On the Accuracy of Online Geocoders

Dirk Ahlers* Susanne Boll†

Geocoding is the conversion of a textual description of a location to geographic coordinates. With online geocoders being freely available to researchers and practitioners alike, their influence on data quality needs to be estimated. To this end, we first describe the basics of address-based geocoding and accuracy issues. We then look at two of the most widely-used geocoders and provide analysis of their automatic geocoder results, discuss their quality, and present a correction methodology to increase the accuracy of geocoding, using both independent sources, heuristics on accuracy metadata and conflation techniques.

1 Introduction

Most geospatial Web applications demand the transformation of a textual description of a place into a geographic coordinate by a geocoding process, to support the mapping of data or user queries.

The major mapping providers also offer online geocoders to prepare textual data for mapping. They are provided free of charge and are readily available and are therefore a preferred source for researchers and practitioners alike. These geocoder services are used in a multitude of applications such as rapid prototyping, research tools, and the vast amount of geographic mashup services on the Web.

We look into the two most widely-used online geocoders, namely those of Google and Yahoo!, which are both freely available and have coverage for Germany. The aim is to identify and quantify inaccuracies as well as differences and similarities within the heterogeneous results to derive a correction method to reduce the overall error in address-level geocoding.

Our application scenario of geographic Web Information retrieval aims to identify and extract location references in unstructured Web pages and to provide spatial search

*OFFIS – Institute for Information Technology
†University of Oldenburg, susanne.boll@uni-oldenburg.de

capabilities on these documents [1], aiming at the high granularity of individual addresses [2]. To actually enable spatial search and analysis capabilities such as vicinity searches, the textual address has to be converted to a geographical coordinate by a geocoder. The retrieved data then is suitable for geospatial search applications and supports assistance to mobile users.

2 Related Work

In the field of geographic information retrieval, the issue of uncertainty is usually discussed with a focus on term disambiguation for placenames in unstructured documents [10], [3], [9]. In the context of IR, the challenge lies in correct identification and assignment of geospatial properties.

A complementary topic is the accuracy of the geographical coordinates that are assigned to the extracted information. [12] discusses the topic of geographic uncertainty, ranging from theoretical aspects over modelling, handling, and mapping up to data acquisition and positioning. [8] gives an overview of the state of the art in geocoding. [6] describes quality indicators but focuses on issues of address identification and parsing while the quality of the actual geocoding is only touched briefly. An analysis of geocoder error levels and an initial correction method can be found in [11]. [5] examines the positional errors of geocoders by the distances between entities. The authors compare automatically geocoded points with the actual positions of both parcel and house locations to derive individual measures of positional error. The distance between house and parcel centroid was found to be much lower than that of the geocoded position to both. [4] describe a correction methodology based on direct access to interpolating line-based reference data and additional external sources for parcel sizes and distributions.

A more detailed insight into how street names and house numbers can be assigned in an administrative process is available in [7] which also discussed different numbering schemes and practical implementations as well as noteworthy examples of street addressing.

3 Geocoder Accuracy

An address in itself is a hierarchical textual description of a certain place and can be geocoded to a geographical point within a small radius. We have discussed some of the issues of address recognition, identification, extraction and verification in our previous work [2]. The location granularity of full addresses is rather high at a building level and is then well useable by pedestrians or other mobile users as seen in Figure 1. Especially at such high granularity, small errors can easily accumulate and become significant. For a consumer of geocoding results, data quality is not always easily to be assessed and needs to be considered within the data processing.

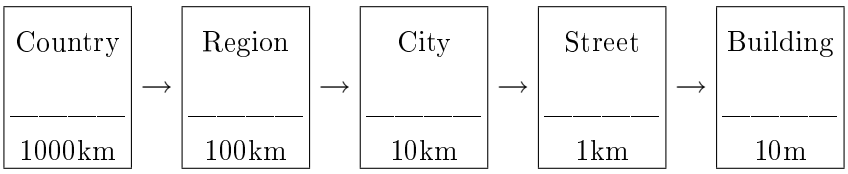

Abbildung 1: Spatial granularity levels by orders of magnitude

3.1 Requirements

The application scenario of pedestrian assistance has accuracy requirements at the granularity of individual addresses. This leads to the interesting question of what exactly makes a 'correct' and 'accurate' coordinate for a geocoded address.

This could be the building or parcel centroid, the entrance to the parcel or the building or point on the road nearest to any of the above. For navigational purposes, a connection to the road network is desireable. Therefore, following the street segment approach, we prefer to pinpoint addresses near the road instead of at the cadastral parcel centroid. However, to distinguish buildings, the clear identification of the side of the road is preferable to a location on the road centerline. Apart from the semantics of the point approximation of an extended geographic entity, we discuss two main requirements. One is the absolute positional accuracy as the congruence to the physical world. The other is the relational accuracy which should maintain spatial relations such as distance and direction between entities and buildings, keeping them clearly distinguishable.

3.2 Analysis

The area of study for the address-level geocoding is the city of Oldenburg in Northern Germany. Preprocessing was done by our validating address parser [2]. We had addresses geocoded by both the Google and Yahoo geocoder APIs which rely on different datasets from TeleAtlas and Navteq and also use different geocoders or at least different parameters. For initial analyses, we have mapped whole streets by iterating through all possible house numbers but the goal is to be able to accurately geocode a single address.

We compared the results by various statistical measures and by visualization and mapping for analysis of the disagreement of both geocoders. Furthermore, we used manual comparison with official cadastral data to estimate the absolute positional error by euclidean distance. For comparison with cadastral data and manual geocoding, we used the closed bitmap-based system of LGN [1]. The manual geocoding is pinpointed to the road-facing side of a building, which corresponds well to the geocoder's results based on the road network.

We additionally examine the accuracy annotations of the geocoders as a granularity and reliability measure. However, this does not capture uncertainty due to the geocoder's reference database, number of street segments, interpolation etc. Still, even this rough measure provides interesting insights into coverage and completeness of the respective

[1] Niedersachsen-Map, `http://www.niedersachsen-karte.de/`

reference databases. The two geocoders prove to have differing coverage, with no single source always preferable to the other. Similarly, the analysis of distances, positional errors and error distribution shows non-uniform character between the two sources.

We have found a variety of mismatches which in part can be traced or attributed to a range of issues. We find geocoding inaccuracies due to line simplification or interpolation errors as well as suspected mismatches between the reference data and the real world, such as missing city areas, out-of-date street directories, overlapping streets, missing street segments, or non-existing house numbers, to name just a few.

For some cases, correct and mismatched results can have very similar properties which are hardly detectable without grounding to the real world. However, in many other cases, outliers or inaccuracies can be detected within street data and can then be used to judge the reliability of that data. Therefore, the challenge is to distinguish and rank results based on their derived accuracy and reliability to support a usage decision and to further provide improved data.

4 Correction Methodology

Correction of inaccuracies in the data sources proves very demanding, as no obvious feature would give an estimate of the correctness or completeness of the geocoder for a specific street without resorting to a ground truth. By combining both online geocoders, strengths of the individual sources can be exploited while their limitations can be alleviated. We can then realise a method to reduce the error and enhance the accuracy.

There is no complete disclosure of the inner workings of the geocoder or the reference databases, but an accuracy indicator for a returned position is usually part of the result. It is very commendable that the APIs provide this accuracy information with the results which is a good step towards reliable quality metadata annotation. However, the accuracy described in Table 1 only gives a granularity measure (cf. Figure 1). The amount of uncertainty due to the geocoder's reference database, number of street segments, interpolation etc. is not captured. It should be noted that both geocoders try very hard to make sense of the input so that even very convoluted or obscure queries get matched to geographic features. However, thereby the accuracy indicator is not always reliable. For out-of-range house numbers, the geocoder might try to match a similar street where that house number exists. It is therefore advisable to carefully examine any error messages and additionally, to check that the given result address matches the one initially queried.

A first – already efficient – error reduction strategy would be to simply select the source with the better accuracy indicator. Of the individual addresses we geocoded, we can determine combinations of accuracy indicators and assign respective correction tasks as shown in Table 2.

In cases with matching accuracy indicators, a comparison of their respective geographic positions is made. The conflation criteria then are the stated granularity of the geocoders, the distance of points, outlier detection, variations and fluctuations in neighbouring house numbers etc. If the results lie within a small threshold from each other, an averaged position can be sufficient. In theses cases, differences may result in different

Google	Yahoo!	Description of accuracy level
0	warning or error	Unknown location.
1	country	Country
2	state	Region (state, province, prefecture, etc.)
3		Sub-region (county, municipality, etc.)
4	city	Town (city, village)
5	zip	Post code
	zip+2	Post code + 2 digits (US).
	zip+4	Post code + 4 digits (US).
6	street	Street
7		Intersection
8	address	Address
9		Premise (building name, property name, shopping center, etc.)

Tabelle 1: Geocoding accuracy indicators from APIs

Geocoder1	Geocoder2	Result
street	street	Granularity is too low. Use the preferred geocoder, an average, or mark for manual geocoding.
address	street	Use geocoder 2
street	address	Use geocoder 1
address	address	calculate combined result

Tabelle 2: Corrections based on combinations of geocoding accuracy indicators

geocoder	mean	σ	min	max
google	179,2	512,3	6,2	2349,1
yahoo	225,8	549,0	0,0	2288,6
avg	169,4	490,4	0,0	2285,6

Tabelle 3: Results as mean error and standard deviation

geocoder	mean	σ	min	max
Artillerieweg_avg	18,9	14,1	0,0	57,7
Artillerieweg_google	25,6	11,8	7,1	53,9
Artillerieweg_yahoo	28,0	20,2	6,0	77,9
Damm_avg	24,7	15,8	0,0	51,4
Damm_google	50,2	33,5	8,1	137,6
Damm_yahoo	21,5	13,7	0,0	49,8
Donnerschweer_avg	111,4	246,8	0,0	802,0
Donnerschweer_google	27,2	20,8	6,2	90,5
Donnerschweer_yahoo	220,6	517,2	2,1	1661,1
Staustrasse_avg	73,7	111,4	0,0	351,2
Staustrasse_google	153,6	362,6	6,2	1413,9
Staustrasse_yahoo	179,9	231,0	0,0	696,4
Prinzessinweg_avg	645,6	938,2	0,0	2285,6
Prinzessinweg_google	647,0	947,2	10,0	2349,1
Prinzessinweg_yahoo	657,4	931,4	10,0	2288,6

Tabelle 4: Results for selected streets

side-of-road offsets or slightly shifted streets. For differences under 20, we use an average over both results. For larger differences, it often remains unclear which geocoder provides the better results since both claim to be address-level accurate. However, due to some properties of the underlying datasets, we can use a small part of the immediate environment of an address by additionally examining the directly surrounding house numbers and subjecting them to spatial analysis. The spatial relation of addresses above and below the currently checked address can show clusters of coordinates at a single point compared to a street-wise distribution. Due to the way numbers can be assigned, this needs to take numbering schemes and spatial distances into account. Exploiting this information allows us to arrive at a more reliable answer.

Using our algorithm for the test area, we find an increase in accuracy in terms of mean error and standard deviation and are able to improve upon the individual sources as seen in Table 4. Note that these are the results as compared to our manually geocoded data, which contain a larger number of incorrect streets as those were the ones we were interested in. If we were to examine a more natural set of streets including more streets with rather accurate geocoder results, this bias would be dampened and the mean errors would drop considerably. While there are cases where we have to note a small decrease,

these are within streets with already higher accuracy and are in part within a few metres which is acceptable for the gain in other areas where we can reduce the error by tens or even hundred meters.

Regarding the results of the evaluation, types of errors can be detected where one geocoder's results indicate a lower accuracy by inconsistent coordinates or outliers and are better within the respective other one. A limit of this method is at non-mapped numbers where one geocoder has only street-level accuracy, but the other, due to interpolation, implies better data. Self-ascribed indicators are usually reliable in a broad sense, but we still notice errors probably due to wrong street assignments in the reference database. Cases where both geocoders agree on an incorrect result or where one geocoder offsets a street cannot be detected by the current method. Since all cases occur, there is no direct subsumption relation between the used geocoders and depending on the address to be geocoded, a decision on usage has to be made.

While the addresses are geocoded to point features, these points are derived from the underlying road network. Therefore, errors induced during geocoding could be rectified based on the observation that the points are not randomly distributed, but well assigned along these roads and usually distinguishable. [11] uses these features to move geocoded nodes further from the road centerline by an offset depending on the side of the road as is obviously also implemented within te Yahoo geocoder. For the limited area we examined, we can note that the current Google data has a more thorough coverage, especially in the downtown area, but that the side-of-road properties of the Yahoo geocoder can provide better accuracy if the underlying street data is correct and consistent. Inconsistencies, however, seem present in both road data sets and currently set limits to the achievable accuracy of these geocoders and derived correction methods.

5 Conclusion

We showed that free geocoding services already support a high level of granularity but that the accuracy at highest granularity levels still introduces some errors. We demonstrated that by combination of multiple sources, we can deliver geocoded locations from full addresses at a better accuracy than from individual sources alone.

By combining several online geocoders, strengths of the individual sources can be exploited while their limitations can be alleviated. Still, some issues persist which cannot be rectified by our approach and currently remain undecidable. We are continuously working on these in our ongoing work to assess and identify the inaccuracies to improve the ranking and selection methods within our correction methods.

Literatur

[1] D. Ahlers and S. Boll. Location-based Web search. In A. Scharl and K. Tochtermann, editors, *The Geospatial Web*. Springer, 2007.

[2] D. Ahlers and S. Boll. Retrieving Address-based Locations from the Web. In C. Jones and R. Purves, editors, *GIR'08*. ACM, 2008.

[3] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging Web Content. In *SIGIR'04*. ACM, 2004.

[4] R. Bakshi, C. A. Knoblock, and S. Thakkar. Exploiting Online Sources to Accurately Geocode Addresses. In *GIS'04*. ACM, 2004.

[5] M. R. Cayo and T. O. Talbot. Positional error in automated geocoding of residential addresses. *Int. Journal of Health Geographics*, 2(1):10, 2003.

[6] C. A. Davis, Jr. and F. T. Fonseca. Assessing the Certainty of Locations Produced by an Address Geocoding System. *Geoinformatica*, 11(1):103–129, 2007.

[7] C. Farvacque-Vitkovic, L. Godin, H. Leroux, F. Verdet, and R. Chavez. Street Addressing and the Management of Cities. Technical report, The World Bank, Washington, DC, USA, 2005.

[8] D. W. Goldberg, J. P. Wilson, and C. A. Knoblock. From Text to Geographic Coordinates: The Current State of Geocoding. *Journal of the Urban and Regional Information Systems Association*, 19(1):33–46, 2007.

[9] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger. Design and implementation of a geographic search engine. In A. Doan, F. Neven, R. McCann, and G. J. Bex, editors, *WebDB 2005*, pages 19–24, Baltimore, Maryland, USA, 2005.

[10] K. S. McCurley. Geospatial mapping and navigation of the web. In *WWW'01*. ACM, 2001.

[11] J. H. Ratcliffe. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science*, 15(5), 2001.

[12] J. Zhang and M. Goodchild. *Uncertainty in Geographical Information*. New York, 2002.