Why Geolocating Written Routes is Harder than it Looks

Ian Turton,

GeoVISTA Center
Walker Building
Pennsylvania State University
University Park
PA 16802
USA
Email:ijt1@psu.edu

1 Introduction

This paper seeks to explain why geolocating (or geocoding) a route description from a web page is a significantly more difficult task than other previously discussed problems in the geocoding domain. The canonical geolocation problem is the extraction and location of city names and countries (Turton et al. 2007, Ourioupina, 2002, Turton 2008, Guo et al. 2008). In this work the aim is to extract and map route descriptions from web pages. This requires the system to recognize streets and roads from the directions as well as determining the location of any addresses found. The task of drawing a route on a map can be broken down into three steps: extraction, disambiguation and mapping. This paper concentrates on the extraction and disambiguation steps as the mapping of a route is a entirely separate process.

Leidner (2008, chap.3) carries out a thorough review of previous work in the field of GIR. Early work by Woodruff & Plaunt (1994) and Amitay et al. (2004) focused on determining the geographic focus of a document, by combining the geographic points of any locations they identified in the text. While that is an important and sometimes hard task (e.g., there are 1042 instances of the name "Columbia" in the Geographic Names Information System), place name extraction is just a small part of the challenge.

2 Problems

While the process outlined above sounds straightforward and even easy, it is far from easy to implement. Currently the GeoCAM project is concentrating on routes in the USA, while limiting the geographic scope of the search for locations helps in the case of cities for streets the USA is more complex than many comparable countries. In the USA roads are often referred to by just their main name with out the prefixes and suffixes that they are stored in the database, e.g. "Turn on to Atherton" as opposed to the more normal "Turn on to Atherton St." and hardly ever the "correct" - "Turn on to South Atherton St.". This means heuristics must be applied to determine if a street is being discussed (or George W. Atherton) as well as trying to guess whether North or South Atherton is meant, (for added problems Atherton St. runs neither north or south).

Another major problem that is encountered is that there are relatively few street names used in the US compared to the number of streets. As can be seen in Table 1 Main St is the most popular followed closely the "numbered" streets such as 2nd St, the next most popular category is tree types (showing a certain poverty of imagination amongst town planners). At number 14 in the table we see the start of another perennial favourite suburban builders people and places, Washington St (Pl, Ave, Blvd) can be either of course, and is often surrounded by other presidents or towns. The final popular grouping is streets named after something nearby such as the river (or water), the church or the park. All of these groups of street names can prove problematic to process, first the named entity extractor has to distinguish between a text about Maple trees and Maple St or Church occurring at the start of a sentence from "turn onto Church". Finally, the sheer number of streets with the same name must be dealt with (remembering that some are segments of the same street, while others are distinct).

Rank	Street Name	Count
1	Main St	12849
2	2nd St	8977
3	1st St	8093
4	3rd St	8027
5	4th St	6945
6	Oak St	6612
7	Elm St	6104
8	Pine St	6069
9	5th St	5677
10	Church St	5662
11	Maple St	5527
12	Walnut St	5041
13	6th St	4698
14	Washington St	4104
15	7th St	3927
16	N Main St	3535
17	Center St	3502
18	River Rd	3493
19	High St	3452
20	S Main St	3421
21	Cedar St	3405
22	North St	3322
23	8th St	3305
24	Park Ave	3305
25	Park St	3277

Table 1: The 25 most popular street names in the USA (from Open Street Map data)

The next problem is that many US street names are difficult to spot in a document as the system is required to distinguish between, for example, E St. NE and E School St and must even handle N N St.[1] or S S Dixon St[2]. There are also streets like the one in Colorado "N ¾ Road" (see Figure 2) (*obviously* a street between N Rd and O Rd.). Even when a street is unambiguously named and correctly referred to it may not match as it has several names (see Figure 1). For example I-99 near State College, PA can be described, correctly, as I-99, US 220 (with added N or S) or PA 150. As can also be seen in Table1 streets such as North St have to be compared to N Rd, this makes the normalization task especially hard.

1 See for example North N Street, Muskogee, OK 74403, USA or North N Rd, Broken Bow, NE 68822, USA

2 S South Dixon St, Milton, FL 32571, USA

Figure 1: Streets with multiple names.
Photography by Joe Mabel, licensed under GFDL

Figure 2: N ¾ Road, Loma, Colorado.
Photograph by Ian Turton, licensed CC BY-NC-SA 2.0

3 Solutions

There are a number of methods that can be used to solve these disambiguation problems, this project has chosen a geographic technique. The system attempts to determine the destination of the route (as this is usually the most specific) it looks for complete addresses, followed by telephone numbers and zip codes. It uses this location to resolve as many locations that have been extracted from the rest of the route description. The system works outward from the "known" location and finds the closest named street segment to the destination, this is then connected to the destination. The routing algorithm repeats this process until it can not make a connection, then the next nearest street name is selected and the process repeats.

Where a clear destination can not be determined the system attempts the same process but using the best match town as the destination. Ambiguous cities are resolved by choosing the group of places that has the smallest bounding polygon. Again once any sort of fixed point can be determined the system then works out from that point trying to fit streets.

4 Conclusions

This paper has outlined some of the problematic elements of route description and geocoding. While the system must recognize all the usual named entities normally worked with, such as countries and cities, it must also extract and locate streets and roads. This second task is much harder than for cities due to the larger size of the database as well as the number of repeated street names (even with in a State). A geographic technique to help resolve this natural language processing problem has been outlined. These techniques will be expanded on and explained in the full paper.

5 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.

6 References

Amitay, E. et al., 2004. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, pp. 280, 273. Available at: http://dx.doi.org/10.1145/1008992.1009040

Guo, Q., Liu, Y. & Wieczorek, J., 2008. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10), 1090, 1067. Available at: http://dx.doi.org/10.1080/13658810701851420.

Leidner, J., 2008. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*, Dissertation.Com. Available at: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/1581123841

Ourioupina, O., 2002. Extracting geographical knowledge from the internet. In *International Workshop on Active Mining, ACDM-AM*. Available at: http://www.coli.uni-saarland.de/~ourioupi/ouri_fin.pdf.

Turton, I., 2008. A system for the automatic comparison of machine and human geocoded documents. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*. Napa Valley, California, USA: ACM, pp. 24, 23. Available at: http://dx.doi.org/10.1145/1460007.1460012.

Turton, I., Gahegan, M. & Jaiswal, A., 2007. Geographic Information Retrieval from Disparate Data Sources. In *GeoComputation'07*. Available at: http://ncg.nuim.ie/geocomputation/sessions/3B/3B5.pdf.

Woodruff, A. & Plaunt, C., 1994. *GIPSY: Georeferenced Information Processing SYstem*, Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.4945