automatically extract geographic information from historical texts (Rupp et al, 2013). In this context, historical gazetteers, such as Pleiades (`https://pleiades.stoa.org`) and World-Historical Gazetteer (`http://whgazetteer.org`), have been developed to provide services for

finding and using information related to ancient places. In addition to supporting direct analysis, geospatial data can be extracted from the various documents used in humanities to build spatial indices which provide an alternative way of exploring textual content from a geographic perspective (McCurley, 2001; Purves et al, 2007; Adams et al, 2015).

0.3.2 Situation awareness for emergency response

Emergency response applications usually need real-time data about the situations on the ground. A lot of such data comes in the form of natural language text. Examples include social media posts, short text messages, texts converted from phone calls (or voice messages), and news reports sent by the journalists at emergency scenes. After an emergency, information from different sources often flood into the emergency operations center, overwhelming first responders. Accordingly, automated methods and tools become very useful for extracting location information (e.g., who needs help at which location) from massive amounts of data.

Many studies have used geospatial data harvested from texts for emergency responses. Social media data, especially Twitter data, has been widely utilized by many researchers (Tsou, 2015; Haworth and Bruce, 2015). For example, De Longueville et al (2009) investigated the spatial, temporal, and social dynamics of tweets during a major forest fire in the South of France in 2009. Crooks et al (2013) examined the spatial and temporal characteristics of tweets after a 5.8 magnitude earthquake occurred on the East Coast of the US in 2011. Nagar et al (2014) used daily geotagged tweets in NYC to investigate the spatiotemporal tweeting behavior related to influenza-like illness (ILI). Although a small percentage of tweets are already geotagged (about 1-2%), it is estimated that more than 10% tweets contain place references in their texts (Wallgrün et al, 2018). Thus, researchers also focused on extracting place reference information from the textual content of tweets. For example, MacEachren et al (2011) developed SensePlace2, a visual analytics system that supports the space-time-theme exploration of Twitter data for situation awareness and crisis management. In SensePlace2, the researchers differentiated *tweets from* (i.e., geotagged location) and *tweets about* (i.e., the locations mentioned in tweet content). Gelernter and Balaji (2013) proposed an algorithm for extracting place names in various forms, such as abbreviated, misspelled, or highly localized names, from the content of tweets posted after the 2011 earthquake in Christchurch, New Zealand. Issa et al (2017) studied the spatial diffusion of tweets about flu in four different cities using both geotagged and non-geotagged tweets. In addition to social media, news articles were also used by researchers to understand the situations related to natural hazards. For example, Wang and Stewart (2015) examined the impact of Hurricane Sandy by extracting place names, timestamps, and emergency information (e.g., power failure) from the news texts.

To give an intuitive idea of using social media data for situation awareness, we show a possible graphic user interface (GUI) of an information system in Figure 3.2

based on a sample of tweets collected during Hurricane Irma in September 2017. In

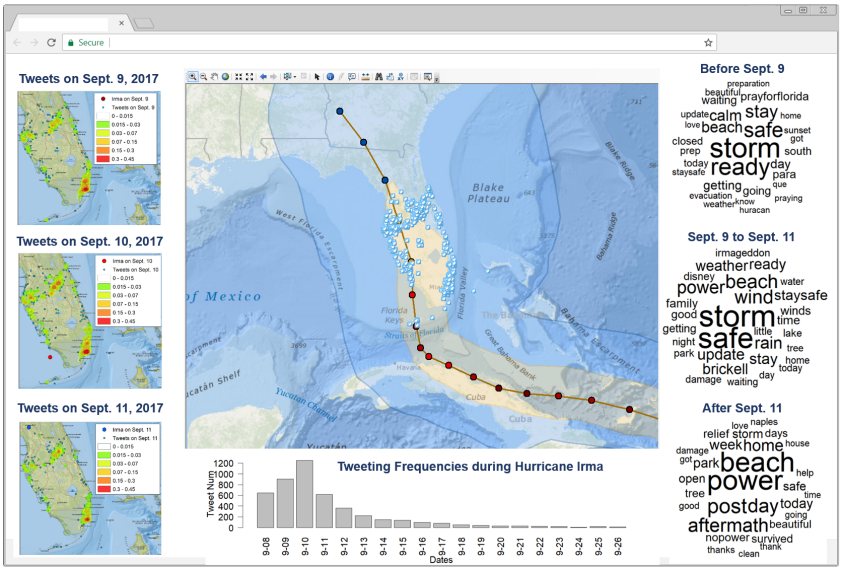


Fig. 0.2 A possible GUI of an information system for using the spatial, temporal, and textual information harvested from tweets for situation awareness using an example of Hurricane Irma.

this user interface, the main map shows the current and predicted trajectory of the hurricane and its impact area. The locations of geotagged tweets are visualized on the ground (one can also visualize the locations mentioned in the content of tweets using an approach by MacEachren et al (2011)). The bar chart at the bottom shows the tweeting intensities on different days. In the case of Hurricane Irma, most tweets were made between September 9th and 11th when Irma made Florida landfall and moved inland. On the left side of the interface, a user can pick three specific days and examine the intensities and geographic distributions of the tweets on those days. On the right side, three word clouds summarize the main topics of the tweets in three different time periods. In the case of Hurricane Irma, the tweets were summarized based on the periods of *before*, *during*, and *after* Irma. As can be seen, there were many words related to preparation and evacuation *before* the hurricane, and words about winds, rain, and trees were seen frequently *during* the event; and *after* the hurricane, the frequent words were about disaster damage and relief. Such information collected from social media and processed in a near real-time manner can help support the decision makings of emergency responders.

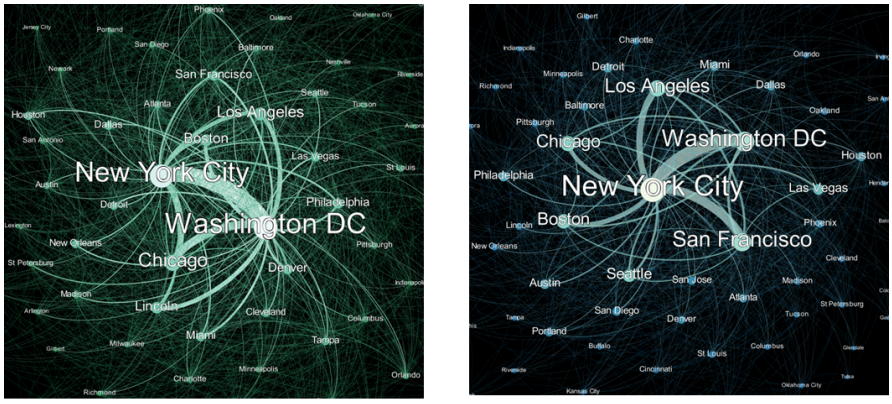
0.3.3 Place relations in virtual or cognitive space

Another special and valuable sort of geospatial information captured by texts is the relationships between places in virtual or cognitive space. Most traditional geographic datasets are organized based on *spatial proximity*. For example, we may have a dataset of land parcels located in the same geographic region. By contrast, texts, such as Web pages, social media posts, and news articles, can mention multiple places that are far apart and even in global scale, thereby relating these places together, often representing social, economic, and historical relationships that are non-spatially determined (Adams, 2018). Place name co-occurrences, thus, are often considered as evidence for these sorts of place relations (Hecht and Raubal, 2008; Twaroch et al, 2009; Ballatore et al, 2014; Liu et al, 2014; Spitz et al, 2016). Depending on application needs, different textual contexts, such as sentences, paragraphs, and even entire articles, can be used for determining place name co-occurrences. Place relations can also be established via hyperlinks, such as those in Wikipedia articles and other Web pages.

Places can be related together in texts for a variety of reasons. News articles can report different events that involve multiple places: a sports team may travel from their hometown to another city for a game; a company based in one country may establish a new branch office in another country (Toly et al, 2012; Sassen, 2016); a natural disaster, such as hurricane and flooding, can impact multiple cities and towns. In addition, Wikipedia pages and online blogs can discuss the similarities and dissimilarities of two places in terms of their climates, populations, geographic locations, and other aspects. In social media posts, people can talk and compare the life styles, food, and cultures in different places. In today's digital society empowered by information and communication technologies, a majority of places are interlinked together in the virtual or cyberspace, forming place networks (Taylor and Derudder, 2015; Shaw et al, 2016). As a result, big geospatial data harvested from natural language texts provide one important source for understanding the diverse and dynamic place relations in the virtual space, as well as the those perceived by people, i.e., the relations in cognitive space.

Many studies have examined place relations using different types of texts. Hecht and Moxley (2009) conducted an early study on place relations using hyperlinks in Wikipedia pages, and found that nearby places are more likely to have relations than distant ones, although places far away can still have relations. Liu et al (2014) examined place name co-occurrences in a set of news articles, and found that place relatedness in news articles has a weaker distance decay effect compared with those derived from human movements. Zhong et al (2017) also looked into place name co-occurrences in news articles, and concluded that places are more likely to be related if they are in the same administrative level or have a part-whole relation (e.g., Seattle is part of Washington State). Salvini and Fabrikant (2016) analyzed place name co-occurrences in Wikipedia pages and examined the semantics of place relations via the categories of Wikipedia pages. Also based on the co-occurrences of place names in Wikipedia articles, Spitz et al (2016) constructed toponym networks for place name disambiguation. Adams and Gahegan (2016) performed spatio-temporal

(*chronotopic*) analysis on Wikipedia corpus by analyzing the co-occurrences of places and times in texts to understand the intrinsic relations between place, space, and time in narrative texts. Hu et al (2017) examined place name co-occurrences in news articles, and employed a topic modeling approach to annotate the semantic topics of place relations. Figure 3.3 shows the relations of places extracted from a corpus of The Guardian newspapers under different semantic topics, as discussed in (Hu et al, 2017). As can be seen, places can have different strengths of relations under different semantic topics and thus different position prominence in the place networks: Washington DC plays a much more important role under the topic of *Politics* than under the topic of *Science and Technology*; by contrast, *San Francisco* has a largely increased prominence in the network under the topic of *Science and Technology* compared with its role under the topic of *Politics*.



Politics Science and Technology

Fig. 0.3 Relations of places under different semantic topics extracted from a corpus of news articles from The Guardian.

0.4 Summary and Future Directions

Geospatial data exist in various types of natural language texts, such as news articles, social media posts, Wikipedia pages, travel blogs, historical archives, housing advertisements, and so forth. Many of these data sources provide large amounts of data (e.g., millions or even billions of social media posts) which are constantly increasing as the time goes by. As a result, it becomes possible to harvest big geospatial data from natural language texts. Compared with the data from more conventional sources, such as the USGS and the US Census, geospatial data from texts capture valuable human experiences toward places, provide near real-time information after a disaster, and record place relations in virtual and cognitive spaces. In this chapter,

we discussed the methods and tools that can be used for harvesting geospatial data from texts. Geoparsing is a major approach that can extract structured geographic information from unstructured texts by recognizing and resolving the place names mentioned in texts. When toponyms are not explicitly contained in texts, other approaches based on language modeling can help us derive geographic information from texts.

A number of research directions can be pursued in the near future. For toponym recognition, the performances of existing approaches still vary depending on the tested datasets. Advancements in deep learning, such as bidirectional recurrent neural networks, can help increase the accuracy of recognizing place names from texts. New NLP methods may also help better identify the metonymies used in the texts. For toponym resolution, most approaches currently still resolve place names only to point-based locations, and there are rivers, countries, and other geographic features whose spatial footprints can be better represented as polylines, polygons, and even polyhedras (in a 3D space). In addition, although a number of geoparsers exist, it is difficult to directly compare the performances of these geoparsers. One reason is a lack of open and annotated corpora. Although researchers have started to address this issue in recent years, it still takes a considerable amount of time and effort to implement existing baselines and run them against common datasets. Thus, a benchmarking platform, such as EUPEG (Wang and Hu, 2019), could be helpful for comparing and evaluating geoparsers. From a perspective of applications, while this chapter has highlighted the use of geospatial data from texts in studies about place, digital humanities, situation awareness, and place relations, other applications are waiting to be explored and examined in the near future.

References

Adams B (2015) Finding similar places using the observation-to-generalization place model. Journal of Geographical Systems 17(2):137–156

Adams B (2018) From spatial representation to processes, relational networks, and thematic roles in geographic information retrieval. In: Proceedings of the 12th Workshop on Geographic Information Retrieval, ACM, New York, NY, USA, GIR'18, pp 1:1–1:2

Adams B, Gahegan M (2016) Exploratory chronotopic data analysis. In: International Conference on Geographic Information Science, Springer, pp 243–258

Adams B, Janowicz K (2012) On the geo-indicativeness of non-georeferenced text. In: Proceedings of the International Conference on Web and Social Media (ICWSM), AAAI Press, pp 375–378

Adams B, McKenzie G (2013) Inferring thematic places from spatially referenced natural language descriptions. In: Crowdsourcing geographic knowledge, Springer, pp 201–221

Adams B, McKenzie G (2018) Crowdsourcing the character of a place: Character-level convolutional networks for multilingual geographic text classification.

Transactions in GIS 22(2):394–408
Adams B, McKenzie G, Gahegan M (2015) Frankenplace: interactive thematic mapping for ad hoc exploratory search. In: Proceedings of the 24th international conference on world wide web, International World Wide Web Conferences Steering Committee, pp 12–22
Alex B, Byrne K, Grover C, Tobin R (2015) Adapting the Edinburgh geoparser for historical georeferencing. International Journal of Humanities and Arts Computing 9(1):15–35
Amitay E, Har'El N, Sivan R, Soffer A (2004) Web-a-where: geotagging web content. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 273–280
Avvenuti M, Cresci S, Nizzoli L, Tesconi M (2018) Gsp (geo-semantic-parsing): Geoparsing and geotagging with machine learning on top of linked data. In: European Semantic Web Conference, Springer, pp 17–32
Awamura T, Aramaki E, Kawahara D, Shibata T, Kurohashi S (2015) Location name disambiguation exploiting spatial proximity and temporal consistency. SocialNLP 2015@ NAACL pp 1–9
Ballatore A, Adams B (2015) Extracting place emotions from travel blogs. In: Proceedings of AGILE, vol 2015, pp 1–5
Ballatore A, Bertolotto M, Wilson DC (2014) An evaluative baseline for geo-semantic relatedness and similarity. GeoInformatica 18(4):747–767
Barbaresi A (2017) Towards a toolbox to map historical text collections. In: Proceedings of the 11th Workshop on Geographic Information Retrieval, ACM, p 5
Bastos MT, Recuero R, Zago G (2014) Taking tweets to the streets: A spatial analysis of the vinegar protests in brazil. First Monday 19(3)
Bateman J, Tenbrink T, Farrar S (2007) The role of conceptual and linguistic ontologies in interpreting spatial discourse. Discourse Processes 44(3):175–212
Bodenhamer DJ, Corrigan J, Harris TM (2010) The spatial humanities: GIS and the future of humanities scholarship. Indiana University Press
Buscaldi D, Rosso P (2008) A conceptual density-based approach for the disambiguation of toponyms. International Journal of Geographical Information Science 22(3):301–313
Campagna M (2014) The geographic turn in social media: opportunities for spatial planning and geodesign. In: International Conference on Computational Science and Its Applications, Springer, pp 598–610
Cataldi M, Ballatore A, Tiddi I, Aufaure MA (2013) Good location, terrible food: detecting feature sentiment in user-generated reviews. Social Network Analysis and Mining 3(4):1149–1163
Caverlee J, Cheng Z, Sui DZ, Kamath KY (2013) Towards geo-social intelligence: Mining, analyzing, and leveraging geospatial footprints in social media. IEEE Data Eng Bull 36(3):33–41
Chen D, Manning C (2014) A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 740–750

Chon Y, Lane ND, Li F, Cha H, Zhao F (2012) Automatically characterizing places with opportunistic crowdsensing using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM, pp 481–490

Cranshaw J, Yano T (2010) Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling. In: CSSWC Workshop at NIPS, vol 10

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

Daiber J, Jakob M, Hokamp C, Mendes PN (2013) Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems, ACM, pp 121–124

De Longueville B, Smith RS, Luraschi G (2009) Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In: Proceedings of the 2009 international workshop on location based social networks, ACM, pp 73–80

DeLozier G, Baldridge J, London L (2015) Gazetteer-independent toponym resolution using geographic word profiles. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI),, AAAI Press, pp 2382–2388

DeLozier G, Wing B, Baldridge J, Nesbit S (2016) Creating a novel geolocation corpus from historical texts. In: Proceedings of The 10th Linguistic Annotation Workshop, Association for Computational Linguistics, pp 188–198

Di Rocco L, Bertolotto M, Catania B, Guerrini G, Cosso T (2016) Extracting fine-grained implicit georeferencing information from microblogs exploiting crowd-sourced gazetteers and social interactions. In: AGILE International Conference on Geographic Information Science

Doytsher Y, Galon B, Kanza Y (2017) Emotion maps based on geotagged posts in the social media. In: Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities, ACM, pp 39–46

Dunn J (2018) Finding variants for construction-based dialectometry: A corpus-based approach to regional CxGs. Cognitive Linguistics 29(2):275–311

Eisenstein J, O'Connor B, Smith NA, Xing EP (2010) A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp 1277–1287

Elwood S, Goodchild MF, Sui DZ (2012) Researching volunteered geographic information: Spatial data, geographic research, and new social practice. Annals of the association of American geographers 102(3):571–590

Finkel JR, Grenager T, Manning C (2005) Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics,, Association for Computational Linguistics, pp 363–370

Freire N, Borbinha J, Calado P, Martins B (2011) A metadata geoparsing system for place name recognition and resolution in metadata records. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, ACM, pp 339–348

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

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

Gelernter J, Ganesh G, Krishnakumar H, Zhang W (2013) Automatic gazetteer enrichment with user-geocoded data. In: Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, ACM, pp 87–94

Ghosh D, Guha R (2013) What are we 'tweeting' about obesity? mapping tweets with topic modeling and geographic information system. Cartography and geographic information science 40(2):90–102

Goldberg DW, Wilson JP, Knoblock CA, Ritz B, Cockburn MG (2008) An effective and efficient approach for manually improving geocoded data. International journal of health geographics 7(1):60

Goodchild MF (2011) Formalizing place in geographic information systems. In: Communities, neighborhoods, and health, Springer, pp 21–33

Goodchild MF, Hill LL (2008) Introduction to digital gazetteer research. International Journal of Geographical Information Science 22(10):1039–1044

Gregory I, Donaldson C, Murrieta-Flores P, Rayson P (2015) Geoparsing, gis, and textual analysis: Current developments in spatial humanities research. International Journal of Humanities and Arts Computing 9(1):1–14

Gritta M, Pilehvar MT, Limsopatham N, Collier N (2018) What?s missing in geographical parsing? Language Resources and Evaluation 52(2):603–623

Grothe C, Schaab J (2009) Automated footprint generation from geotags with kernel density estimation and support vector machines. Spatial Cognition & Computation 9(3):195–211

Hahmann S, Purves R, Burghardt D (2014) Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. Journal of Spatial Information Science 2014(9):1–36

Han B, Cook P, Baldwin T (2014) Text-based twitter user geolocation prediction. Journal of Artificial Intelligence Research 49:451–500

Hao Q, Cai R, Wang C, Xiao R, Yang JM, Pang Y, Zhang L (2010) Equip tourists with knowledge mined from travelogues. In: Proceedings of the 19th international conference on World wide web, ACM, pp 401–410

Haworth B, Bruce E (2015) A review of volunteered geographic information for disaster management. Geography Compass 9(5):237–250

Hecht B, Moxley E (2009) Terabytes of tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge. In: International conference on spatial information theory, Springer, pp 88–105

Hecht B, Raubal M (2008) Geosr: Geographically explore semantic relations in world knowledge. The European Information Society pp 95–113

Hecht BJ, Stephens M (2014) A tale of cities: Urban biases in volunteered geographic information. ICWSM 14:197–205

Hill LL (2000) Core elements of digital gazetteers: placenames, categories, and footprints. In: International Conference on Theory and Practice of Digital Libraries, Springer, pp 280–290

Hollenstein L, Purves R (2010) Exploring place through user-generated content: Using flickr tags to describe city cores. Journal of Spatial Information Science 2010(1):21–48

Hu Y, Janowicz K, Prasad S (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, ACM, pp 1–8

Hu Y, Ye X, Shaw SL (2017) Extracting and analyzing semantic relatedness between cities using news articles. International Journal of Geographical Information Science 31(12):2427–2451

Hu Y, Mao H, McKenzie G (2018) A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. International Journal of Geographical Information Science pp 1–25

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

Inkpen D, Liu J, Farzindar A, Kazemi F, Ghazi D (2015) Location detection and disambiguation from twitter messages. Journal of Intelligent Information Systems pp 1–17

Intagorn S, Lerman K (2011) Learning boundaries of vague places from noisy annotations. In: Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems, ACM, pp 425–428

Issa E, Tsou MH, Nara A, Spitzberg B (2017) Understanding the spatio-temporal characteristics of twitter data with geotagged and non-geotagged content: two case studies with the topic of flu and ted (movie). Annals of GIS 23(3):219–235

Janowicz K, Keßler C (2008) The role of ontology in improving gazetteer interaction. International Journal of Geographical Information Science 22(10):1129–1157

Jiang Y, Li Z, Ye X (2018) Understanding demographic and socioeconomic biases of geotagged twitter users at the county level. Cartography and Geographic Information Science pp 1–15

Jones CB, Purves RS (2008) Geographical information retrieval. International Journal of Geographical Information Science 22(3):219–228

Jones CB, Purves RS, Clough PD, Joho H (2008) Modelling vague places with knowledge from the web. International Journal of Geographical Information Science 22(10):1045–1065

Ju Y, Adams B, Janowicz K, Hu Y, Yan B, McKenzie G (2016) Things and strings: Improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In: 20th International Conference on Knowledge Engineering and Knowledge Management, Springer

Karimzadeh M, Huang W, Banerjee S, Wallgrün JO, Hardisty F, Pezanowski S, Mitra P, MacEachren AM (2013) Geotxt: a web api to leverage place references

in text. In: Proceedings of the 7th workshop on geographic information retrieval, ACM, pp 72–73

Keßler C, Janowicz K, Bishr M (2009a) An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, pp 91–100

Keßler C, Maué P, Heuer J, Bartoschek T (2009b) Bottom-up gazetteers: Learning from the implicit semantics of geotags. GeoSpatial semantics pp 83–102

Ladra S, Luaces MR, Pedreira O, Seco D (2008) A toponym resolution service following the ogc wps standard. In: International Symposium on Web and Wireless Geographical Information Systems, Springer, pp 75–85

Leidner JL (2008) Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. Universal-Publishers

Li H, Srihari RK, Niu C, Li W (2002) Location normalization for information extraction. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics, pp 1–7

Li L, Goodchild MF (2012) Constructing places from spatial footprints. In: Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information, ACM, pp 15–21

Lieberman MD, Samet H (2011) Multifaceted toponym recognition for streaming news. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, pp 843–852

Lieberman MD, Samet H, Sankaranarayanan J (2010) Geotagging with local lexicons to build indexes for textually-specified spatial data. In: Data Engineering (ICDE), 2010 IEEE 26th International Conference on, IEEE, pp 201–212

Liu Y, Wang F, Kang C, Gao Y, Lu Y (2014) Analyzing relatedness by toponym co-occurrences on web pages. Transactions in GIS 18(1):89–107

MacEachren AM, Jaiswal A, Robinson AC, Pezanowski S, Savelyev A, Mitra P, Zhang X, Blanford J (2011) Senseplace2: Geotwitter analytics support for situational awareness. In: Visual analytics science and technology (VAST), 2011 IEEE conference on, IEEE, pp 181–190

Madden DJ (2017) Pushed off the map: Toponymy and the politics of place in new york city. Urban Studies p Online First

Malik MM, Lamba H, Nakos C, Pfeffer J (2015) Population bias in geotagged tweets. People 1(3,759.710):3–759

Mani I, Hitzeman J, Richer J, Harris D (2008) ACE 2005 english spatialML annotations. Linguistic Data Consortium, Philadelphia

McCurley KS (2001) Geospatial mapping and navigation of the web. In: Proceedings of the 10th international conference on World Wide Web, ACM, pp 221–229

McKenzie G, Adams B (2018) A data-driven approach to exploring similarities of tourist attractions through online reviews. Journal of Location Based Services 12(2):94–118

McKenzie G, Liu Z, Hu Y, Lee M (2018) Identifying urban neighborhood names through user-contributed online property listings. ISPRS International Journal of Geo-Information 7(10):388

Mei Q, Liu C, Su H, Zhai C (2006) A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proceedings of the 15th international conference on World Wide Web, ACM, pp 533–542

Melo F, Martins B (2017) Automated geocoding of textual documents: A survey of current approaches. Transactions in GIS 21(1):3–38

Mendes PN, Jakob M, García-Silva A, Bizer C (2011) Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems, ACM, pp 1–8

Merschdorf H, Blaschke T (2018) Revisiting the role of place in geographic information science. ISPRS International Journal of Geo-Information 7(9):364

Monteiro BR, Davis Jr CA, Fonseca F (2016) A survey on the geographic scope of textual documents. Computers & Geosciences 96:23–34

Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, Brownstein JS (2014) A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. Journal of medical Internet research 16(10)

Nesi P, Pantaleo G, Tenti M (2016) Geographical localization of web domains and organization addresses recognition by employing natural language processing, pattern matching and clustering. Engineering Applications of Artificial Intelligence 51:202–211

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

Purves RS, Clough P, Jones CB, Arampatzis A, Bucher B, Finch D, Fu G, Joho H, Syed AK, Vaid S, 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):717–745

Purves RS, Clough P, Jones CB, Hall MH, Murdock V, et al (2018) Geographic information retrieval: Progress and challenges in spatial search of text. Foundations and Trends® in Information Retrieval 12(2-3):164–318

Rahimi A, Cohn T, Baldwin T (2017) A neural model for user geolocation and lexical dialectology. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol 2, pp 209–216

Rahmani K, Gnoth J, Mather D (2017) Tourists' participation on web 2.0: A corpus linguistic analysis of experiences. Journal of Travel Research p 0047287517732425

Roller S, Speriosu M, Rallapalli S, Wing B, Baldridge J (2012) Supervised text-based geolocation using language models on an adaptive grid. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, pp 1500–1510

Rupp C, Rayson P, Baron A, Donaldson C, Gregory I, Hardie A, Murrieta-Flores P (2013) Customising geoparsing and georeferencing for historical texts. In: Big Data, 2013 IEEE International Conference on, IEEE, pp 59–62

Salvini MM, Fabrikant SI (2016) Spatialization of user-generated content to uncover the multirelational world city network. Environment and Planning B: Planning and Design 43(1):228–248

Santos J, Anastácio I, Martins B (2015) Using machine learning methods for disambiguating place references in textual documents. GeoJournal 80(3):375–392

Sassen S (2016) The global city: Strategic site, new frontier. In: Managing Urban Futures, Routledge, pp 89–104

Shaw SL, Tsou MH, Ye X (2016) Human dynamics in the mobile and big data era. International Journal of Geographical Information Science 30(9):1687–1693

Southall H (2014) Rebuilding the great britain historical gis, part 3: integrating qualitative content for a sense of place. Historical Methods: A Journal of Quantitative and Interdisciplinary History 47(1):31–44

Speriosu M, Baldridge J (2013) Text-driven toponym resolution using indirect supervision. In: ACL (1), ACL, pp 1466–1476

Spitz A, Geiß J, Gertz M (2016) So far away and yet so close: augmenting toponym disambiguation and similarity with text-based networks. In: Proceedings of the third international ACM SIGMOD workshop on managing and mining enriched geo-spatial data, ACM, p 2

Stock K, Yousaf J (2018) Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. International Journal of Geographical Information Science 32(6):1087–1116, DOI 10.1080/13658816.2018.1432861

Sui D, DeLyser D (2012) Crossing the qualitative-quantitative chasm i: Hybrid geographies, the spatial turn, and volunteered geographic information (vgi). Progress in Human Geography 36(1):111–124

Taylor PJ, Derudder B (2015) World city network: a global urban analysis. Routledge

Tenbrink T, Kuhn W (2011) A model of spatial reference frames in language. In: Egenhofer M, Giudice N, Moratz R, Worboys M (eds) Spatial Information Theory, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 371–390

Toly N, Bouteligier S, Smith G, Gibson B (2012) New maps, new questions: global cities beyond the advanced producer and financial services sector. Globalizations 9(2):289–306

Tsou MH (2015) Research challenges and opportunities in mapping social media and big data. Cartography and Geographic Information Science 42(sup1):70–74

Tuan YF (1977) Space and place: The perspective of experience. U of Minnesota Press

Twaroch FA, Jones CB (2010) A web platform for the evaluation of vernacular place names in automatically constructed gazetteers. In: Proceedings of the 6th Workshop on Geographic Information Retrieval, ACM, p 14

Twaroch FA, Jones CB, Abdelmoty AI (2009) Acquisition of vernacular place names from web sources. In: King I, Baeza-Yates R (eds) Weaving Services and People on the World Wide Web, Springer, pp 195–214

Wallgrün JO, Karimzadeh M, MacEachren AM, Pezanowski S (2018) Geocorpora: building a corpus to test and train microblog geoparsers. International Journal of Geographical Information Science 32(1):1–29

Wang J, Hu Y (2019) Enhancing spatial and textual analysis with eupeg: An extensible and unified platform for evaluating geoparsers. Transactions in GIS

Wang M, Zhou X (2016) Geography matters in online hotel reviews. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences pp 573–576

Wang W, Stewart K (2015) Spatiotemporal and semantic information extraction from web news reports about natural hazards. Computers, environment and urban systems 50:30–40

van Weerdenburg D, Scheider S, Adams B, Spierings B, van der Zee E (2019) Where to go and what to do: Extracting leisure activity potentials from web data on urban space. Computers, Environment and Urban Systems 73:143–156

Wing B, Baldridge J (2014) Hierarchical discriminative classification for text-based geolocation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 336–348

Wing BP, Baldridge J (2011) Simple supervised document geolocation with geodesic grids. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, pp 955–964

Woodruff AG, Plaunt C (1994) Gipsy: Automated geographic indexing of text documents. Journal of the American Society for Information Science 45(9):645–655

Xiang Z, Schwartz Z, Gerdes Jr JH, Uysal M (2015) What can big data and text analytics tell us about hotel guest experience and satisfaction? International Journal of Hospitality Management 44:120–130

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

Zhong X, Liu J, Gao Y, Wu L (2017) Analysis of co-occurrence toponyms in web pages based on complex networks. Physica A: Statistical Mechanics and its Applications 466:462–475