Is This Statement About A Place? Comparing two perspectives*†

Alan M. MacEachren
GeoVISTA Center, Dept. of Geography, Penn State University, University Park, PA, USA
maceachren@psu.edu

Richard Caneba
College of Information Science & Technology, Penn State University, University Park, PA, USA

Hanzhou Chen
Department of Geography, Penn State University, University Park, PA 16802, USA

Harrison Cole
Department of Geography, Penn State University, University Park, PA 16802, USA

Emily Domanico
Department of Geography, Penn State University, University Park, PA 16802, USA

Nicholas Triozzi
Dept. of Anthropology, Penn State University, University Park, PA 16802, USA

Fangcao Xu
Department of Geography, Penn State University, University Park, PA 16802, USA

Liping Yang
Department of Geography, Penn State University, University Park, PA 16802, USA

Abstract

Text often includes references to places by name; in prior work, more than 20% of a sample of event-related tweets were found to include place names. Research has addressed the challenge of leveraging the geographic data reflected in text statements, with well-developed methods to recognize location mentions in text and related work on automated toponym resolution (deciding which place in the world is meant by a place name). A core issue that remains is to distinguish between text that mentions a place or places and text that is about a place or places. This paper presents the first step in research to address this challenge. The research reported here sets the conceptual and practical groundwork for subsequent supervised machine learning research; that research will leverage human-produced training data, for which a judgment is made about whether a statement is or is not about a place (or places), to train computational methods to do this classification for large volumes of text. The research step presented here focuses on three questions: (1) what kinds of entities are typically conceptualized as places, (2) what features of a statement prompt the reader to judge a statement to be about a place (or not about a place) and (3) how do judgments of whether or not a statement is about a place compare between a group of experts who have studied the concept of "place" from a geographic perspective and a cross-section of individuals recruited through a crowdsourcing platform to make these judgments.

2012 ACM Subject Classification Information systems → Information retrieval, Computing methodologies → Natural language processing

* The first author led the research within a graduate seminar on Place & Big Data that the next 6 authors participated in (with equal contribution, listed in alphabetical order). The last author leads the machine learning process to follow. All authors contributed to writing and/or editing parts of the paper.

† The crowdsourced data collection component of the research was reviewed by the Penn State Institutional Review Board and determined to not require formal IRB review, as it meets criteria for exempt research.

10th International Conference on Geographic Information Science (GIScience 2018).
Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 44; pp. 44:1–44:6
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Keywords and phrases geographic information retrieval, spatial language, crowdsourcing

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.44

Category Short Paper

1 Introduction

The research reported here has two primary goals. The first extends beyond and motivates the present paper – to develop geographic information retrieval (GIR) methods to retrieve place-focused unstructured information from text. Our longer term project related to this goal is to explore the potential of machine/deep learning methods to categorize statements into those "about place" (or not). Work reported here is a precursor to that objective. Our second goal, the primary focus of the project reported on here, is to explore the concept of place and what it means for a statement to be "about" a place. To address this objective, we: (a) consider examples of places and attributes that lead to an entity being considered to be a place or not, (b) assess the extent to which a set of individuals with scientific understanding of place as a concept agree on whether short statements (in Twitter) are about place or not and (c) evaluate the potential to use Amazon Mechanical Turk (MTurk) crowdsourcing to build large corpora of statements classified into those that are or are not about a place (for subsequent use in training and testing of machine/deep learning).

2 What is a place?

Place has been a core concept of Geography for centuries. Trying to define "place" in a way that appeals across multiple disciplines has been a beguiling problem for geographers [5]. From a humanist perspective, Tuan [11] defined place as "spatial locations that have been given meaning by human experience." Golledge [3], from a behavioral science perspective, contended that "although place is a dimensionless spatial term, it is conventionally interpreted as a multidimensional phenomenon (emphasis added)." From a social perspective, place can be characterized as an emergent phenomenon, its evolution is non-linear and shaped by many, varying perspectives, constructed and made tangible by social processes and historical narratives, see: [8]. In spite of many efforts to define place, the concept has been difficult to formalize sufficiently to leverage digital data for understanding place as a dynamic construct [4]. Here, we focus on exploring place-related discourse in language. For a broader overview of place in the context of GIScience and Big Data, see: [7].

3 Typical "places"

As a discussion starting point in a Place & Big Data seminar, 6 students (co-authors) completed two tasks in successive weeks. The first focused on listing and categorizing "places," the second on listing attributes that distinguish places from other entities. Entities proposed as places varied in scale (from the Treaty Oak, through countries, to The Universe). Some entities were uniquely personal (e.g., "the secret fort near my house growing up"). Others, while personally relevant were also prototypical examples of local places (e.g., "Flightpath Coffee"). Some entities, while locations one can be at or in, are also prominent landmarks (e.g., "Golden Gate Bridge", Taj Mahal).

One parsing of entities listed is to apply Montello's [9] four Scales of Psychological Space: figural (smaller than the body), vista (potentially apprehended from one place – single rooms,

Table 1

Vista (43)	Environmental Scale (63)	Geographical Scale (25)
Treaty Oak	Museum of Modern Art	Pennsylvania
This classroom	Lake Michigan	The Great Basin
Hubble telescope	16801	Midwest
secret fort near my house growing up	Yahoo! Inc. Headquarters	Mesopotamia
The bathroom	Grand Central	United States
Craig O's Pastaria walk-in freezer	JFK International Airport	Mordor
Times Square	Boalsburg, PA	I-99
Intersection Allen St and College Ave.	Korean town in LA	Yugoslavia
Golden Gate Bridge	Manhattan	Africa
My hallway closet	Wall Street	The universe

town squares, small valleys), environmental (requiring locomotion to experience – buildings, neighborhoods, cities), and geographical (much larger than the body, understood through symbolic means). Among 140 entities listed collectively, five (arguably) are figural (e.g., the atom in my foot; my shoe). The table below provides 10 examples each for the other three categories (with totals). Those at vista scale include many personal places. Most environmental and geographical scale entities are named places experienced or known by many people. Geographical scale places were least frequent, suggesting that "place" is more easily associated with locations that can be experienced; it also included the only instances of fictional (Mordor) or historical (Mesopotamia) places. Overall, few linear features were named (e.g., 2 streets, 1 wildlife drive, 1 freeway, 1 interstate, and 1 river – the Nile).

4 Statements about places: expert classification

Understanding which entities count as places is a step toward recognizing statements "about" a place. Addressing the about component is closely related to GIR research on document relevance, (e.g., [1], [10]) and on document geographic focus (e.g., [2],[6]), but focuses on statements, not documents. In this section, we present results of a classification task carried out by the 6 graduate student co-authors. The objective was to explore factors leading to statements (in tweets) being conceptualized as "about a place" (or not), and to analyze differences in opinion among individuals who have studied the concept of place formally.

4.1 Procedure

For this task, 104 tweets were sampled from a large repository, with 8 tweets each from 13 subsets related to different event types (earthquake, ebola, fire, flood, flu, malaria, measles, protest, rebels, riot, tornado, violence, womensmarch). Each sample of 8 included 4 tweets containing a formal place name and 4 tweets without a formal place. Tweets with strong offensive language, unintelligible language, or primarily hashtags and/or URLs were omitted. The sampling goal was to select tweets (whether containing formal place names or not) that varied in likelihood of being considered to be about place. Tasks were presented via Google Forms with a form heading of Is this Tweet about a place? followed by, The goal of this task is to distinguish between tweets that are "about" places (thus that are "on the subject of; concerning" places) and those that are not. Tweets appeared to participants in random order, with two choices: "Yes, it is about a place" or "No, it is not about a place."

4.2 Results and interpretation

Of the 104 tweets, 20 were judged unanimously to be about a place, with 24 more about a place by a majority (>= 4 of 6). At the other extreme, 28 tweets were judged unanimously to be not about a place, with 25 more by a majority. Seven tweets resulted in a 3-3 tie.

At the extremes, there are clear characteristics that prompt unanimity in judgments about whether a statement is or is not "about a place." For those judged as about a place, the statement is often about an event, focused on something local in geographical scale, and/or from the perspective of being on the ground. Linguistic cues in the form of locative prepositions also are common. Examples (with RT and @ references removed) include:

- ... about 20,000 people are here in Santa Ana for Orange County #womensmarch2018
- Apparently it's testing day for the tornado sirens. Skerd me to death. They're much louder at 101st and Sheridan!??

For statements judged consistently as not about a place, the most common feature is absence of reference to a geographic scale entity (thus without a name or description). This is the case even if an event probably occurring in a place is mentioned; examples include:

- Proud supporter of this & other groups trying to save this democracy.. #dontbackdown . #unitedwewin . #womensmarch2018
- ... and the government want to send arms for the rebels but not a democracy

That said, statements with place names are not always judged to be about a place; e.g., when a government is the intended meaning rather than the territory as well as when it is clear that the geographic entity mentioned is not the focus of the statement; one example is:

- It would cost $1 billion a year to eradicate malaria which kills $1 million people per year, the U.S. spends 10 billion ...

Minority views in near-unanimous "is a place" judgments (5-1) can result from too-quick reading (e.g., not noticing a place name due to abbreviation of unfamiliarity). Other factors leading to a minority view that a statement is not about a place are: statements naming more than one location, interpreting "about" strictly, or a geographic entity with indistinct boundaries. At the other extreme, a liberal definition of "about" (e.g., any mention of a proper place name counts) or considering virtual/social "places" to count (e.g., twitisphere) prompts judgments that a statement is about a place when most individuals feel it is not.

Statements with a 4-2 majority for place typically included a formal place name or abbreviation (e.g., "... about the ebola existing in jhb") and/or use a preposition tied to an event or a proper noun (e.g., "... the 0749 from Radlett cancelled due to no driver..."). Lack of unanimity, however, is prompted by many factors: unfamiliar abbreviations (jhb for Johannesburg), symbolic interpretation of a name (e.g., White House), unclear connection of name to overall statement (e.g., for hashtags), mention of multiple places (thus not ***a*** place), context points to other focus (e.g., mentions China, but tweet is "about" measles), or too little context to distinguish place from object (e.g., "the fire hydrant outside my building").

In contrast to the set above, some statements resulted in a 2-4 minority judging them to be about a place. Factors include: use of negation (Not Baghdad), unclear place reference ("Miss", could be a person's title or an abbreviation for the U.S. state), place entities mentioned as context for something else ("If TB Joshua want to heal the Ebola Victims Sierra Leone and Liberia isn't far away let him take his crusade there pls!we"), place names standing for a person (the White House, as above) or a government (Russia protests ...), vague reference (e.g., the world), description of an event, but with no place name to locate it (e.g., "I want them to stop rioting now"), use of a place name without a corresponding event

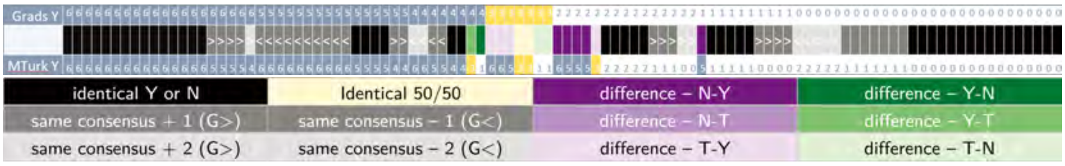

Figure 1 The top figure section (from Grads Y through MTurk Y) depicts comparison of judgments by 6 graduate students (co-authors) and 6 MTurk workers. The bottom section is a legend for the middle row of the top figure section. In the top section of the figure, the "Grads Y" row contains the number of graduate students (out of 6) who judged each of the 104 tweets to be about a place (each column signifies one tweet); the "MTurk Y" row contains the same information for the 6 MTurk workers. The tweets are ordered from those with unanimous agreement by the graduate students as being about a place (6), through those with a 3-3 split judgment, to those with unanimous agreement that the tweet is not about a place (0). Slate gray highlights all tweets with a consensus (4-2, 5-1, or 6-0) that the tweet is about a place; yellow highlights the 3-3 disagreements, and white with slate gray numbers highlights consensus (2-4, 1-5, 0-6) that the tweet is not about a place. For those that agree on consensus, but differ in number, a ">" indicates that more graduate students than MTurk workers judged the tweet to be about a place and a "<" indicates that fewer graduate students than MTurk workers judged the tweet to be about a place. The same color coding is applied to judgments by MTurk workers on each tweet. The middle row highlights agreements and disagreements between the graduate students and the MTurk workers. All that are black or gray signify that the majority in both groups agreed on 'yes' or 'no'. The two in light yellow represent 3-3 judgments by both groups. Only those tweets in purple or green have a disagreement in majority judgment. Dark green indicates a consensus on 'yes' for graduate students and 'no' for MTurk workers; dark purple indicates the reverse. Medium green indicates a consensus on 'yes' for graduate students and a 3-3 judgment by MTurk workers with the lightest green indicating a 3-3 judgment by graduate students and a 'no' by MTurk workers. The medium and light purples indicate 'no' compared to 3-3 and 3-3 compared to 'yes' for graduate students compared with MTurk workers.

(e.g., a hashtag such as #bristol but no clear connection to the rest of the text), and reference to imaginary, virtual, or fictitious places (dreams, computer games such as Minecraft).

The greatest disagreement (3 for, 3 against) are with statements referring to a location that is not specifically named (e.g., "the airport" or "the mountains"). In addition, vague locations (e.g., "We want snow here") also lead to contrasting views. In addition, a difference of opinion can result from anthropomorphizing the place or perhaps treating the statement as a metaphorical one (e.g., "Happy Independence Day Indonesia! ...").

5 Comparing crowdsourced judgment of place to expert judgment

We repeated the tweet classification activity with MTurk workers as participants. The same 104 tweets were used, grouped in eight Human Intelligence Tasks (HITs) with 13 tweets each (systematically sorted to mix the 13 event types across HITs). Instructions were identical to those for the grad students (plus the requisite informed consent statement). Google Forms was used again, to provide the tweets in random order to avoid any order effects. Each HIT was completed by 6 workers to match the 6 graduate students who initially classified the same tweet (17 workers did 1 or more HITs). Work time varied widely (from about 3min. to 50min with a median of 13/HIT or about 1min/tweet).

Data from MTurk and the 6 grad students was integrated, with tweets sorted from high to low grad "about a place" rating. This supported assessment of the extent to which crowdsourced and expert data matched and an examination of between group differences. Results are summarized graphically in Figure 1, with a detailed explanation in the caption.

6 Discussion

The research reported is a part of a larger effort focused on understanding characteristics of language related to place and creating computational methods to recognize statements (and documents) that are about places. While the initial research (focused on entities considered to be places and place attributes) was carried out in a semi-formal way as part of an ongoing course, results provide a starting point to explore the diverse characteristics that define place, including how place is related to geographic scale, personal experience, and function.

The second two parts of the research together provide insight on the challenges and possibilities for building computational methods to enable large volumes of text to be explored for place-related information. It is clear (from analysis of agreement and disagreement among a group of individuals studying place), that judging whether a statement is "about a place" depends on how "about" is interpreted as well as on the individual's view of what constitutes a "place". But, the small number of statements that resulted in a stalemate of conflicting judgments suggests that statements can be reliably categorized as being about a place (or not). The subsequent repeat of the experiment using crowdsourcing shows that reliable results are likely using this approach for all statements except those on which even experts disagree (situations with differences in what "about" means, abbreviated names, symbolic places, or imprecise/vague place references). Thus, we expect that it will be possible to build a large corpus of statements classified as being about place or not and to use them to train and test machine/deep learning methods to carry out this task with large volumes of text.

References

1 Paul D Clough, Hideo Joho, and Ross Purves. Judging the spatial relevance of documents for gir. In *European Conference on Information Retrieval*, pages 548–552. Springer, 2006.
2 Yong Gao, Dan Jiang, Xiang Zhong, and Jingyi Yu. A point-set-based footprint model and spatial ranking method for geographic information retrieval. *ISPRS International Journal of Geo-Information*, 5(7):122, 2016.
3 Reginald G Golledge. Place recognition and wayfinding: Making sense of space. *Geoforum*, 23(2):199–214, 1992.
4 Michael F Goodchild. Formalizing place in geographic information systems. In *Communities, neighborhoods, and health*, pages 21–33. Springer, 2011.
5 Jay T Johnson. Place-based learning and knowing: critical pedagogies grounded in indigeneity. *GeoJournal*, 77(6):829–836, 2012.
6 Michael D Lieberman, Hanan Samet, Jagan Sankaranarayanan, and Jon Sperling. Steward: architecture of a spatio-textual search engine. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, page 25. ACM, 2007.
7 Alan M MacEachren. Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier. In *Spatial Data Handling in Big Data Era*, pages 139–155. Springer, 2017.
8 Doreen Massey. Places and their pasts. In *History workshop journal*, volume 39, pages 182–192. JSTOR, 1995.
9 Daniel R Montello. Scale and multiple psychologies of space. In *European conference on spatial information theory*, pages 312–321. Springer, 1993.
10 Tumasch Reichenbacher, Stefano De Sabbata, Ross S Purves, and Sara I Fabrikant. Assessing geographic relevance for mobile search: A computational model and its validation via crowdsourcing. *Journal of the Association for Information Science and Technology*, 67(11):2620–2634, 2016.
11 Yi-Fu Tuan. *Space and place: The perspective of experience.* U of Minnesota Press, 1977.