
Spatial knowledge discovery from volunteered spatial language documents on the Web

PhD Candidate: Sen Xu
GeoVista Center, Department of Geography,
Pennsylvania State University, U.S.A.

August 19th, 2010 Portland

Outline

- Motivation
 - Methods
 - A case study
 - cardinal direction vs. relative direction usage in route directions
 - Work in progress
-

Motivation

- Sprouting volunteered geographic information on the WWW:
 - Human generated route directions: “how to get here”, “transportation & directions” from hotels, companies, churches...
 - Other text with spatial information: twitter, news stories (BP oil spill), ...
 - Limitations of human participant involved experiment methods
 - Expensive
 - Sample size limit
 - Spatial coverage limit
- } Scale
-

Screenshot of a route direction web page (from Nittany Inn website)

The screenshot shows a web browser window with the following content:

- Header: "Hotel Suite in State College PA : Pen..."
- Contact info: "Toll-Free: (800) 233-7505 · Phone: 814-863-5000 · Fax: 814-863-5002"
- Section: "Nittany Lion Inn"
- Map: A digital map showing the location of "THE NITTANY LION INN" in State College, PA. A callout box says "Click to Enlarge Map >>".
- Address: "200 West Park Avenue · State College, PA 16803-3598"
- Contact info: "Toll-Free: (800) 233-7505 · Phone: (814) 865-8500 · Fax: (814) 865-8501"
- Directions section:
 - Driving from New York City and the East:** "Take I-80 W in PA to Exit 161 (Bellefonte). Follow PA 26 S to US 220 S, and take Exit 74 for Innovation Park/Penn State University. For the Nittany Lion Inn: Stay in the left lane of the exit and follow the sign for Penn State University. This will become Park Avenue. Travel Park Avenue to the stop light at Allen Street. Turn left on Allen Street and then right on Fischer Road. The Nittany Lion Parking Deck will be on your left followed by the Nittany Lion Inn. For the Penn Stater Conference Center Hotel: Stay in the right lane of the exit and follow the sign for Innovation Park. Turn right at the end of the ramp onto the Park Avenue extension, and turn left at the stop sign onto Innovation Boulevard. The hotel is on the right."
 - From Philadelphia:** "Take the PA Turnpike/I-76 W..."
- Right sidebar:
 - "Visit Our Events CALENDAR"
 - "The Latest NITTANY NEWS! Click Here!"
 - "HOSPITALITY SERVICES GREEN LEADERSHIP"
 - "2009 Best of STATE COLLEGE" badge

Annotations on the left side of the page:

- "Accompanied with digital map" points to the map area.
- "Address of destination" points to the address "200 West Park Avenue · State College, PA 16803-3598".
- "origin" points to the "Driving from New York City and the East:" section.
- "landmarks" points to "Penn State University" in the directions text.
- "Relative directions" points to "Turn left at the stop sign onto Innovation Boulevard. The hotel is on the right."

Footer: "Read www.pshs.psu.edu"

BP oil spill

- This application allows you to add points with links to online photos, Web sites, and YouTube videos.

Analyze “region-of-interest”

The screenshot shows the Esri website interface for the Gulf of Mexico Oil Spill. At the top, the Esri logo and tagline "GIS Software that Gives You THE GEOGRAPHIC ADVANTAGE" are visible. Below the navigation menu, the main heading is "Gulf of Mexico Oil Spill". A sub-heading "Interactive Social Media Map" is followed by a brief description: "April 20, 2010, Transocean Ltd. reported an explosion and subsequent fire on board the semi-submersible drilling rig Deepwater Horizon leased and operated by BP. The incident has resulted in a massive oil spill and has been declared a Spill of National Significance by President Obama." Below this, a map of the Gulf of Mexico region is displayed, showing the coastline of Louisiana, Mississippi, Alabama, Georgia, and Florida. The map is overlaid with numerous blue location pins, each representing a point of interest. A red circle highlights a specific area on the map. On the left side of the map, there is a "Locate" search bar and an "Explore" panel with various layers and filters. The "Explore" panel includes options for News, Unlabeled, YouTube Videos, Tweets, Threatened Sites, Nautical Chart, Sensitive Habitat, Precipitation, Estimated Spill Areas, and Shared Content. The "Shared Content" panel allows users to add features like Web Link, Photo, Video, and News. At the bottom of the map, there is a social media sharing section with a "Like" button and a notification that "785 people like this. Be the first of your friends." The Esri logo is also present in the bottom right corner of the map area.

Advantages of collecting spatial (language) data from the Web

- ***Inexpensive***
 - ***Fast***
 - ***Easy accessibility***
 - ***Wide spatial coverage***
-
- WWW is a potential data source for spatial cognition study; an alternative for collecting data through human participant involved experiments
-

Challenges

- Data Collection - Web documents
 - High precision
 - Spatially unbiased
 - Data Analysis
 - Automatic process text corpora of large sizes
 - Geovisualization
-

Related Methodologies & Research

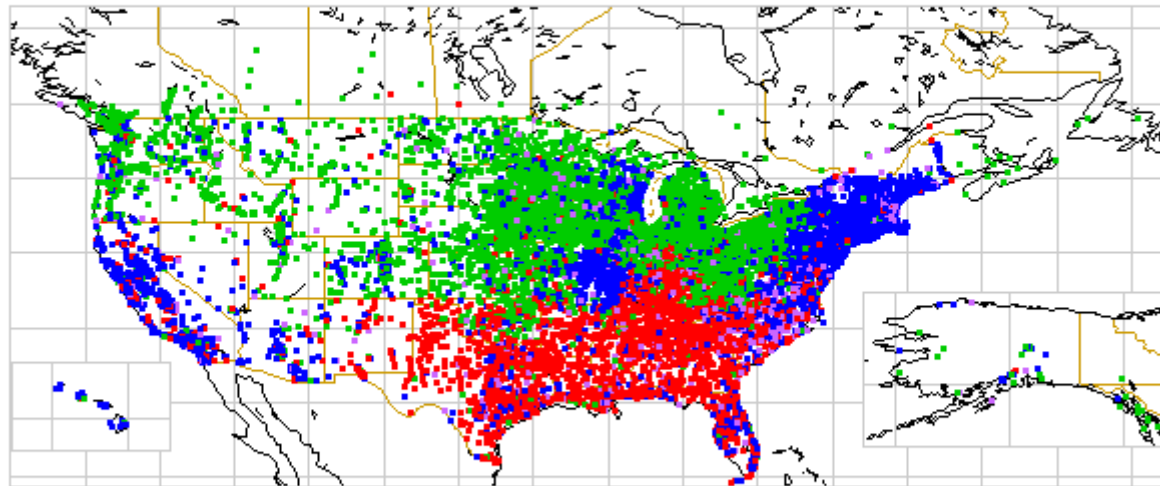
- Data Collection

- IR/KDD [Luhn, 1957]
- Crowdsourcing [Howe, 2006], [McConchie, 2002]
- Text classification

- Data Analysis

- Corpus linguistics
 - Text analytics
-

The Great Pop vs. Soda Controversy



Pop vs. Soda data as of October 3, 2002

■ "Pop" ■ "Soda" ■ "Coke" ■ Other

[show all](#) [pop](#) [soda](#) [coke](#) [other](#)

Your name: (optional)

Your e-mail address: (optional)

What generic word do you use to describe carbonated soft drinks? (Note that these could be of any brand or type, Coca-Cola, Pepsi, 7-Up, etc. We are concerned with the overall word, not a specific brand.)

- Pop
 Soda
 Coke
 Other (please specify word)

Please remember to give the city, state and zip code of your *home town*, the town where you learned the dialect of English you speak, even if it is not where you live now.

Your hometown:

State/Province: Zip/Postal Code:

Forgot the zip code of where you grew up?

These sites will tell you the zip/postal code for any address:

[United States Postal Service](#)

[Canada Post](#)

A Case study

—

Regional variation in cardinal vs. relative
direction usages in route directions

Observation

- Difference exists in spatial language choices in route directions
 - “go ahead on Atherton Street, turn *left* at the traffic light”
 - “go *north* on Atherton Street, turn *west* at the traffic light”
 - Individual, group, or sex-related difference has been investigated on the difference stated above (Montello et al., 1999, Ishikawa and Kiyomoto, 2008)
 - Regional difference has also been noted (Davies and Pederson, 2001)
-

Development of methodology

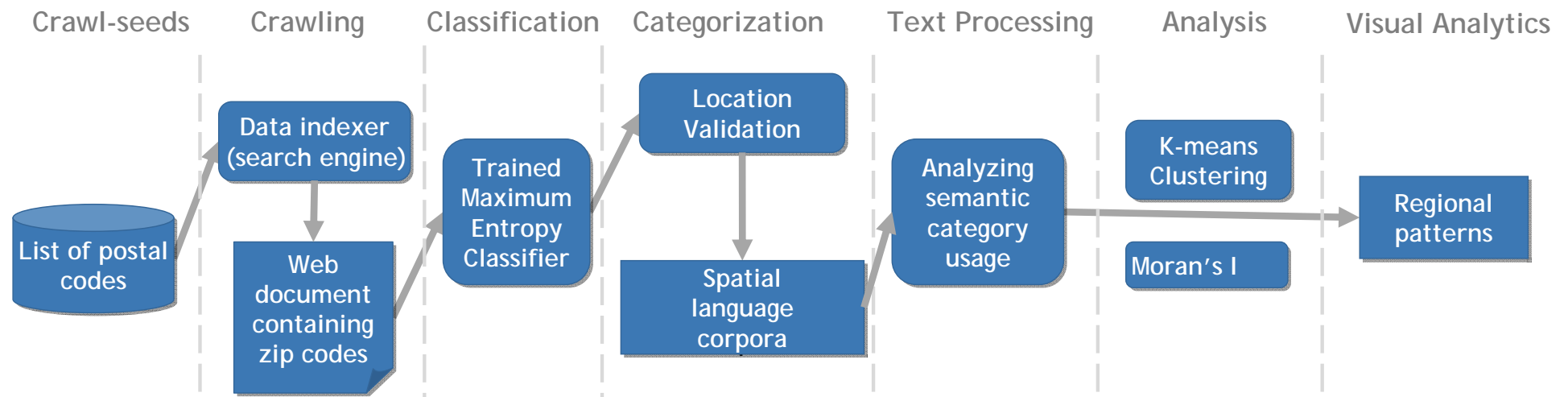


Figure 2. Overview of the methodology for building and analyzing the SARD Corpus

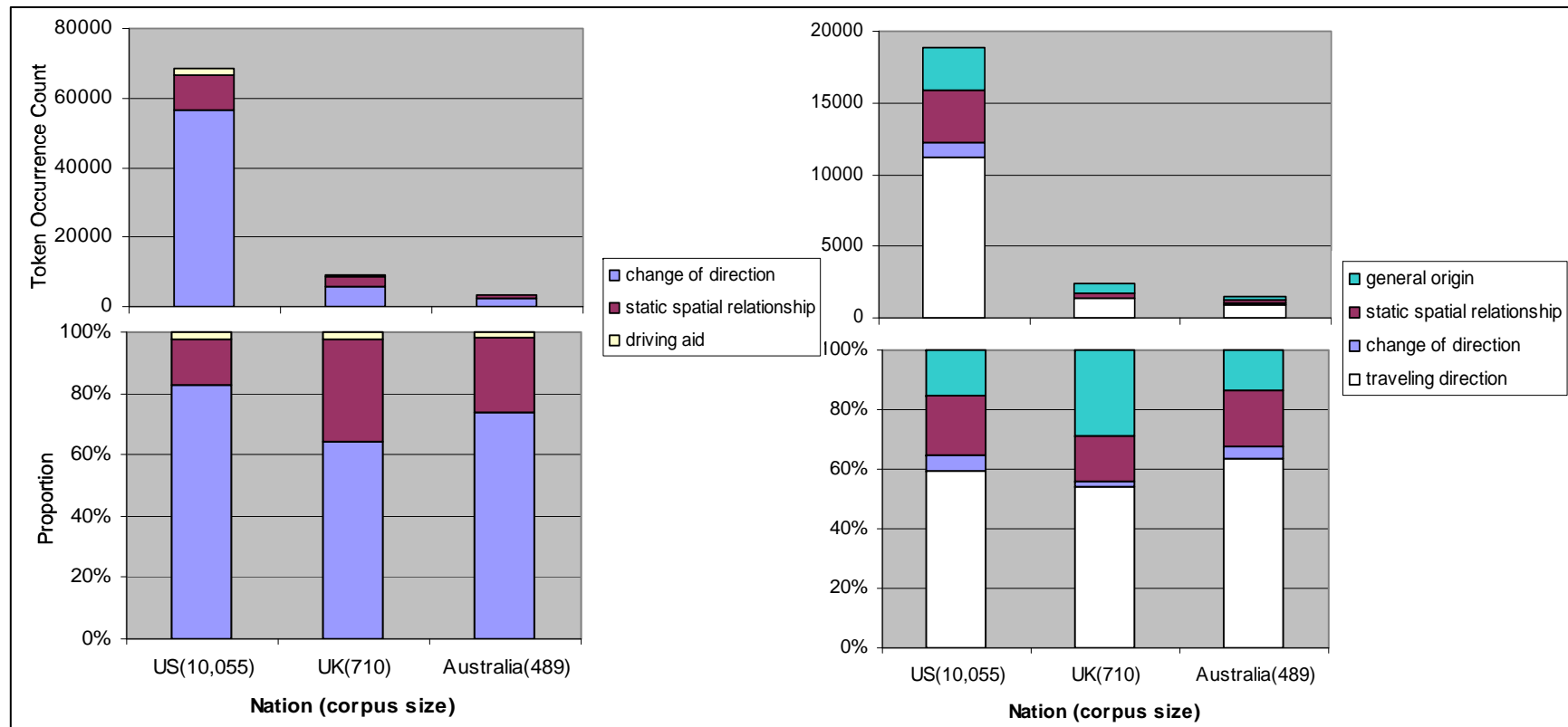
Attributes of the SARD Corpus (Spatially-stratified Route Direction Corpus)

Attributes	Value
Corpus topic	Route Directions
Document format	HTML
Language	English
Data source	WWW
Spatial coverage	The U.S., the U.K. and Australia
Size (total of documents)	11,254 documents (10,055 in the U.S., 710 in the U.K., and 489 in Australia)
Size (mega-byte)	203 MB
Percentage of true route directions	93%
Organization	Nation — Postal Region — Postal code
Other	does not contain documents with postal codes from different postal regions.

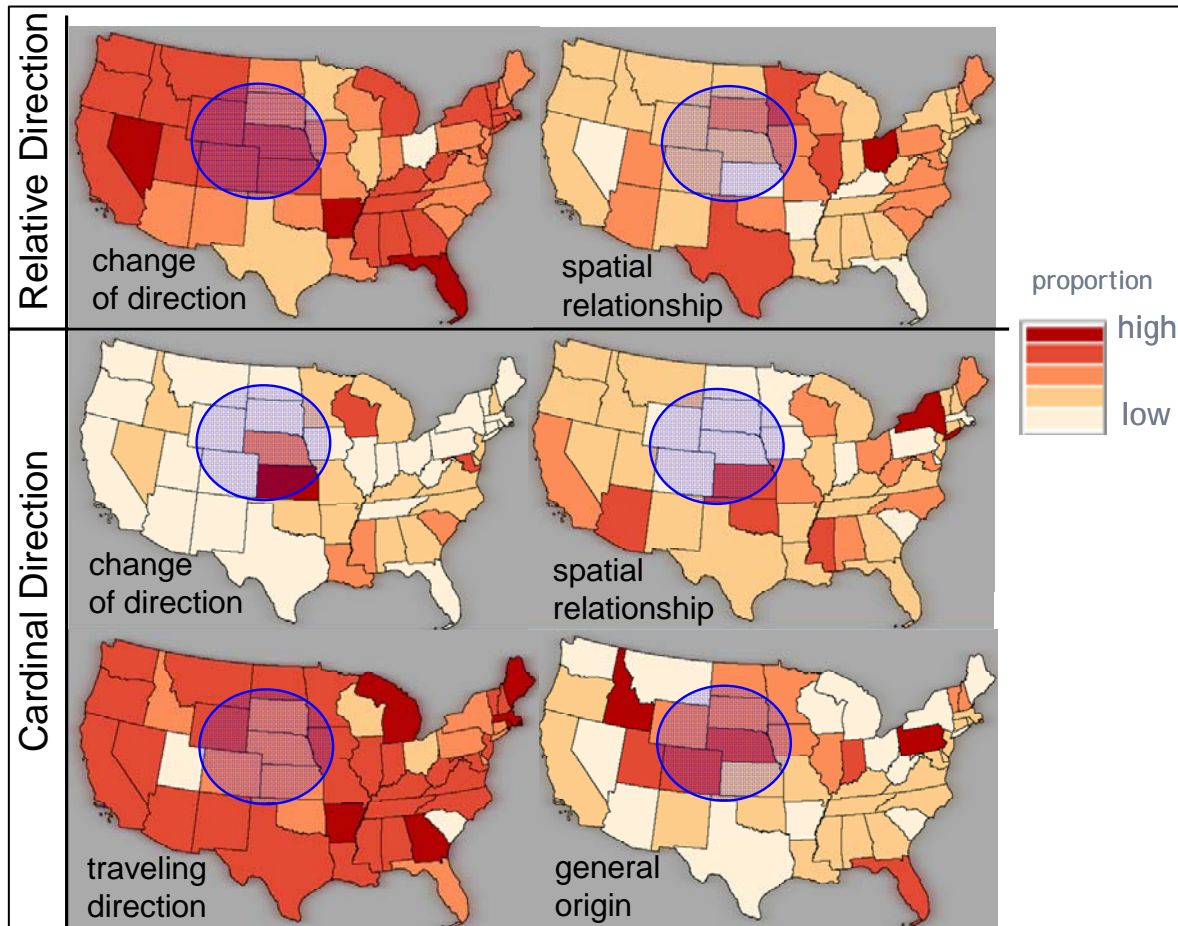
Semantic categories for cardinal and relative directions

	Semantic categories	examples
Relative Direction	1. Change of direction	take a left, bear right
	2. Static spatial relationship	see a landmark on your right, the destination is left to a landmark
	3. Driving aid	keep to the left lane, merge to the right lane
Cardinal Direction	1. Traveling direction	head north, traveling south
	2. Change of direction	veer southwest on US Hwy 24, turn north
	3. Static spatial relationship	2 blocks east of landmark
	4. General origin	from North, coming from South of New York
	*used in POI names	<u>North Atherton Street</u> , <u>West Street</u> .

National level histogram of relative direction (RD) (left) and cardinal direction (CD) (right) usages (Top: token occurrence count, Bottom: Proportion)



Region-level comparison of RD and CD usages in the U.S.



Case study - discussion and conclusion

- Regional variations: statistically significant
 - Pros and cons:
 - Data source: relatively unbiased
 - Route directions online
 - Postal code based data collection sacrifice certain types of route directions
 - Scale of analysis
-

Work in Progress

- The GeoCAM project
 - Representing, extracting, mapping, and interpreting movement references in text
 - More analysis possibilities:
 - Mode-of-transportation
 - Scale
 - Use of landmarks
 - Comparison of Human-generated vs. machine-generated route directions
-

References

- [Chen et al., 2007] Chen, J., MacEachren, A. M., and Guo, D. (2007). Visual inquiry toolkit - an integrated approach for exploring and interpreting space-time, multivariate patterns. Technical report, GeoVista Center and Department of Geography Pennsylvania State University, Department of Geography University of South Carolina.
 - [Davies and Pederson, 2001] Davies, C. and Pederson, E. (2001). Grid patterns and cultural expectations in urban wayfinding. In Montello, D., editor, *Spatial Information Theory*, volume LNCS 2205, pages 400–414, Morro Bay, CA, USA. Springer.
 - [Goodchild, 2007] Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.
 - [Ishikawa and Kiyomoto, 2008] Ishikawa, T. and Kiyomoto, M. (2008). Turn to the left or to the west: verbal navigational directions in relative and absolute frames of reference. In Cova, T. J., Miller, H. J., Beard, K., Frank, A. U., and Goodchild, M. F., editors, *Geographic Information Science*, volume LNCS 5266, pages 119–132, Park City, UT, USA. Springer.
 - [Jaiswal, 2010] Jaiswal, A. R. (2010). Automatic semantic tagging of transportation mode for digital documents. Technical report, College of Information Science and Technology, Pennsylvania State University.
 - [Montello et al., 1999] Montello, D. R., Lovelace, K. L., Golledge, R. G., and Self, C. M. (1999). Sex-related differences and similarities in geographic and environmental spatial abilities. *Annals of the Association of American Geographers*, 89(3):515–534.
-

Questions & Comments?
