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
  \[ \text{Scale} \]
Screenshot of a route direction web page
(from Nittany Inn website)

Accompanied with digital map
Address of destination
origin
landmarks
Relative directions
BP oil spill

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

Analyze “region-of-interest”
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

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: Not Applicable  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  

Send  Clear
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)

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

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


Questions & Comments?