Geo-Text Data and Data-Driven Geospatial Semantics

Yingjie Hu

GSDA Lab, Department of Geography, University of Tennessee, Knoxville, TN, 37996, USA

Abstract

Many datasets nowadays contain links between geographic locations and natural language texts. These links can be geotags, such as geotagged tweets or geotagged Wikipedia pages, in which location coordinates are explicitly attached to texts. These links can also be place mentions, such as those in news articles, travel blogs, or historical archives, in which texts are implicitly connected to the mentioned places. This kind of data is referred to as *geo-text data*. The availability of large amounts of geo-text data brings both challenges and opportunities. On the one hand, it is challenging to automatically process this kind of data due to the unstructured texts and the complex spatial footprints of some places. On the other hand, geo-text data offers unique research opportunities through the rich information contained in texts and the special links between texts and geography. As a result, geo-text data facilitates various studies especially those in data-driven geospatial semantics. This paper discusses geo-text data and related concepts. With a focus on data-driven research, this paper systematically reviews a large number of studies that have discovered multiple types of knowledge from geo-text data. Based on the literature review, a generalized workflow is extracted and key challenges for future work are discussed.

Keywords: geo-text data, spatial analysis, natural language processing, spatial and textual data analysis, data-driven geospatial semantics, spatial data science.

1. Introduction

Recent years have witnessed an unprecedented increase in the volume, variety, and velocity of data from different sources (Miller and Goodchild, 2015). Thanks to the advancements in sensors and information technologies, authoritative organizations, such as the U.S. Geological Survey, are continuing to produce many datasets with often richer content and higher precision. Meanwhile, general individuals, with the support of GPS-enabled smart devices, are also contributing large amounts of data via social media platforms, online blogs, review websites, and others (Goodchild, 2007; Haklay et al., 2008). As a result, various types of datasets have been generated.

Among these datasets, there is one kind that contains interesting links between geographic locations and natural language texts. Some of these links are geotags, such as geotagged

Email address: `yjhu.geo@gmail.com` (Yingjie Hu)

How to cite this article: Hu Y. Geo-text data and data-driven geospatial semantics. Geography Compass. 2018, e12404. `https://doi.org/10.1111/gec3.12404`

tweets or geotagged Wikipedia pages, in which location coordinates are directly attached to texts. Some other links are in the form of place mentions, such as those in news articles or travel blogs, in which the texts mention one or multiple place names. This kind of data is referred to as *geo-text data* in this paper.

To a certain degree, geo-text data already exists in many GIS applications. In a vector crime map, the locations of crimes can be linked to crime types, such as "Property Crime" and "Violent Crime", which are represented as text strings. In a raster land-use map, pixels representing geographic locations are linked to land-use categories, such as "Commercial" and "Agricultural". Although these crime types and land-use categories are texts, they are pre-defined by a schema and can only be chosen from a set of pre-defined text strings. This type of texts with a pre-defined schema is considered as *structured data*. By contrast, the texts in geo-text data (e.g., geotagged tweets) can be composed freely, and should be considered as *unstructured data*. Structured data can be handled relatively easily, since they can be converted to numeric indices based on the related schema. It is more challenging to process unstructured data due to language flexibility and text ambiguity. Despite these challenges, natural language texts offer rich information, such as keywords, topics, entities, and sentiments, and enable various new research when linked to geographic locations.

Geo-text data has been used in many empirical studies. For example, geotagged tweets were employed for supporting disaster response (Huang and Xiao, 2015; Zhang et al., 2015) and investigating public health issues (Widener and Li, 2014; Stefanidis et al., 2017). Geotagged Flickr photos were used to enrich gazetteers (Keßler et al., 2009) and to represent vague places (Hollenstein and Purves, 2010). Natural language descriptions about landscapes were collected to understand the perceptions of people and their sense of place (Mark et al., 2011; Wartmann et al., 2018). Meanwhile, methods and tools were developed in geographic information retrieval (GIR) for recognizing and geo-locating place names from texts (Jones and Purves, 2008). Examples of such tools, also called *geoparsers*, include MetaCarta (Frank et al., 2006), GeoTxt (Karimzadeh et al., 2013), Edinburgh Geoparser (Alex et al., 2015), and TopoCluster (DeLozier et al., 2015). Open and labeled datasets, such as the corpora annotated using SpatialML (Mani et al., 2010), WikToR (Gritta et al., 2017), and GeoCorpora (Wallgrün et al., 2018), were made available for training and testing geoparsers. While many empirical studies exist, there lacks a systematic discussion on the concept of geo-text data, its formal representation, and the types of knowledge that can be extracted. This paper fills such a gap by bringing together the insights from different studies and organizing them using a coherent framework. Specifically, this paper makes the following contributions:

- A formal representation of geo-text data based on a general GIS theory.
- A systematic discussion on the knowledge that can be extracted from geo-text data.
- A generalized workflow for processing and analyzing geo-text data.
- A set of key challenges for future research based on geo-text data.

The remainder of this paper is organized using a series of questions. Section 2 addresses the question "*what is geo-text data?*" by presenting a formalization of geo-text data and discussing related concepts. Section 3 examines the question "*what can we get from geo-text data?*" by reviewing a large number of studies that have discovered various types of knowledge from geo-text data. Section 4 answers the question "*what is a general workflow that we can follow to analyze geo-text data?*" by presenting a workflow generalized from the literature. Section 5 addresses the question "*what are some key challenges for future research?*". Finally, Section 6 summarizes this work.

2. Geo-text data

Geo-text data can be considered as a collective term that encompasses many specific types of data, such as geotagged social media, geotagged Wikipedia pages, news articles, historical archives, location-focused online reviews, geotagged housing posts, and others that contain links between locations and texts. While these different types of data have been used in separate studies, they share similar characteristics and can be processed and analyzed using similar methods. This paper abstracts from the specific studies and data formats, and identifies the core concepts and methods that can be applied to geo-text data in general. The identified knowledge could be integrated to educational programs on *spatial data science.*

Based on a general theory of geographic representation in GIS (Goodchild et al., 2007), geo-text data can be formalized as Equation 1.

$$\langle f, [t], W \rangle \tag{1}$$

where:

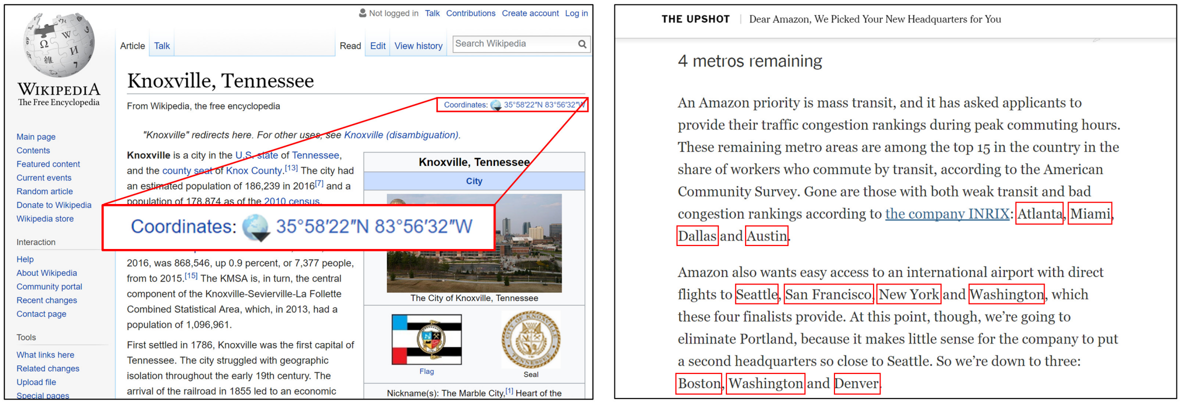
- f is the geographic location, or *spatial footprint*, of geo-text data. A common spatial footprint is a point defined by a pair of coordinates, such as the location of a geotagged tweet. However, f can also be lines (e.g., roads and rivers), polygons (e.g., cities and states), and polyhedras (e.g., a 3D geographic feature). In addition, f can be vague places, cognitive regions, or other forms that lack crisp boundaries (Montello et al., 2003; Jones et al., 2008; Montello et al., 2014).

- t is the timestamp associated with the data record. t is optional and therefore is surrounded by the square brackets in Equation 1. Many geo-text data have timestamps, such as the posting time of a tweet, the publishing time of a news article, and the editing time of a Wikipedia page. Such time information can enable valuable time series analysis.

- W is the textual content of geo-text data. W can be modeled using a simple bag-of-words model, while more sophisticated methods, such as parse trees (Schuster and Manning, 2016), can be employed to provide more accurate modeling of texts. Depending on the specific needs of an application, different types of information can be extracted from W, such as entities (e.g., persons, places, and events), topics (e.g., music and travel), and sentiments (e.g., happy and sad).

Compared with typical GIS data, such as temperature measurements and digital elevation models (DEM), in which numeric values are attached to locations, geo-text data can have a sentence, a paragraph, or even an entire article attached to a location. Geo-text data also includes Flickr photo tags (Hollenstein and Purves, 2010; Tardy et al., 2016), geotagged surnames of people (Longley et al., 2011; Cheshire and Longley, 2012), street or mountain names (Hill, 2000; Alderman, 2016), and others, in which words and phrases, instead of complete sentences, are attached to locations.

Depending on how locations are linked to texts, we can differentiate *explicit* and *implicit* geo-text data. *Explicit* geo-text data have spatial footprints explicitly attached to texts, such as geotagged Wikipedia articles (Figure 1(a)), while *implicit* geo-text data do not have explicit spatial footprints, but mention place names in their texts, such as the news article in Figure 1(b). Implicit geo-text data can be converted to explicit data through geoparsing (Gregory et al., 2015).



Knoxville, Tennessee

From Wikipedia, the free encyclopedia

Coordinates: 35°58′22″N 83°56′32″W

"Knoxville" redirects here. For other uses, see Knoxville (disambiguation).

Knoxville is a city in the U.S. state of Tennessee, and the county seat of Knox County.[13] The city had an estimated population of 186,239 in 2016[7] and a population of 178,874 as of the 2010 census.

Coordinates: 35°58′22″N 83°56′32″W

2016, was 868,546, up 0.9 percent, or 7,377 people, from to 2015.[15] The KMSA is, in turn, the central component of the Knoxville-Sevierville-La Follette Combined Statistical Area, which, in 2013, had a population of 1,096,961.

First settled in 1786, Knoxville was the first capital of Tennessee. The city struggled with geographic isolation throughout the early 19th century. The arrival of the railroad in 1855 led to an economic

Knoxville, Tennessee

City

The City of Knoxville, Tennessee

Flag

Seal

Nickname(s): The Marble City,[1] Heart of the

(a)

THE UPSHOT | Dear Amazon, We Picked Your New Headquarters for You

4 metros remaining

An Amazon priority is mass transit, and it has asked applicants to provide their traffic congestion rankings during peak commuting hours. These remaining metro areas are among the top 15 in the country in the share of workers who commute by transit, according to the American Community Survey. Gone are those with both weak transit and bad congestion rankings according to the company INRIX: Atlanta, Miami, Dallas and Austin.

Amazon also wants easy access to an international airport with direct flights to Seattle, San Francisco, New York and Washington, which these four finalists provide. At this point, though, we're going to eliminate Portland, because it makes little sense for the company to put a second headquarters so close to Seattle. So we're down to three: Boston, Washington and Denver.

(b)

Figure 1: Explicit and implicit geo-text data: (a) a geotagged Wikipedia page; (b) a news article containing city names.

Two major processes can generate geo-text data, each of which can generate both explicit and implicit geo-text data. In the first process, people learn about a place and express words related to it (Figure 2(a)). For example, an Instagram user may take a photo at a location and describe what he sees (which generates explicit geo-text data); or a travel blog writer may write her experience after a day of touring in a foreign city (which generates implicit data). In the second process, people express thoughts and opinions which are not directly related to their current locations (Figure 2(b)). For example, a Twitter user may post a tweet irrelevant to his location (explicit geo-text data are generated); or a journalist in her office at Washington D.C. may write a news article about an event happened in Las Vegas (implicit data are generated). The two processes also suggest that the text in a geo-text data record can be linked to two locations: one *about* location and one *from* location. These two locations can be the same, partially overlap, or completely different. Depending on the application needs, we may choose one location over the other or use both simultaneously. In a study on geotagged Twitter data, MacEachren et al. (2011) made a distinction between tweets *about* and *from* locations. The two processes discussed here are similar to their distinction, but are generalized to geo-text data beyond only geotagged tweets.

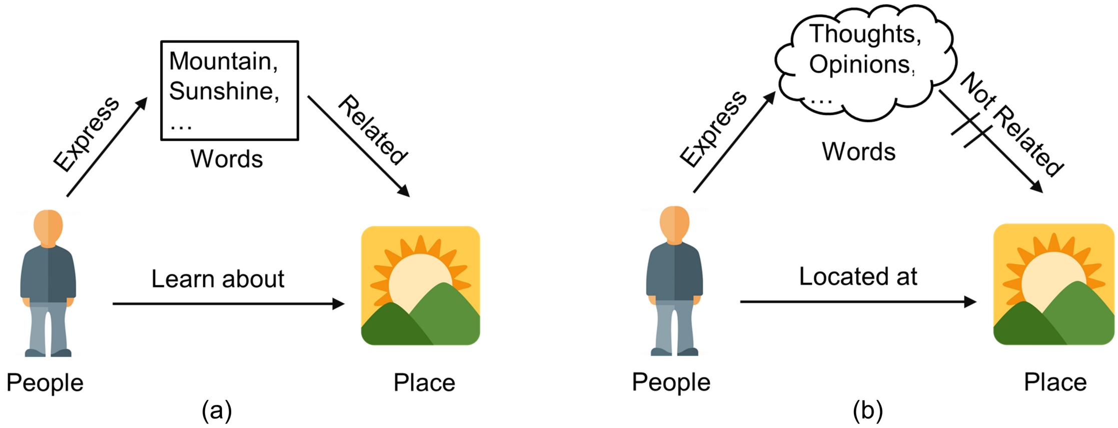


Figure 2: Two major processes that generate geo-text data: (a) people learn about a place and express words related to this place; (b) people are located at a place and express words not directly related to this place.

In summary, this section addresses the question "*what is geo-text data?*" by providing a formal representation and discussing related concepts. Geo-text data can be considered as a collective term that encompasses various types of data that contain links between locations and texts. This section discusses the spatial footprints, timestamps, and texts of geo-text data, and compare it with typical GIS data. Explicit and implicit geo-text data are differentiated, and two major processes that generate geo-text data are discussed.

3. Knowledge discovery from geo-text data and data-driven geospatial semantics

The large volume and rich variety of geo-text data enable the discovery of knowledge that can support disaster response, urban planning, transportation management, and many other applications. Particularly, geo-text data facilitates research in *data-driven geospatial semantics*, a bottom-up approach for studying the meaning of geographic features and terms. *Data-driven geospatial semantics* can be distinguished from the *expert-driven* or top-down approach (Kuhn, 2005; Hu, 2018). Consider measuring the semantic similarity between two words, "road" and "street". We can take an *expert-driven* approach by inviting a group of experts to assign scores between 0 and 1 and then taking an average. Alternatively, we can use a *data-driven* approach by harvesting millions of Web pages that contain either "road" or "street" and measuring their semantic similarity using context words. Both approaches have their pros and cons, and can sometimes be combined (Hu and Janowicz, 2016). This section focuses on data-driven research. It reviews a large number of studies and organizes them based on the types of knowledge discovered from geo-text data.

3.1. Place names

Extracting place names, or *toponyms*, from texts is a topic frequently studied in GIR (Jones and Purves, 2008; Wing and Baldridge, 2011; Vasardani et al., 2013; Gelernter and Balaji, 2013; Li et al., 2014; Laurini, 2015; Nesi et al., 2016). This process is often referred to as *geoparsing* which involves two main steps: toponym recognition and toponym resolution. In toponym recognition, the goal is to identify the words and phrases that can represent place names. Gazetteers (Lieberman and Samet, 2011; Zhang and Gelernter, 2014) and linguistic features (Freire et al., 2011; Inkpen et al., 2017) are often utilized for this step. In toponym resolution, the goal is to disambiguate and geo-locate the identified place names. Place name disambiguation is necessary due to both geo/geo ambiguity (i.e., the same term, such as *London*, can refer to different places) and geo/non-geo ambiguity (i.e., the same term, such as *Washington*, can refer to both places and persons) (Amitay et al., 2004; Leidner, 2008). Methods, such as co-occurrences (Overell and Rüger, 2008), conceptual density (Buscaldi and Rosso, 2008), and topic modeling (Ju et al., 2016), were proposed for place name disambiguation.

Geo-text data can be used for identifying the spatial footprints of place names as well. Keßler et al. (2009) studied vague place names, such as "Soho", using geotagged photo data from Flickr, Panoramio and Picasa, and developed a clustering method based on Delaunay triangulation to construct their spatial footprints. Hollenstein and Purves (2010) and Li and Goodchild (2012) used geotagged Flickr photos to derive the spatial footprints of city names using kernel density estimation (KDE). Jones et al. (2008) used a search engine to harvest Web pages that contain a target vague place name (e.g., "Mid-Wales"), extracted and geo-located the place names contained in these Web pages, and delineated the spatial footprint of the target place name using KDE. Table 1 summarizes the studies discussed above. Good reviews focusing on the topic of geoparsing and GIR can also be found in Monteiro et al. (2016), Melo and Martins (2017), and Purves et al. (2018).

Table 1: Summary of the discussed studies on extracting place names and their spatial footprints.

Study	Main Task	Methods

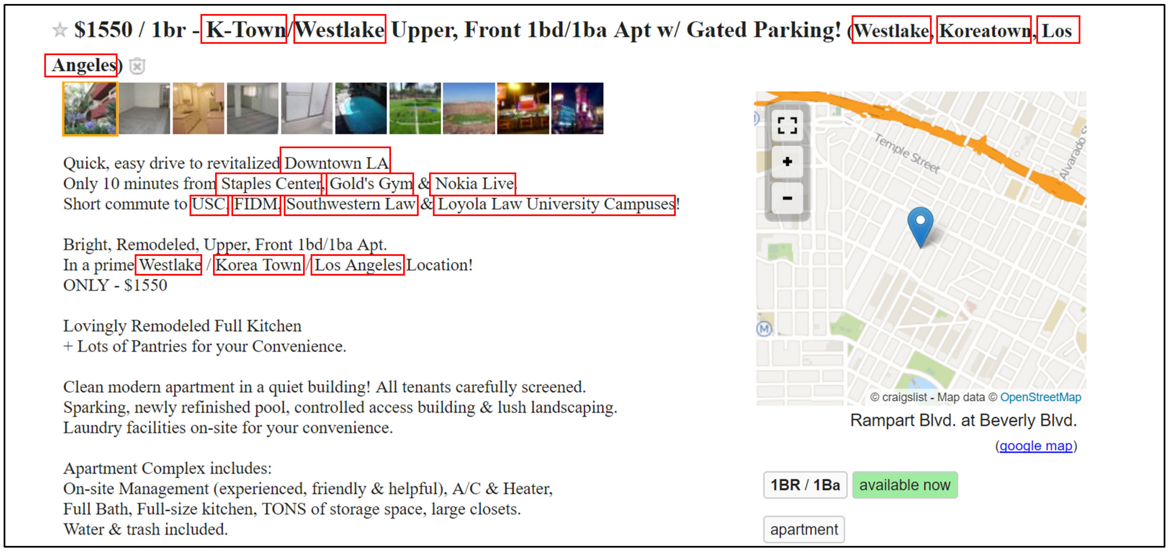
Lieberman and Samet (2011)
Freire et al. (2011)
Zhang and Gelernter (2014)
Inkpen et al. (2017)
	Place name recognition and resolution	Digital gazetteers;
		Linguistic features;
		Machine learning models

Overell and Rüger (2008)
Buscaldi and Rosso (2008)
Ju et al. (2016)
	Place name disambiguation	Co-occurrence models;
		Conceptual density;
		Topic modeling

Keßler et al. (2009) Hollenstein and Purves (2010) Li and Goodchild (2012) Jones et al. (2008)	Delineating spatial footprints of (vague) place names	Kernel density estimation; Delaunay triangulation

Geotagged housing posts are a type of geo-text data that have been rarely studied. They often contain local place names that can be extracted for enriching gazetteers. Figure 3 shows a geotagged housing post published on a local-oriented website. Such a post contains local place names, such as "K-Town" and "USC", and the location of the advertised property.

One can design methods to extract these local place names and their spatial footprints from these geotagged housing posts.



$1550 / 1br - K-Town/Westlake Upper, Front 1bd/1ba Apt w/ Gated Parking! (Westlake, Koreatown, Los Angeles)

Quick, easy drive to revitalized Downtown LA
Only 10 minutes from Staples Center, Gold's Gym & Nokia Live
Short commute to USC, FIDM, Southwestern Law & Loyola Law University Campuses!

Bright, Remodeled, Upper, Front 1bd/1ba Apt.
In a prime Westlake / Korea Town / Los Angeles Location!
ONLY - $1550

Lovingly Remodeled Full Kitchen
+ Lots of Pantries for your Convenience.

Clean modern apartment in a quiet building! All tenants carefully screened.
Sparking, newly refinished pool, controlled access building & lush landscaping.
Laundry facilities on-site for your convenience.

Apartment Complex includes:
On-site Management (experienced, friendly & helpful), A/C & Heater,
Full Bath, Full-size kitchen, TONS of storage space, large closets.
Water & trash included.

Rampart Blvd. at Beverly Blvd.
(google map)

1BR / 1Ba
available now
apartment

Figure 3: A geotagged housing post published on a local-oriented website in Los Angeles, USA.

3.2. *Place relations/sequences*

The ability of extracting place names from texts enables further examinations on place relations/sequences based on geo-text data. Such examinations can reveal interesting and sometimes intangible connections among places. Two places can be considered as related if they co-occur in texts or if there exists Web links between them (Ballatore et al., 2014; Liu et al., 2014; Spitz et al., 2016). Hecht and Moxley (2009) conducted an early study using the hyperlinks among geotagged Wikipedia pages to empirically verify Tobler's First Law. They found that nearby places are indeed more likely to have relations than distant ones, although places far away may still have relations. Salvini and Fabrikant (2016) analyzed place name co-occurrences in Wikipedia pages, and used the categories of Wikipedia pages to annotate the semantics of place relations. Adams and Gahegan (2016) performed chronotopic analysis on Wikipedia corpus by analyzing the co-occurrences of places and times in texts. Using news articles, Liu et al. (2014) examined place name co-occurrences and found that place relatedness in news articles decreases less rapidly with the increase of distance, compared with the results from human movement analysis. Hu et al. (2017) took a topic modeling approach to understand the semantics of place relations using news articles, and found that geographic distance has a non-uniform impact on place relatedness under different topics. City network research also analyzed place name co-occurrences in news articles often based on small data samples and using manual content analysis (Taylor, 1997; Beaverstock et al., 2000). Place relations can be visualized on maps (Figure 4 shows a simple example), and further analysis can be performed based on these relations.

Place sequences can also be extracted from geo-text data. For example, travel trajectories can be extracted from travel blogs, while life trajectories can be derived from biographic documents, such as one's born, study, marriage, and work places. A related research is from

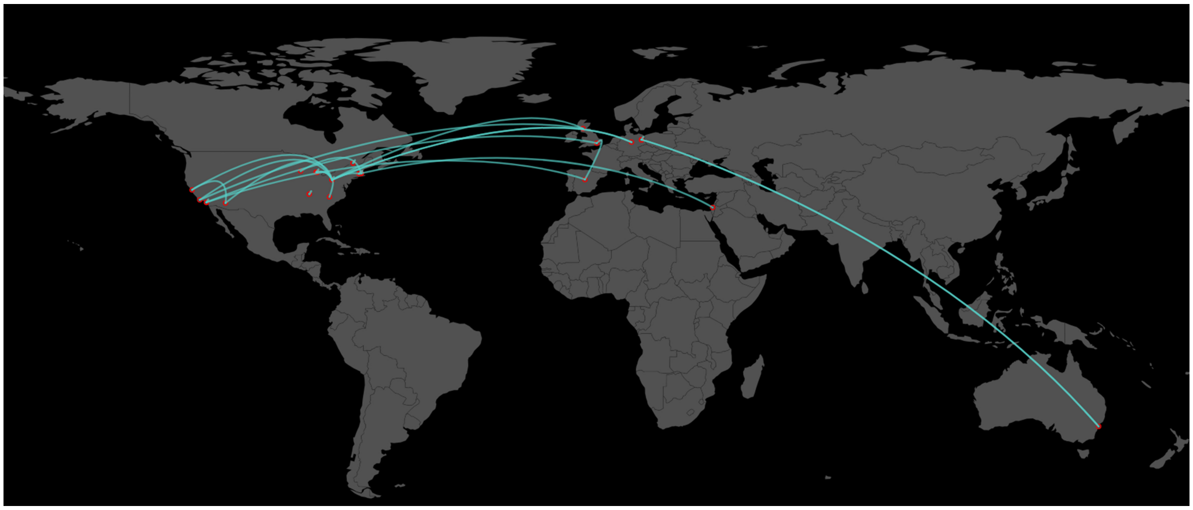


Figure 4: A simple visualization for place relations.

Keßler et al. (2012) who analyzed the academic trajectories of GIScience scholars based on their affiliation changes in publications. Table 2 summarizes the studies discussed above.

Table 2: Summary of the discussed studies on extracting place relations and sequences.

Study	Main Task	Methods
Hecht and Moxley (2009) Salvini and Fabrikant (2016) Adams and Gahegan (2016)	Exploring and quantifying place and time relations	Place co-occurrences or hyperlinks in Wikipedia pages
Liu et al. (2014) Hu et al. (2017)	Extracting place relations and analyzing distance decay effects	Place co-occurrences in news articles and gravity models
Taylor (1997) Beaverstock et al. (2000)	Analyzing city relations and city networks	Content analysis based on sampled newspapers in cities
Keßler et al. (2012)	Extracting place sequences and trajectories	Author affiliation analysis based on publication records

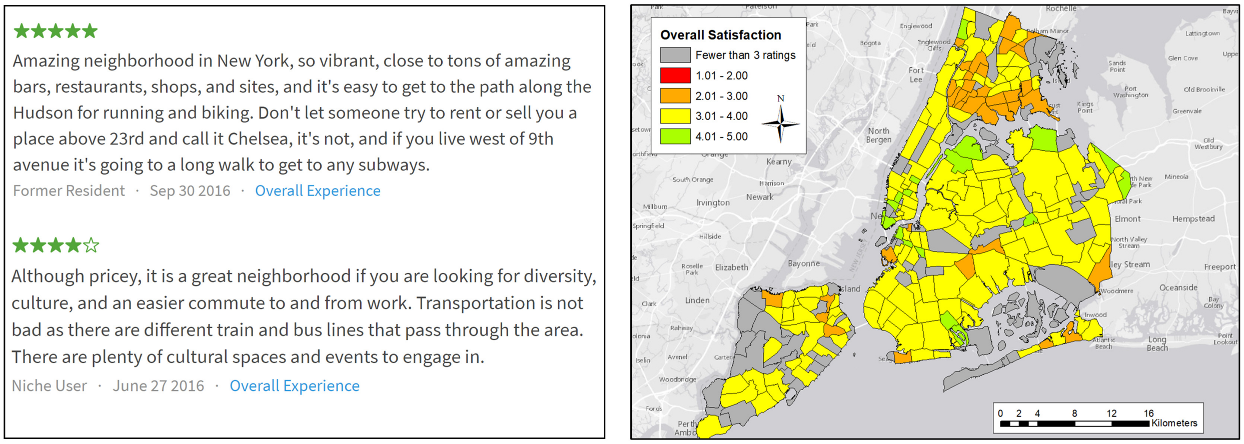
3.3. Place opinions/emotions

Geo-text data contains words expressed by people, which make it less suitable for studying environmental variables but more appropriate for examining the opinions and emotions of people. Sentiment analysis is a subfield in natural language processing (NLP) (Pang et al., 2008; Liu, 2012), which also attracted the interests of GIScience researchers. Using travel blog data, Ballatore and Adams (2015) analyzed the emotions of place types and constructed a vocabulary that associates sentiment words with place types. Based on geotagged tweets, Nelson et al. (2015) developed a geovisual analytics tool, called *SPoTvis*, and applied it to the tweets related to the debate on the Affordable Care Act in the US. Wang et al. (2016) performed text mining on geotagged tweets in the response to a wildfire, and detected the attitudes of people, such as their appreciation to fire fighters. Looking into TripAdvisor hotel reviews, Cataldi et al. (2013) proposed an approach for detecting the sentiments of people toward different aspects of a hotel, such as its location convenience and food quality. Wang and Zhou (2016) performed sentiment analysis on TripAdvisor hotel reviews within the same city, and found that spatial dependence exists in the satisfaction of customers. Wartmann and Purves (2018) investigated the sense of place of people towards landscape features (e.g., mountains and rivers) by collecting free listings and place descriptions from visitors, and found that the elicited sense of place was similar across landscape types. Table 3 summarizes the discussed studies.

Table 3: Summary of the discussed studies on extracting place opinions and emotions.

Study	Main Task	Methods
Ballatore and Adams (2015)	Extracting the emotions associated with place types	Sentiment analysis based on travel blog data
Nelson et al. (2015)	Visualizing and analyzing opinions on political events	Geovisual analytics based on geotagged tweets
Wang et al. (2016)	Understanding the attitudes of people in disaster response	Text analysis based on geotagged tweets
Cataldi et al. (2013) Wang and Zhou (2016)	Examining the sentiments of people toward hotels	Sentiment analysis based on TripAdvisor hotel reviews
Wartmann and Purves (2018)	Investigating sense of place towards landscape features	Interviews, frequent terms, and similarity comparisons

Neighborhood reviews are a relatively new type of geo-text data. In recent years, there is an emergence of websites, e.g., StreetAdvisor and Niche, designed to help people find suitable neighborhoods to live. On these websites, current or previous residents can review their neighborhoods. Figure 5(a) shows two reviews on a neighborhood in New York City (NYC), and Figure 5(b) shows the overall satisfaction levels based on the review ratings. Analyzing these reviews can help understand the perceptions of people, and can benefit urban planning and quality of life studies (Keßler et al., 2005; Das, 2008).



★★★★★

Amazing neighborhood in New York, so vibrant, close to tons of amazing bars, restaurants, shops, and sites, and it's easy to get to the path along the Hudson for running and biking. Don't let someone try to rent or sell you a place above 23rd and call it Chelsea, it's not, and if you live west of 9th avenue it's going to a long walk to get to any subways.

Former Resident · Sep 30 2016 · Overall Experience

★★★★☆

Although pricey, it is a great neighborhood if you are looking for diversity, culture, and an easier commute to and from work. Transportation is not bad as there are different train and bus lines that pass through the area. There are plenty of cultural spaces and events to engage in.

Niche User · June 27 2016 · Overall Experience

(a) (b)

Figure 5: (a) Two reviews from a neighborhood review website; (b) the overall satisfaction levels of people toward neighborhoods (the highly-satisfied are in green while the less-satisfied are in orange).

3.4. Place zones

Place zones are another type of knowledge that can be extracted from geo-text data. While detecting hot zones from location data is a common GIS operation, the uniqueness of using geo-text data lies in its rich semantics: we can understand the diverse reasons underlying the formation of these zones based on the words of people. There exist many empirical studies on extracting place zones using geo-text data. Based on geotagged Flickr photos, Rattenbury and Naaman (2009) identified point clusters using K-means clustering, and detected representative textual tags for each cluster using an algorithm called TagMaps. Andrienko et al. (2010b) proposed a visual analytics framework which detects special place zones containing periodic or irregular events. Hu et al. (2015) used geotagged Flickr photos to extract urban areas of interest (AOI) by performing DBSCAN clustering and chi-shape algorithm (Figure 6(a)). They employed term frequency and inverse document frequency (TF-IDF) to identify representative words for the extracted AOI (Figure 6(b)). Velasco et al. (2017) identified place zones in the city of Quito, Ecuador based on venue locations and textual reviews on TripAdvisor, and labeled these zones with user-generated texts. Some research combined multiple types of data. For example, Jenkins et al. (2016) examined both geotagged tweets and Wikipedia pages, and applied topic modeling, semantic analysis, and geospatial clustering to find locations with a collective sense of place. Gao et al. (2017) synthesized geotagged social media data from Twitter, Flickr, and Instagram to extract cognitive zones such as "SoCal" and "NorCal". Using the data from Foursquare, Twitter, and Yik Yak, McKenzie and Adams (2017) compared the regions identified based on place instances and place mentions. Table 4 summarizes the discussed studies.

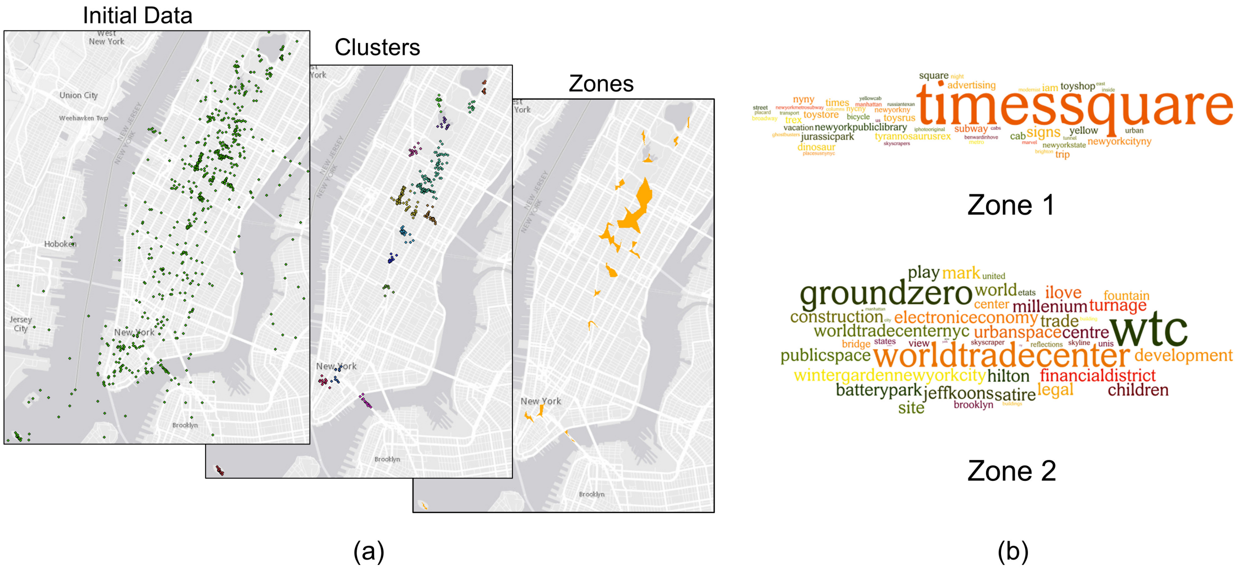


Figure 6: (a) Place zones extracted from geotagged Flickr photo data in NYC; (b) representative words of two zones.

3.5. Place impacts

Geo-text data can help reveal the impact of an event, such as a natural disaster, a public policy, an infectious disease, or others (all are referred to as *target event*). From texts, we can understand the attitudes of people; from locations, we can examine the geographic areas where people are reacting to the event. Here, we focus more on the *from* location rather

Table 4: Summary of the discussed studies on extracting place zones from geo-text data.

Study	Main Task	Methods
Rattenbury and Naaman (2009)	Extracting representative tags for special zones within a city	K-means clustering and TagMaps on geotagged photos
Andrienko et al. (2010b)	Detecting special place areas that contain events	A visual analytics framework with event detection
Hu et al. (2015)	Extracting urban AOI and representative words	DBSCAN clustering, chi-shape algorithm, and TF-IDF
Velasco et al. (2017)	Identifying place zones in the city of Quito, Ecuador	K-means clustering and TF-IDF on TripAdvisor data
Jenkins et al. (2016)	Identifying locations with a collective sense of place	Topic modeling, semantic analysis, and spatial clustering
Gao et al. (2017)	Extracting vague cognitive zones and semantic topics	Clustering and topic modeling based on multiple types of data
McKenzie and Adams (2017)	Comparing the regions extracted in different ways	KDE on multiple types of social media data

than the *about* location of geo-text data. In addition, the impacts examined here are *social* rather than *physical* impacts of events.

Many studies have utilized geo-text data to investigate the impacts of events. A notable example is Google's Flu Trends (GFT) (Ginsberg et al., 2009), in which the search keywords from people were linked to their locations based on IP addresses to predict the intensities of influenza-like illness (ILI) in different geographic areas. While GFT eventually failed, it nevertheless demonstrated a novel idea by linking texts to locations. Based on a sample of news articles, Wang and Stewart (2015) examined the impact of Hurricane Sandy by extracting place names, timestamps, and emergency information (e.g., power failure). Also using news articles, Peuquet et al. (2015) developed a computational method that extends the T-pattern analysis, and applied this technique to discovering the event associations during the Arab Spring. Geotagged tweets are widely used for studying place impacts (Tsou, 2015; Haworth and Bruce, 2015). Focusing on public health issues, Issa et al. (2017) studied the spatial diffusion of tweets about flu in four different cities, while Nagar et al. (2014) used daily geotagged tweets in NYC to investigate the spatiotemporal tweeting behavior related to ILI. Looking into disaster responses, Crooks et al. (2013) examined the spatial and temporal characteristics of tweets after an earthquake, while De Longueville et al. (2009) investigated the tweeting activities during a major forest fire. There exist other methods and visual analytics systems for detecting anomalies (Chae et al., 2012; Thom et al., 2012; Andrienko et al., 2013), extracting topics and events (Cho et al., 2016), and analyzing place-time-attribute information (Pezanowski et al., 2017). Table 5 summarizes the discussed studies.

With its ability of capturing real-time public reactions, geotagged tweets have become a convenient resource for exploring the impact of an event. One simple approach is to first retrieve related tweets using keywords, and then examine the spatiotemporal patterns of the retrieved tweets. Figure 7 shows an example of using this approach for exploring the impact of Hurricane Irma in September 2017. From the tweet counts on different days (at the bottom of the figure), we can see that most tweets were posted between Sept. 9th and 11th when Irma was landing and moving inland Florida. By dividing the entire dataset into three groups (based on the two red dotted lines), we can see different frequent words at different stages

Table 5: Summary of the discussed studies on examining place impacts based on geo-text data.

Study	Main Task	Methods
Ginsberg et al. (2009)	Predicting the locations and intensities of ILI	Examining the spatiotemporal patterns of search keywords
Wang and Stewart (2015)	Understanding the impacts of Hurricane Sandy	Extracting place and event information from news articles
Peuquet et al. (2015)	Discovering the associations of social and political events	A computational method that extends the T-pattern analysis
Issa et al. (2017) Nagar et al. (2014) Crooks et al. (2013) De Longueville et al. (2009)	Understanding the spatial and temporal impacts of diseases and disasters	Spatial, temporal, and content analysis on geotagged tweets

of the hurricane. By visualizing the tweet locations on three specific days (at the top of the figure), we can see the major areas of the tweets. This spatial-temporal-thematic analysis also shows that the place impact examined here is *social* rather than *physical* impact, since most tweets came from the cities rather than the areas lying on the hurricane path.

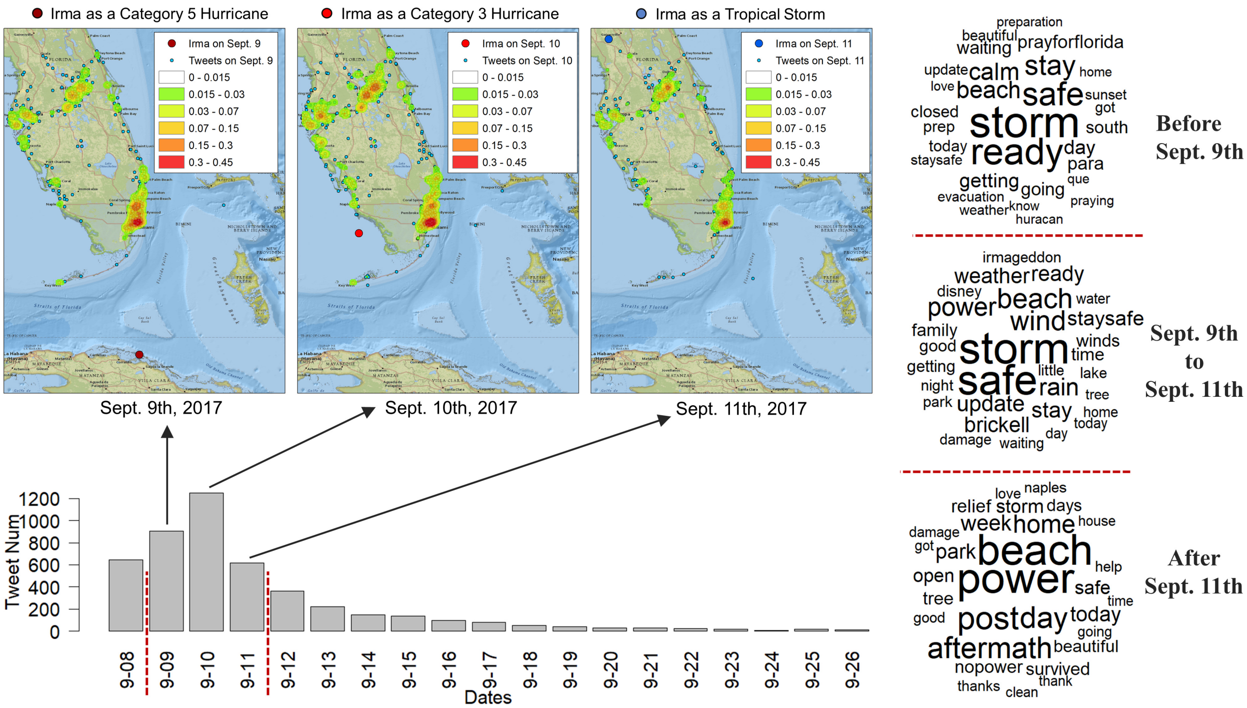


Figure 7: Exploratory analysis on the impact of Hurricane Irma based on a sample of geotagged tweets.

3.6. Summary

This section answers the question "*what can we get from geo-text data?*" by reviewing a large number of studies, discussing their used datasets and methods, and organizing them based on the types of knowledge discovered. Place names are important geographic information that can be extracted from texts, and are necessary for more advanced tasks, such as place relation or sequence analysis. We can investigate the opinions and sentiments of people attached to places, and can identify place zones and understand their meanings. We can also study the impact of an event by examining the attitudes of people in different geographic areas. While five types of knowledge are identified here, this list is not exhaustive and other types of knowledge can be discovered as well.

4. Towards a generalized workflow for analyzing geo-text data

While previous studies used different types of geo-text data and examined problems in various domains, they share similar procedures in data processing and analysis. This section extracts a workflow from the previous studies in order to provide a general reference for future research. Figure 8 illustrates this workflow.

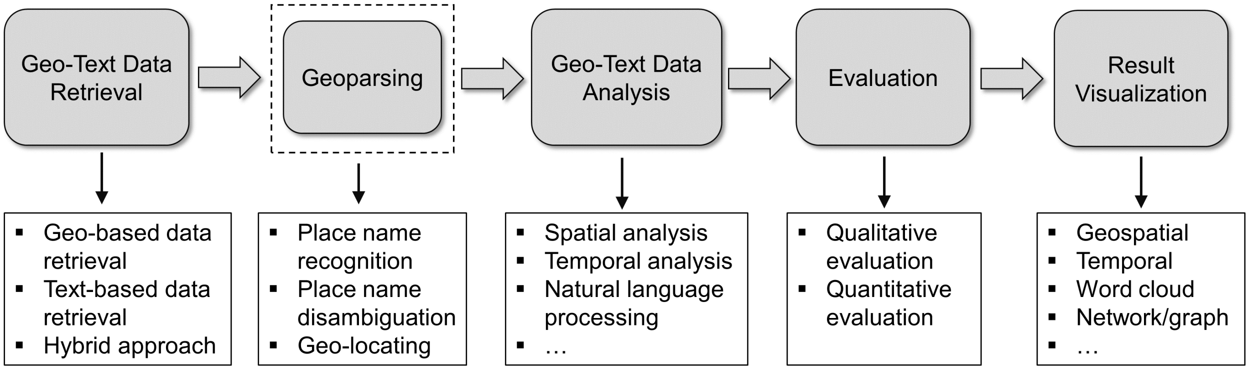


Figure 8: A generalized workflow for analyzing geo-text data.

Geo-text data retrieval. Retrieving relevant geo-text data is the first step for conducting a study. The retrieval process can be performed in three major ways. The first one is based on location, in which a bounding extent is often used. For example, we can retrieve all geotagged Wikipedia pages within the boundary of California. The second approach is text-based data retrieval, in which keywords or topics are used to retrieve data. For example, we can collect a set of news articles related to certain keywords. The third one is a hybrid approach which retrieves geo-text data using both locations and texts.

Geoparsing. This step recognizes the words that represent place names from texts, resolves the ambiguous names, and geo-locates the place names to their corresponding spatial footprints. This step is within a dotted rectangle in Figure 8, since explicit geo-text data may not need geoparsing. While multiple geoparsers exist, they can have very different performances when applied to different testing corpora (Monteiro et al., 2016). Gritta et al. (2017) tested five geoparsers using the same datasets, and their performances based on one dataset are provided in Table 6. When applied to a corpus with many ambiguous place names, the performance of a geoparser can decrease dramatically (Ju et al., 2016).

Geo-text data analysis. This is a key step in the workflow. Many methods can be utilized, such as named entity recognition, topic modeling, and sentiment analysis for the text part, and KDE, geospatial clustering, spatial autocorrelation for the location part. When timestamps are available, temporal analysis can also be performed. More systematically, we can categorize the analysis process into *geo-first* and *text-first*. In *geo-first*, we start from the locations of geo-text data by segmenting or grouping them. Figure 9 shows three

Table 6: Performances of five geoparsers applied to the same dataset (Gritta et al., 2017).

	Precision	Recall	F-score
GeoTxt (Karimzadeh et al., 2013)	0.80	0.59	0.68
Edinburgh (Alex et al., 2015)	0.71	0.55	0.62
Yahoo! PlaceSpotter	0.64	0.55	0.59
CLAVIN	0.81	0.44	0.57
TopoCluster (DeLozier et al., 2015)	0.81	0.64	0.71

examples of segmenting data using the administrative boundary, a grid-based tessellation, and a clustering technique respectively. Sometimes, the identified data groups can also spatially overlap. With the grouped data, we can use the texts associated with each group to examine its semantics. The work of Andrienko et al. (2010b) is an example of the *geo-first* approach. In *text-first*, we begin with the text part by extracting information from it, and then investigate the spatial or spatiotemporal patterns of the extracted information. For example, we can first extract place relations based on place name co-occurrences in texts, and then explore their patterns in geographic space. The work of Peuquet et al. (2015) is an example of the *text-first* approach.

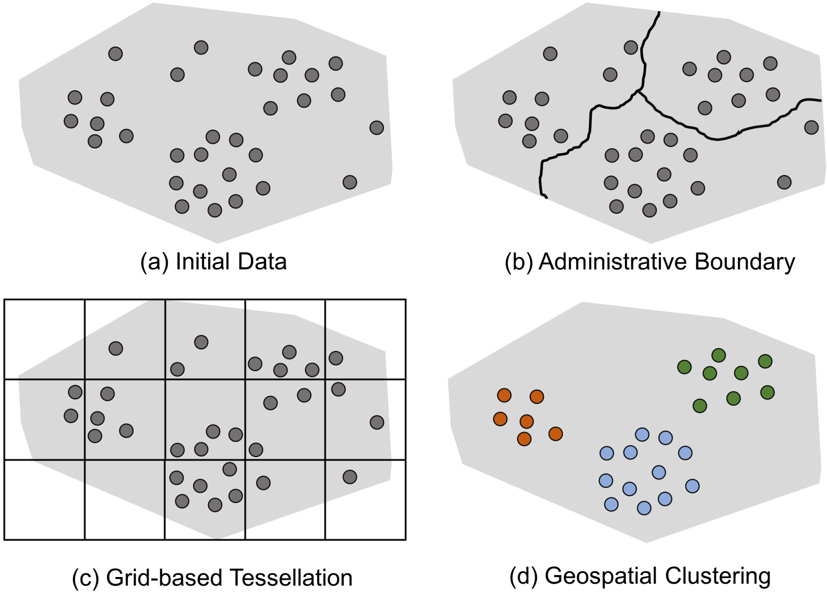

(a) Initial Data
(b) Administrative Boundary
(c) Grid-based Tessellation
(d) Geospatial Clustering

Figure 9: Three methods for grouping geo-text data based on their locations.

Evaluation. Evaluation is critical for ensuring that the extracted knowledge is valid and useful. Evaluations for geo-text data can benefit from both qualitative and quantitative assessments. While qualitative assessment is sometimes criticized for its lack of representativeness, it provides intuitive understandings on the obtained results. Meanwhile, quantitative assessment is necessary for robust evaluations. Suitable quantitative metrics are project specific, but common ones include accuracy, error rate, and correlation coefficient. It is usually better to combine qualitative and quantitative assessments to provide both intuitive and robust evaluations.

Result visualization. Many techniques can be employed to visualize the knowledge extracted from geo-text data. Given the locations and timestamps, we can visualize the results as maps, temporal sequences, or spatiotemporal cubes (Andrienko et al., 2010a; Luo and MacEachren, 2014; Nelson et al., 2015). With texts, techniques, such as word clouds, multi-dimensional scaling, or self-organization maps, can be used (Skupin and Fabrikant, 2003). Network or graph visualizations can be employed to show the extracted place relations. There also exist geovisual analytics systems, such as SensePlace2 (MacEachren et al., 2011), SensePlace3 (Pezanowski et al., 2017), STempo (Robinson et al., 2017), and VAiRoma (Cho et al., 2016), that support overview and detailed visualization of data.

In summary, this section answers the question "*what is a general workflow that we can follow to analyze geo-text data?*" by extracting a step-by-step data analysis skeleton from the literature. One can flesh out a specific workflow by choosing a data retrieval approach, selecting a geoparser, deciding data analysis methods, designing evaluation experiments, and choosing visualization techniques. Such a workflow also carries important implications that should be noted. For the retrieved geo-text data, the data source can directly affect the analysis results. For example, depending on whether the data are generated by tourists or local residents, the identified place zones may reflect the different interests of the two groups of people. For the step of geoparsing, while methods and geoparsers have been developed, they are not perfect and errors can propagate to the downstream of data analysis. For geo-text data analysis, the chosen methods can have certain assumptions. For example, term frequency analysis assumes that the importance of a term is reflected in its mentioning frequency. However, it is possible that some important terms are mentioned only a few times. Finally, the selected data visualization methods can also distort the perceptions of readers. These implications do not mean a decreased value of geo-text data analysis. In fact, they make some problems, e.g., local versus tourist places, more interesting. However, these implications should be kept in mind when we interpret the analysis results.

5. Challenges for future research

This section discusses some of the key challenges for future research based on geo-text data. The discussion is organized based on the dual parts of geo-text data, the special link between them, and the development of future methods.

Uncertainty of spatial footprints. As discussed previously, the spatial footprints of geo-text data can be points, lines, polygons, and even polyhedra. Accordingly, geo-text data can be affected by the same uncertainty issues like other GIS data (Ehlschlaeger et al., 1997). When the spatial footprint is a vague region, suitable representation methods need to be selected and ideally verified with human perceptions (Montello et al., 2003; Jones et al., 2008). In addition, some existing geo-text data have only point-based footprints (e.g., the point location of Knoxville in Figure 1(a)), whereas the geographic features may be better represented using polylines or polygons given the application scale. In short, how can we represent the spatial footprint of geo-text data more accurately based on application needs?

Ambiguity of texts. Despite the advancements in NLP, it is still challenging to accurately understand natural language texts. The commonly used bag-of-words model ignores word orders and cannot capture the structures of sentences. Other text processing methods, such as parse trees and smoothing windows (Mikolov et al., 2013; Schuster and Manning, 2016), as well as the recent deep neural nets provide more advanced approaches for modeling texts (Tang et al., 2015; Li et al., 2017), but often require large amounts of labeled corpora and can take much longer time to train than traditional models. Even when the state-of-the-art methods are employed, some of the entities, topics, and sentiments extracted from geo-text data can still be incorrect. In short, how can we improve the accuracy of understanding texts given reasonable computing and data resources?

Unclear Links between locations and texts. The link between geography and text makes geo-text data special. However, such links can be unclear and can bring challenges in two areas. First, the text can be related to multiple locations. As discussed in Section 2, geo-text data can have both *about* location and *from* location. Besides, some geo-text data, such as news articles, can mention multiple place names in the same textual context. In these situations, a suitable method needs to be selected to link the text with the right geographic reference. Second, the text may link to only part of a referred geographic feature. For example, one may be referring to only the peak of a mountain or the mouth of a river, and in such situations, it can be inappropriate to link the text to the whole feature. Likewise, it can also be inappropriate to link the text to only part of a geographic feature when the text in fact refers to the whole feature. In short, how can we correctly and accurately link locations and texts for geo-text data?

Loose integration between spatial and text analysis. Because of the dual parts of geo-text data, many previous studies integrated spatial and text analysis. Such integrations, however, were usually in a loose manner, which is also reflected in the *geo-first* or *text-first* approaches identified in Section 4. While such loose integration can already discover useful knowledge, methods closely integrating spatial and text analysis may better address problems in text understanding and geoparsing. For example, the geographic context of a person can help interpret the words of this person; meanwhile, the topics that people talk about can help infer their locations. Some research has examined such a close integration (Cocos and Callison-Burch, 2017). In addition, using geographic knowledge to improve computational models, rather than simply taking methods from other fields, can help increase the impact of GIScience overall. In short, how can we more closely integrate locations and texts to develop methods with potential cross-domain impacts?

In summary, this section answers the question "*what are some key challenges for future research?*" by identifying the difficulties in four areas. The first two areas, spatial footprint uncertainty and text ambiguity, are related to not only geo-text data but also spatial data and linguistic data in general. Thus, they can benefit from the advancements in the corresponding fields. The second two areas are related to the special links between locations and texts, which are more unique to geo-text data and need integrated thinking rather than separated studies. These challenges provide great opportunities for future research.

6. Conclusions and summary

Geo-text data is a collective term referring to the kind of data that contains links between geographic locations and natural language texts. Compared with typical GIS data, such as temperature measurements and DEM, in which numeric values are attached to locations, geo-text data links unstructured texts to locations. Such special links make geo-text data unique, and enable new research topics on place names, place relatedness, sense of place, vague spatial footprints, cognitive regions, the attitudes of people toward events, and many other topics that require an additional dimension of human experience. Geo-text data greatly facilitates research in data-driven geospatial semantics by enhancing our understanding on the semantics of geographic features and terms. This paper provides a formal representation of geo-text data based on a general GIS theory. Explicit and implicit geo-text data are differentiated, and two major processes that generate geo-text data are discussed. A systematic literature review is conducted which organizes previous studies based on the types of knowledge discovered. A generalized workflow is then extracted from the literature, and key challenges for future research are discussed. Overall, geo-text data and the related research are situated at the intersection of geography, computer science, linguistics, cognitive science, statistics, and other related fields. Interdisciplinary collaboration is and will continue playing an important role in fostering advancements in this growing and exciting area.

Acknowledgements

The author would like to thank Dr. James Cheshire, Dr. Michael Goodchild, and the anonymous reviewers for their constructive comments and suggestions. This work is supported by the Professional and Scholarly Development Award (Award Number: R011038-002) from the University of Tennessee.

References

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

Alderman, D. H. (2016). Place, naming and the interpretation of cultural landscapes. *Heritage and Identity, edited by Brian Graham and Peter Howard*, pages 195–213.

Alex, B., Byrne, K., Grover, C., and 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., and 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*, pages 273–280. ACM.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., and Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering*, 15(3):72–82.

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., Jern, M., Kraak, M.-J., Schumann, H., and Tominski, C. (2010a). Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10):1577–1600.

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M., and Pölitz, C. (2010b). Discovering bits of place histories from people's activity traces. In *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 59–66. IEEE.

Ballatore, A. and Adams, B. (2015). Extracting place emotions from travel blogs. In *Proceedings of AGILE*, volume 2015, pages 1–5.

Ballatore, A., Bertolotto, M., and Wilson, D. C. (2014). An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica*, 18(4):747–767.

Beaverstock, J. V., Smith, R. G., Taylor, P. J., Walker, D., and Lorimer, H. (2000). Globalization and world cities: some measurement methodologies. *Applied geography*, 20(1):43–63.

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

Cataldi, M., Ballatore, A., Tiddi, I., and Aufaure, M.-A. (2013). Good location, terrible food: detecting feature sentiment in user-generated reviews. *Social Network Analysis and Mining*, 3(4):1149–1163.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., and Ertl, T. (2012). Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 143–152. IEEE.

Cheshire, J. A. and Longley, P. A. (2012). Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science*, 26(2):309–325.

Cho, I., Dou, W., Wang, D. X., Sauda, E., and Ribarsky, W. (2016). Vairoma: A visual analytics system for making sense of places, times, and events in roman history. *IEEE transactions on visualization and computer graphics*, 22(1):210–219.

Cocos, A. and Callison-Burch, C. (2017). The language of place: Semantic value from geospatial context. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 99–104.

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

Das, D. (2008). Urban quality of life: A case study of guwahati. *Social Indicators Research*, 88(2):297–310.

De Longueville, B., Smith, R. S., and 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*, pages 73–80. ACM.

DeLozier, G., Baldridge, J., and London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *AAAI*, pages 2382–2388.

Ehlschlaeger, C. R., Shortridge, A. M., and Goodchild, M. F. (1997). Visualizing spatial data uncertainty using animation. *Computers & Geosciences*, 23(4):387–395.

Frank, J. R., Rauch, E. M., and Donoghue, K. (2006). Spatially coding and displaying information. US Patent 7,117,199.

Freire, N., Borbinha, J., Calado, P., and 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*, pages 339–348. ACM.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., and Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, 31(6):1245–1271.

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

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Goodchild, M. F., Yuan, M., and Cova, T. J. (2007). Towards a general theory of geographic representation in gis. *International Journal of Geographical Information Science*, 21(3):239–260.

Gregory, I., Donaldson, C., Murrieta-Flores, P., and 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, M. T., Limsopatham, N., and Collier, N. (2017). What's missing in geographical parsing? *Language Resources and Evaluation*, pages 1–21.

Haklay, M., Singleton, A., and Parker, C. (2008). Web mapping 2.0: The neogeography of the geoweb. *Geography Compass*, 2(6):2011–2039.

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

Hecht, B. and Moxley, E. (2009). Terabytes of Tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge. *Spatial information theory*, pages 88–105.

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

Hollenstein, L. and 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. (2018). Geospatial semantics. In Huang, B., editor, *Comprehensive Geographic Information Systems*, pages 80–94. Elsevier, Oxford.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., and Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54:240–254.

Hu, Y. and Janowicz, K. (2016). Enriching top-down geo-ontologies using bottom-up knowledge mined from linked data. *Advancing Geographic Information Science: The Past and Next Twenty Years*, pages 183–198.

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

Huang, Q. and 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., and Ghazi, D. (2017). Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49(2):237–253.

Issa, E., Tsou, M.-H., Nara, A., and 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.

Jenkins, A., Croitoru, A., Crooks, A. T., and Stefanidis, A. (2016). Crowdsourcing a collective sense of place. *PloS one*, 11(4):e0152932.

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

Jones, C. B., Purves, R. S., Clough, P. D., and 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., and McKenzie, G. (2016). Things and strings: Improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20*, pages 353–367. Springer.

Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J. O., Hardisty, F., Pezanowski, S., Mitra, P., and MacEachren, A. M. (2013). Geotxt: a web api to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval*, pages 72–73. ACM.

Keßler, C., Janowicz, K., and Kauppinen, T. (2012). spatial@ linkedscience–exploring the research field of giscience with linked data. In *International Conference on Geographic Information Science*, pages 102–115. Springer.

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

Keßler, C., Rinner, C., and Raubal, M. (2005). An argumentation map prototype to support decision-making in spatial planning. In *Proceedings of AGILE*, volume 5, pages 26–28.

Kuhn, W. (2005). Geospatial semantics: why, of what, and how? *Journal on data semantics III*, pages 587–587.

Laurini, R. (2015). Geographic ontologies, gazetteers and multilingualism. *Future Internet*, 7(1):1–23.

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

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

Li, P.-H., Dong, R.-P., Wang, Y.-S., Chou, J.-C., and Ma, W.-Y. (2017). Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2654–2659.

Li, W., Goodchild, M. F., and Raskin, R. (2014). Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth*, 7(1):17–37.

Lieberman, M. D. and 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*, pages 843–852. ACM.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

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

Longley, P. A., Cheshire, J. A., and Mateos, P. (2011). Creating a regional geography of britain through the spatial analysis of surnames. *Geoforum*, 42(4):506–516.

Luo, W. and MacEachren, A. M. (2014). Geo-social visual analytics. *Journal of spatial information science*, 2014(8):27–66.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., and Blanford, J. (2011). Senseplace2: Geotwitter analytics support for situational awareness. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 181–190. IEEE.

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

Mark, D. M., Turk, A. G., Burenhult, N., and Stea, D. (2011). *Landscape in language: Transdisciplinary perspectives*, volume 4. John Benjamins Publishing.

McKenzie, G. and Adams, B. (2017). Juxtaposing thematic regions derived from spatial and platial user-generated content. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 86, pages 1–13. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, H. J. and Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4):449–461.

Monteiro, B. R., Davis, C. A., and Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96:23–34.

Montello, D. R., Friedman, A., and Phillips, D. W. (2014). Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, 28(9):1802–1820.

Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-3):185–204.

Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., and Brownstein, J. S. (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).

Nelson, J. K., Quinn, S., Swedberg, B., Chu, W., and MacEachren, A. M. (2015). Geovisual analytics approach to exploring public political discourse on twitter. *ISPRS International Journal of Geo-Information*, 4(1):337–366.

Nesi, P., Pantaleo, G., and 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. and Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22(3):265–287.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Peuquet, D. J., Robinson, A. C., Stehle, S., Hardisty, F. A., and Luo, W. (2015). A method for discovery and analysis of temporal patterns in complex event data. *International Journal of Geographical Information Science*, 29(9):1588–1611.

Pezanowski, S., MacEachren, A. M., Savelyev, A., and Robinson, A. C. (2017). Senseplace3: a geovisual framework to analyze place–time–attribute information in social media. *Cartography and Geographic Information Science*, pages 1–18.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., 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.

Rattenbury, T. and Naaman, M. (2009). Methods for extracting place semantics from Flickr tags. *ACM Trans. Web*, 3(1):1–30.

Robinson, A. C., Peuquet, D. J., Pezanowski, S., Hardisty, F. A., and Swedberg, B. (2017). Design and evaluation of a geovisual analytics system for uncovering patterns in spatio-temporal event data. *Cartography and Geographic Information Science*, 44(3):216–228.

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

Schuster, S. and Manning, C. D. (2016). Enhanced English universal dependencies: An improved representation for natural language understanding tasks. In *LREC*.

Skupin, A. and Fabrikant, S. I. (2003). Spatialization methods: a cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science*, 30(2):99–119.

Spitz, A., Geiß, J., and 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*, page 2. ACM.

Stefanidis, A., Vraga, E., Lamprianidis, G., Radzikowski, J., Delamater, P. L., Jacobsen, K. H., Pfoser, D., Croitoru, A., and Crooks, A. (2017). Zika in twitter: Temporal variations of locations, actors, and concepts. *JMIR Public Health and Surveillance*, 3(2).

Tang, D., Qin, B., and Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432.

Tardy, C., Falquet, G., and Moccozet, L. (2016). Semantic enrichment of places with vgi sources: a knowledge based approach. In *Proceedings of the 10th Workshop on Geographic Information Retrieval*, page 6. ACM.

Taylor, P. J. (1997). Hierarchical tendencies amongst world cities: a global research proposal. *Cities*, 14(6):323–332.

Thom, D., Bosch, H., Koch, S., Wörner, M., and Ertl, T. (2012). Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *2012 IEEE Pacific visualization symposium (PacificVis)*, pages 41–48. IEEE.

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

Vasardani, M., Winter, S., and Richter, K.-F. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27(12):2509–2532.

Velasco, M., San Lucas, C., Ortiz, K., Vélez, J., and Vaca, C. (2017). Secrets of quito: Discovering a city through tripadvisor. In *2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 251–255. IEEE.

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

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

Wang, W. and Stewart, K. (2015). Spatiotemporal and semantic information extraction from web news reports about natural hazards. *Computers, Environment and Urban Systems*, 50:30–40.

Wang, Z., Ye, X., and Tsou, M.-H. (2016). Spatial, temporal, and content analysis of twitter for wildfire hazards. *Natural Hazards*, 83(1):523–540.

Wartmann, F. M., Acheson, E., and Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach. *International Journal of Geographical Information Science*, pages 1–21.

Wartmann, F. M. and Purves, R. S. (2018). Investigating sense of place as a cultural ecosystem service in different landscapes through the lens of language. *Landscape and Urban Planning*, 175:169–183.

Widener, M. J. and Li, W. (2014). Using geolocated twitter data to monitor the prevalence of healthy and unhealthy food references across the us. *Applied Geography*, 54:189–197.

Wing, B. P. and 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*, pages 955–964. Association for Computational Linguistics.

Zhang, C., Zhao, T., and Li, W. (2015). Towards an interoperable online volunteered geographic information system for disaster response. *Journal of Spatial Science*, 60(2):257–275.

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