Sampling volunteered spatial language documents on the Web

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

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

Abstract. Spatial language is one important material for spatial cognitive studies, where usually participant involved experiments are carried out to collect data on a small scale. Because of its convenience for information sharing and retrieval, the World Wide Web (WWW) becomes a hotbed for volunteered spatial language data, such as "Directions" on hotel websites. Retrieving such web documents would enable a scaled-up spatial language corpora, which could provides material for spatial cognitive study on a large scale. A data collection method based on web crawling and text classification is designed; two type of crawling schema (keywords, zip codes) are tested and compared. Implementing the zip codes crawling schema, spatial language corpora for the US, UK, and Australia are built with higher than 93% accuracy. Potential research direction is discussed.

Keywords: Volunteered geographic data, corpus collection, geo-referenced web sampling

1 Introduction

Spatial language, as one important medium through which we study spatial cognition, has been collected primarily by individuals via time-consuming human participant involved experiments or interviews. Rapidly growing on-line map services has attracted users to provide spatial language on-line, which poses a new data source for spatial language study. Collecting spatial language data from the WWW have the great advantages: *inexpensive*, *fast*, most importantly, with easy *accessibility* of the volunteered spatial language documents from all over the globe. Research questions such as exploring regional variances in spatial language can be studied with a large quantity, high spatial coverage spatial language corpus. The capacity of collecting massive amount of text and filter text that carries spatial information can also contribute for intelligence purposes. However, spatial language web document are mixed with a great variety of other types of web documents. For example, machine-generated route directions (text output from google map's direction search) is very hard to be distinguished from human generated route directions, while spatial cognitive study is only interested in the latter.

2 Methods

Data collection schema is designed and tested as illustrated in Figure 1. By querying meta-data into web search engines, the returned hits greatly narrows down the documents set to be processed. To overcome the challenge of retrieving only spatial language text documents, a maximum entropy model based classifier with precision of over 93% was built to pick spatial language documents from the returned web documents [1]. After two steps of crawling and classification, spatial language corpora are built.

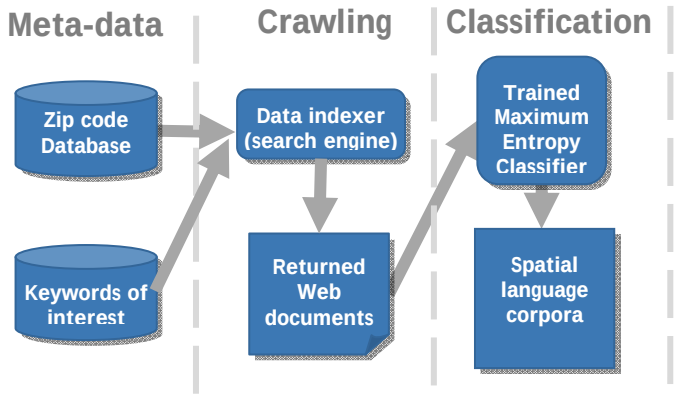

Fig. 1. Data flow chart for building spatial language corpus in the U.S.

Two types of meta-data are tested and compared. Keywords (for spatial language text with a focus on route directions, "direction, turn, mile, go, follow, take, exit etc.") is a simple crawling schema as the returned hits from web search engines usually contains a high volume of target documents. However, it would 1) introduce linguistic bias; 2) the more keywords being used, the more it will limit the variety of linguistic phenomenon in the returned documents, 3) it cannot guarantee the result corpora is spatially distributed. Crawling with zip code database is a convenient way to get vaguely geo-referenced data. Comparing with crawling with keywords, zip code introduces no linguistic bias as it only serves as a "geostamp" for addresses (which usually appears with spatial language texts). With a complete zip code database as meta-data, it offers maximum spatial coverage that WWW can provides on a zip code level.

3 Discussion

The large-scale, spatially-distributed spatial language corpora schema overcome the bottleneck of individual experiment data collection and provides a valid alternative for spatial cognitive studies. Using zip code database for crawling, we've build a spatial language corpora with over 10,000 documents from U.S., U.K, and Australia. The question of spatial referencing preference on national level and regional level are analyzed and compared [2]. 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.

4 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.

5 Reference

[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] Xu, S., Klippel, A., MacEachren, A., Mitra, P., Turton, I., Zhang, X., Jaiswal, A.: Exploring regional variation in linguistically encoded movement patterns - a case study. In: Annual meeting of Association of American Geographers - Cognition, Behavior, and Representation III: Methods, Las Vegas, U.S.A. (2009)