Geographic Analysis of Linguistically Encoded Movement Patterns – A Contextualized Perspective

Alexander Klippel[1], Alan MacEachren[1], Prasenjit Mitra[2], Ian Turton[1], Xiao Zhang[2], Anuj Jaiswal[2], Kean Soon[1], Jared Oyler[1], Rui Li[1]

[1]GeoVISTA Center, Department of Geography, The Pennsylvania State University, PA, USA

[2]Information Science and Technology, The Pennsylvania State University, PA, USA

In this paper, we present an approach and outline the initial steps toward a system to support theoretically-informed, analyst-guided, extraction and contextualization of statements about movement from text. Linguistic geographic references in various text corpora (e.g. web blogs) provide potentially important information about movement of entities (people, vehicles, etc) and about their underlying spatial behaviors. These qualitative geographic references can only be interpreted *if put in an appropriate context*. While progress has been made on geographic information retrieval from text, the progress has been relatively slow and has focused primarily on extracting and disambiguating place names. While that is an important and sometimes hard task (e.g., there are 1042 instances of the name "Columbia" in the Geo-graphic Names Information System), place name extraction is just a small part of the challenge.

Our focus is placed on the analysis of movement patterns characterized linguistically. The initial focus is on route directions provided on the web. This comparatively restricted domain allows us to create a first system for the analysis (visually and conceptually) of movement patterns as characterized in natural language descriptions. A goal for the work reported is to demonstrate the effectiveness of our approach before extending it into other domains (such as adding a temporal component, multiple agents, etc.). We will characterize the basic components of our proposed analytical framework briefly.

Extracting route directions

Classically, route directions contain an *origin*, a *destination*, and a *body of route information*. The flexibility of natural language, however, makes it possible to 'hide' information regarding, for example, origin and destination 'somewhere' on a website, or to have one destination but several origins (or vice versa). Our first task therefore was to identify individual route directions on websites. For this purpose, we crawled over 11,000 thousand reference documents containing route directions using different key words. These documents were analyzed (for example using word trees and document stemming) to define regular expressions that allow for identifying individual route directions.

Mapping geographic way marks

Once the route directions are extracted and stored as an XML file, a second component of our system parses them to identify *geographic way marks* that can be displayed on a map for an analyst to inspect. First, origin, destination and route direction body are tokenized. Second, the algorithm passes through the token stream extracting geocodable tokens such as zip codes and telephone numbers that can be recognized by using a simple regular pattern matching expression. Third each token is matched against a database of street name suffixes generated from the US Census TIGER dataset (e.g., Road, Blvd, St, etc.). Fourth, the proper street names are identified and looked up using the GeoNames system (geonames.org). Finally, a pass is made through the token stream to extract addresses which are defined as all the tokens between a number followed by a street token until a zip code token. Addresses are geocoded at geocoder.us, populated places at geonames.org and roads are resolved using a local database of the Open Street Map data.

Finally, the token stream is then passed to a display renderer, which includes a text window linked dynamically to a map window. Brushing over a geographically encoded text token highlights the related map element, allowing the analyst to see if the system selected the right place or to choose the correct one where multiple places have been returned.

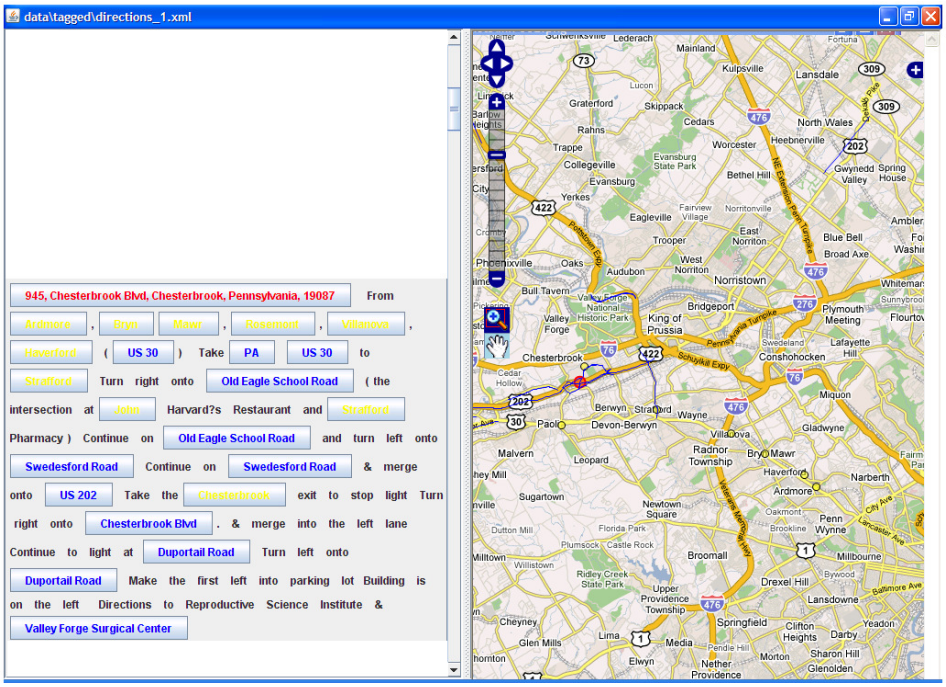

Fig. 1. Screen shot of tagged route directions (left) and there mapping (right, multi-road example).

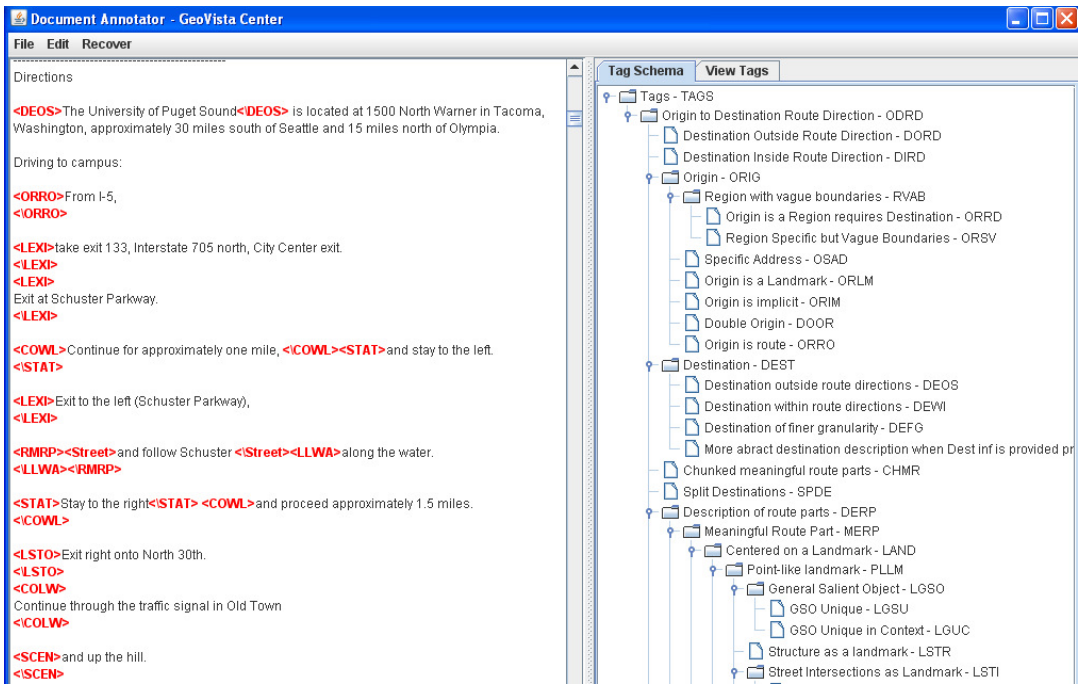

Fig. 2. Screenshot tagging tool.

An ontology of meaningful units of route parts

So far we have discussed the identification of web documents containing route directions, an initial extraction of directions and parsing into origin, destination, and the body of route directions, and the identification of geocodable route parts that can be mapping to support iterative analysis of the linguistically encoded movement pattern and refinement of the route specified. To deepen our analysis and to pursue our long term goal of automatically extracting and geo-locating movement patterns, we randomly selected 30 documents (from the 11,000 thousand identified) to be hand tagged. To support this step, we developed a document tagging tool (Figure 2) that supports the creation and application of a taxonomy of route direction concepts, i.e. the conceptually primitive / meaningful route parts (as shown on the right part of Figure 2). This taxonomy of route direction elements is the basis for an ontological characterization and can be exported as an OWL file for further processing using ontology editors such as ConceptVISTA. We will analyze each meaningful route part as a *spatial scene* and formally characterize the spatial information the linguistic expressions provide, i.e. the characteristics of the route and the spatial environment it is embedded in. Linguistic underspecificity may be resolved using contextual information such as provided by the map.

Conclusions and Outlook

We present a first stage implementation of our approach that supports trial applications to validate effectiveness of system components in supporting the analysis of linguistically characterized movement patterns. Our next step is to implement an initial, integrated, visual workbench for analysis of linguistic movement patterns. We will use the workbench to compare different visual-computational tools to each other and to the base case of manual text analysis.

A further aspect of validation focuses on the more theoretical aspects of movement patterns. The created ontology that is developed and used to identify conceptual primitives in route directions will be tested in an inter-rater agreement task by having multiple users annotate a randomly selected test set of websites containing route directions. We expect the result to confirm the cognitive validity of our approach, i.e. that the identified route parts are indeed meaningful.

Our long term goal focuses on refining each part of our analytical framework to address two specific objectives: (1) build the conceptual/ theoretical/ data model framework needed to represent, extract, map, and interpret geographic accounts of movement found in text; and (2) apply the

framework to creating methods, tools, and a geovisual analytics workspace to accomplish these goals.

Acknowledgements
Research for this paper was funded by the National Geospatial-Intelligence Agency/NGA through the NGA University Research Initiative Program/NURI program. The views, opinions, and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the National Geospatial-Intelligence Agency or the U.S. Government.