Visualizing Unstructured Text Documents using Trees and Maps: Analyzing Verbal Directions

Ian Turton, Alan M. MacEachren
GeoVISTA Center, Pennsylvania State University,
University Park, PA, 16802
{ijt1, maceachren}@psu.edu
Summary

- What did we do?
- Why did we do it?
- How did we do it?
- Did it work?
- What did it look like?
- What we plan to do next
Introduction

- Implicit geographic references are common.
- Text sources provide potentially important geographic information.
- If and only if we can extract the implicit geography and make it explicit.
Why bother?

While humans are quite good at determining the context needed to interpret text, the web provides orders of magnitude more potentially relevant documents that there are human analysts to process.
Methods

- Large system made of sub modules:
  - Route detection
  - Tokenization
  - Landmark extraction and identification
  - Route Sketching

- Tools discussed today
  - *TermTree* aids route detection and tokenization
  - *RouteSketcher* displays landmarks on map
Output of Route Detector

<Routes>
<Route>
......
</Route>
<Route>
<Destination>
Directions to Reproductive Science Institute &amp;
Valley Forge Surgical Center 945 Chesterbrook Blvd. Chesterbrook, Pennsylvania 19087
</Destination>
<Origin>
From Philadelphia, Schykill Expressway West ( I  76)
</Origin>
<RouteDesc>
Take I-76 West ( toward Valley Forge ) Take Exit #328A to 202 South Take the Chesterbrook Exit to
stop light Turn right onto Chesterbrook Blvd. &amp ; merge into the left lane At first light, Duportail Road, turn left Make
the first left into parking lot Building is on the left.
</RouteDesc>
</Route>
......
</Route>
......
</Route>
......
</Routes>
Tokenizing the Route

- Start with destinations, origins and then route descriptions
- Within each segment work backwards looking for keywords
  - Street, road, avenue etc
  - Numbers (zipcode, phone number)
  - Numbers + direction (80 East)
  - Route, highway + number
  - Airport, bridge, river, park
  - Other proper nouns
If we find a keyword

- Lookup road names in street database.
- Lookup telephone numbers at whitepages.com
- Lookup zipcodes in geonames.org DB
- Lookup proper nouns and other locations in geonames.org DB
- If found replace tokens with special place token.
<Whitespace />
<Word>Institute</Word><Whitespace />
<Whitespace />
<Punctuation>&amp;</Punctuation><Whitespace />
<Whitespace />
<City> state='PA'>Valley Forge</City>
<Whitespace />
<Word>Surgical</Word><Whitespace />
<Whitespace />
<Word>Center</Word><Whitespace />
<Whitespace />
<Number>945</Number><Whitespace />
<Whitespace />
<Street state='PA'>Chesterbrook Blvd</Street><Punctuation>
</Punctuation><Whitespace />
<City state='PA'>Chesterbrook</City><Punctuation><Punctuation>.</Punctuation>
</Punctuation><Whitespace />
<State abbr='PA'>Pennsylvania</State><Whitespace />
<Whitespace />
<ZipCode>19087</ZipCode><Whitespace />
<Whitespace />
</Destination>
Displaying Routes as Trees
TermTrees

- Root at user selected term or phrase
- Branches to left and right show words preceding and following that term within the document(s).
- Use standard regular expressions to build searches
Wild card patterns
TermTree: analysis of terms in linguistic context – prepositions
TermTree: analysis of terms in linguistic context - prepositions
TermTree: analysis of terms in linguistic context - prepositions
Benefits

- Improved semantic matching of route detection
- Understanding of variations in term use in linguistic context, and frequency of different variants (e.g., on and onto)
- Supports rapid exploration of term use, including comparison among terms
- Feeds into design of algorithms for landmark detection
Displaying Routes as Maps

- Landmarks detected computationally in previous section are geocoded using the same databases as before.
- Where multiple places (or streets) match all are chosen at this stage.
- These are then passed to a JavaScript based mapping system.
Directions to Reproductive Science Institute & Valley Forge Surgical Center 945
Chesterbrook Blvd, Chesterbrook, Pennsylvania 19087

From Philadelphia, Schuylkill Expressway West (I-76)

Take I-76 West (toward Valley Forge) Take Exit #328A to 202 South Take the Chesterbrook Exit to stop light Turn right onto Chesterbrook Blvd. & merge into the left lane At first light, Duportail Road, turn left Make the first left into parking lot Building is on the left.

Directions to Reproductive Science Institute & Valley Forge Surgical Center 945

-74.46119, 37.22754
Pan, Zoom and Highlight

Directions to Reproductive Science Institute & Valley Forge Surgical Center 945 Chesterbrook Blvd. Chesterbrook, Pennsylvania 19087

From Philadelphia, Schuykill Expressway West (I-76)

Take I-76 West (toward Valley Forge) Take Exit #328A to 202 South Take the Chesterbrook Exit to stop light Turn right onto Chesterbrook Blvd. & merge into the left lane. At first light, Duportail Road, turn left Make the first left into parking lot Building is on the left.

Directions to Reproductive Science Institute & Valley Forge Surgical Center 945 Chesterbrook Blvd. Chesterbrook, Pennsylvania 19087

-75.16157, 39.98510

Scale = 1 : 865K
Direction to Reproductive Science Institute & Valley Forge Surgical Center 945
Chesterbrook Blvd, Chesterbrook, Pennsylvania 19087

From Philadelphia, Schuykill Expressway West (I-76)

Take I-76 West (toward Valley Forge). Take Exit #328A to 202 South. Take the Chesterbrook Exit to stop light. Turn right onto Chesterbrook Blvd. & merge into the left lane. At first light, Duportail Road; turn left. Make the first left into parking lot. Building is on the left.

Directions to Reproductive Science Institute & Valley Forge Surgical Center 945

-75.45505, 40.06892
Benefits

- Allows analyst
  - to determine which landmark or place is the correct one.
  - Follow the route on the map
  - Spot errors in geocoding
Future Work

- Improve interface for user feedback
- Improve links between text and map
- Improve links between TermTree and RouteSketcher
- Support highlighting of whole route + on demand highlighting by user query.
- Add capability to learn from user feedback, thus use feedback to improve the system rule-base
- Conduct formal test of interpretation precision (the proportion of identified entities that are identified correctly) and recall (proportion of entities that exist in the text that are found)
Conclusions

- Initial results are very promising.
- It is possible to automatically extract routes from unstructured text documents.
- Landmarks can be detected algorithmically.
- Once geocoded the display of landmarks on an interactive map allows users unfamiliar with the region to explore the routes mentioned in the document.