From Data Collection to Analysis – Exploring Regional Linguistic Variation in Route Directions by Spatially-Stratified Web Sampling

Sen Xu[1], Anuj Jaiswal[2], Xiao Zhang[3], Alexander Klippel[1], Prasenjit Mitra[2] and Alan MacEachren[1]

[1] GeoVista Center, Department of Geography, Pennsylvania State University, U.S.A.
[2] College of Information Science and Technology, Pennsylvania State University, U.S.A.
[3] Department of Computer Science and Engineering, Pennsylvania State University, U.S.A.

Abstract. How spatial language varies regionally? This study investigates the possibility of exploring regional linguistic variations in spatial language by collecting and analyzing a Spatially-strAtified Route Direction Corpus (SARD Corpus) from volunteered spatial language text on the Web. Because of the fast content sharing functionality of the World Wide Web, it quickly becomes a hotbed for volunteered spatial language text, such as directions on hotels' Websites. These route directions can serve as a representation of everyday spatial language usage on the WWW. The spatial coverage and abundance of the data source is appealing while collecting and analyzing large quantities of spatially distributed data is still challenging. Through automated crawling, classifying and geo-referencing web documents containing route directions from the web, the SARD Corpus has been built covering the U.S., the U.K. and Australia. We implement a semantic categorical analysis scheme to explore regional variations in cardinal versus relative direction usages. Preliminary results show both similarity and differences at national level and geographic patterns at regional level. The design and implementation of building a geo-referenced large-scale corpus from Web documents offers a methodological contribution to corpus linguistics, spatial cognition, and the GISciences.

Keywords: Spatial language analysis, volunteered spatial information, geo-referenced web sampling, regional linguistic variation, cardinal directions

1 Introduction

Spatial language is an important medium through which we study the representation, perception, and communication of spatial information. Research has approached spatial language from various perspectives. From the cognitive perspective, research has focused on group or individual differences, on how language affects way-finding behaviour, or on how regional context affects spatial language usage. From the computational perspective, modelling and reasoning has been applied to spatial language interpretation. The spatial language samples used in these studies have been mostly collected by individuals via time consuming experiments or interviews. This data collection method could provide samples that offer understanding on small-scale phenomenon through manual interpretation by analysts.

However, studying the regional linguistic patterns in spatial language—such as regional variations in route directions—requires a spatially distributed corpus. Spatial language data available from the WWW has great potential for this study because of its unrivaled coverage and easy accessibility. For example, it is common to find hotels, companies and institutions offering route directions on their website which provides spatial way-finding instructions to travelers from different places. Harnessing these human generated route directions on-line and analyzing them is the major focus of this study.

2 Methods

To harness route direction documents from the WWW and ensure the spatial coverage of the resulting corpus, a data collection scheme involving web crawling, text classification, and geo-referencing has been developed. Computational tools have been applied for assisting processing the Spatially-strAtified Route Direction Corpus (SARD Corpus) and interpretation of the results.

Collecting route direction documents from the WWW has two main challenges. First, route directions have a high linguistic complexity that makes it difficult to separate the route direction documents from a variety of irrelevant web documents. This challenge can be solved by applying a machine learning algorithms for text classification [1]. The precision of this route direction document classifier used in this study reaches 93% (from 438 positive classified documents, 407 are hand examined to be spatial language documents). Second, exploring regional variation in spatial language usage requires geo-referencing each document in the corpus, which is not an easy task (i.e., Geographic Named Entity Disambiguation). However, postal code, which commonly appears in destination addresses in route directions, can be used to coarsely geo-reference a route direction document on a postal code level. The data collection scheme first utilizes lists of postal codes for crawling web documents. The returned web documents are fed into the route direction classifier, where only positively classified route direction documents are stored in the result corpus. This data collection scheme maximizes the spatial coverage of the SARD Corpus at a postal code level. To prepare the corpus for extracting region linguistic attributes, the SARD Corpus is organized first by nation, then by region (states in the U.S. and Australia, postal district in the U.K.).

The data analysis of spatial language usage in route directions focuses on the regional linguistic variation, which is addressed by analyzing the semantic usages of cardinal directions (i.e.: *north*, *south*, *east*, *west*, *northeast*, *northwest*, *southeast* and *southwest*) and relative directions (i.e., *left* and *right*). The semantic categories used are detailed in Table 1. The scale and size of the corpus makes corpus linguistic tools a necessity for processing the regional linguistic characteristics. The TermTree tool [2], which is a text processing tool with the capacity to handle regular expressions, is used for assisting an analyst to manually evaluate the semantic usages of direction terms. The semantic categorical data is considered regional linguistic characteristics for each region in the SARD Corpus. Visual Inquiry Toolkit [3] is used for geovisualization of the regional linguistic characteristics (Fig. 3) to interpret the analysis result.

Table 1. Semantic categories for cardinal directions 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. **Change of direction**	*head north, traveling south*
	2. **Static spatial relationship**	*veer southwest on US Hwy 24, turn north*
	3. Traveling direction	*2 blocks east of landmark*
	4. General origin	*from North, if coming from South of New York*
	*used in POI names	*North Atherton Street, West Street.*

As a result of the data collection, the SARD Corpus has been built with 11,254 web documents covering the U.S., the U.K., and Australia. Overview of the workflow is presented in Fig. 1.

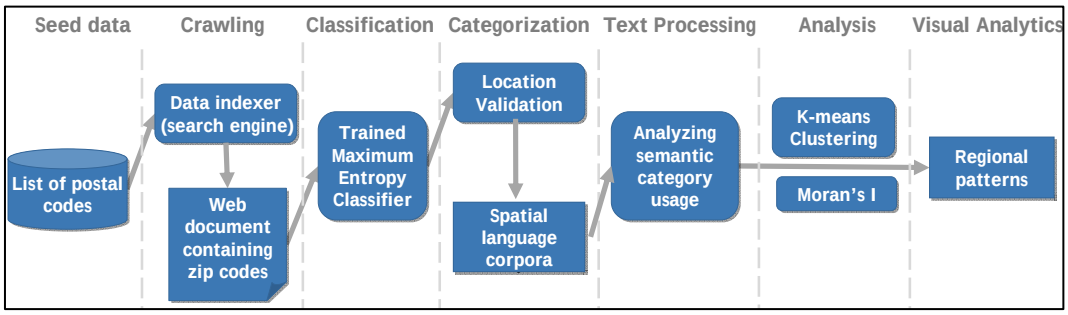

Fig. 1. Overview of the data collection and analysis schemes for building and analyzing the SARD Corpus

3 Results

Regional pattern analysis demonstrates how cardinal/relative directions usage varies at both national level (Fig. 2) and regional level (Fig. 3). On a national level, relative directions in all three nations are mostly used to represent "change of direction" (the blue bar on the left). Similarly cardinal directions are mostly used to represent "travelling direction" (The white bar on the right). On the other hand, the preference for relative direction when representing "change of direction" is much more common in the U.K. than in the U.S. and Australia. Correspondingly we find that cardinal directions are used more often in the U.S. and Australia than in the U.K. (the blue bars on the right) to represent "change of direction".

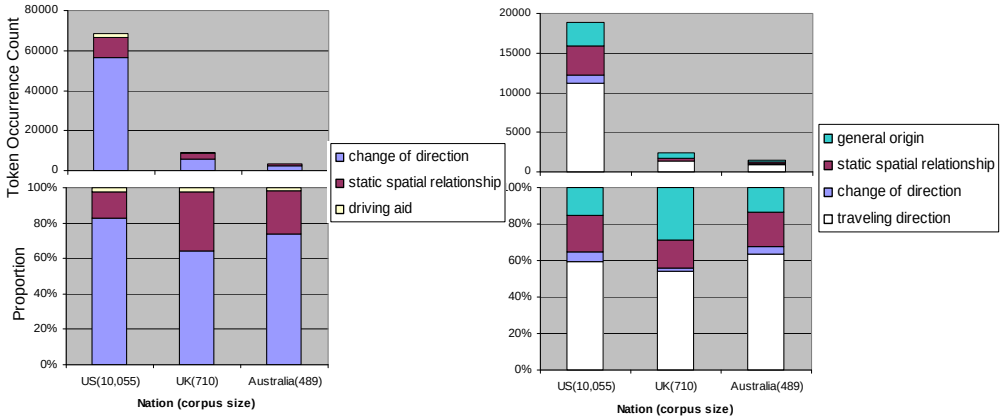

Fig. 2. Nation-level comparison of relative directions and cardinal directions usage

To get a better understanding of the regional variation of relative versus cardinal direction usages, the proportion of each semantic category is plotted on a map for comparison. The plotted map can provide geographical knowledge about the regions, such as adjacency, which helps the analyst to detect regional patterns. Fig. 3 shows that the two most dominant usages as noted at the national-level (relative directions used for "change of direction", cardinal directions used as "travelling direction") are used more frequently in most states in the U.S. For cardinal direction usage, there is a geographic pattern (South Dakota to Kansas, Wyoming to Iowa, blue circle) that differs from its surroundings states in every semantic category. The regional pattern detected is comparable to the Colorado West and Central West region in the map of U.S. dialect [4, p.186]. A possible explanation for this observation may lie in the correlation between the regional linguistic preference and regional geographical features, which is yet to be investigated.

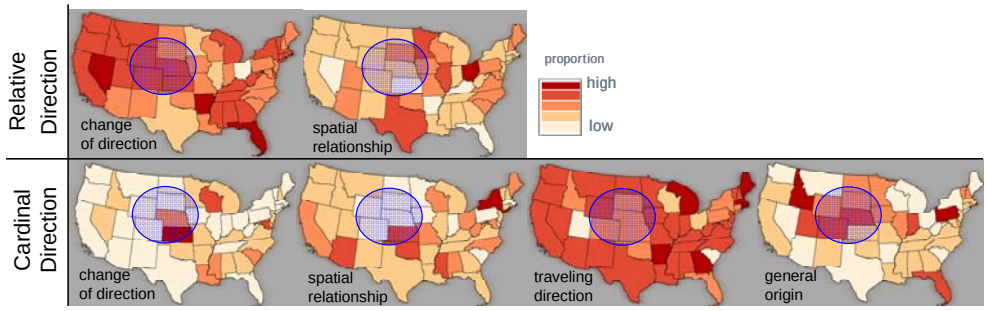

Fig. 3. Regional-level comparison of relative directions and cardinal directions usage (the U.S.).

4 Summary

This paper presents a first step toward an effective and scalable data collection method for spatial language study. It enables spatial cognitive researches to scale-up the spatial language data sets and answer spatial cognitive questions (such as the regional spatial language difference) at a large scale. This study shows promise for effective spatial cognitive research through processing and analyzing volunteered spatial language data, which is an alternative compared to collecting data by designing human participant involved experiments. The presented workflow can also be extended to languages other than English to assist in cross-language comparisons.

The language preference at the nation-level and region-level are both explored, offering 1) a better understanding of how people tend to use spatial language to communicate spatial information; 2) how people differ in using spatial language from different regions; and 3) a guideline to develop a localized, use-specific natural language generation system for navigational devices. Regional patterns of cardinal and relative direction usages in route directions are observed and analyzed, offering a novel perspective for spatial linguistic studies. The design and implementation of building a geo-referenced large-scale corpus from Web documents in this study offers a methodological contribution to corpus linguistics, spatial cognition, and GISciences.

5 Acknowledgement

Research for this paper is based upon work supported 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.

References

[1] Zhang, X., Mitra, P., Xu, S., Jaiswal, A.R., Klippel, A., MacEachren, A.M.: Extracting route directions from web pages. In: Twelfth International Workshop on the Web and Databases (WebDB 2009), Providence, Rhode Island, USA. (2009)

[2] Turton, I., MacEachren, A.: Visualizing unstructured text documents using trees and maps. In: GIScience workshop, Park City, Utah (2008)

[3] Chen, J., MacEachren, A.M., Guo, D.: 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 (2007)

[4] Smith, J.: Bum bags and fanny packs: a British-American, American-British dictionary. Carroll & Graf Publishers (2006)