Geocoding Place Names from Historic Route Descriptions

Daniel Blank
Media Informatics Group
University of Bamberg
96045, Bamberg, Germany
daniel.blank@uni-bamberg.de

Andreas Henrich
Media Informatics Group
University of Bamberg
96045, Bamberg, Germany
andreas.henrich@uni-bamberg.de

ABSTRACT

Many gazetteers contain only a small amount of historic place name information and spelling variants of places. Even more focused historic gazetteers are far from being complete and often specialize on certain geographic regions or particular time periods. On the other hand, there are huge amounts of historic route descriptions, so called itineraries. They represent massive knowledge sources from which historic place names and spelling variants can be deduced. The analysis and geocoding of those route descriptions—often done by hand—is an important task in the humanities. To cope with these problems, we present preliminary ideas how to automatically deduce historic place names and thus travel routes from historic route descriptions.

CCS Concepts

•**Information systems** → *Geographic information systems;*

Keywords

itinerary resolution, geocoding, gazetteer enrichment

1. SCENARIO

We address the problem of *itinerary resolution* where ancient route descriptions have to be resolved. Itineraries are regarded as tables consisting of historic toponym-distance-pairs. Each pair $\langle h_i, d_i \rangle$ represents a route stop i ($1 \leq i \leq n$) at a certain location described by the corresponding (historic) toponym h_i. Distance information d_i indicates how far the current location i is away from the previous stop $i-1$. Distances can be explicit (e.g. medieval miles) or implicit (e.g. route stops indicating day trips of similar distance).

Itinerary resolution is a problem with four sub-tasks: (1) *optical character recognition* (OCR), (2) *itinerary parsing*, (3) *toponym resolution*, and (4) *route finding*. As a starting point of our research, we address step (3) with the task of grounding places mentioned in the itineraries. As preliminary steps, we assume an OCR and itinerary parsing, including the extraction of distance information whenever possible. In a fourth step, which we will target in future work as well as steps (1) and (2), we must identify the concrete paths which underly a particular route description.

The goal of our work is twofold: the enhancement of gazetteers and thus geographic information retrieval (GIR) applications in general, and an automatic suggestion tool which geocodes and visualizes ancient travel routes for researchers in the humanities.

2. RELATED WORK

There is plenty of work on *geoparsing* and *geocoding* in GIR applications (see [5]). However, the task of toponym resolution for (structured) historic route descriptions has to our knowledge not been addressed. [8] deals with the *disambiguation of road names* in text route descriptions using an exact-all-hop shortest path algorithm. Graph weights are geospatial distances whereas we use string distances. [8] allows subpaths as solutions, i.e., certain stops on a route may be skipped. In contrast, we only allow routes of full length. We also make use of vertex filtering when traversing our graph. [1] presents the task of *itinerary retrieval* which aims for the identification of travel itineraries found on the web. A grounding algorithm is applied. However, it does not make use of the itinerary structure itself and does not focus on historic itineraries with spelling variants. The focus lies on machine learning to classify tables as itineraries. Itinerary resolution is clearly different from itinerary retrieval. Instead of identifying itineraries, we want to derive historic place names from itinerary data. Ideas from [1] could complement our work as a preprocessing step for the automatic identification of itineraries in large text corpora. Addressing *spelling variations* is essential for IR (for references see [3]). Considering historic spelling variants is an important task, too. [3] analyzes the retrieval with historic spelling in a broader context independent from GIR. *String similarity measures* are important for name and address matching. Their use in GIR applications is less frequent. [6] compares *string similarity measures for toponym matching* and provides a good source for the selection of adequate string similarity measures in future work. Closely related to our work is gazetteer integration where it is common to use distance filters in addition to string matching (see e.g. [2]).

3. HISTORIC PLACE NAME DETECTION

Figure 1 shows the conceptual decision graph on which we operate with a virtual start and end node. We iterate over the potentially unknown toponyms in the historic route

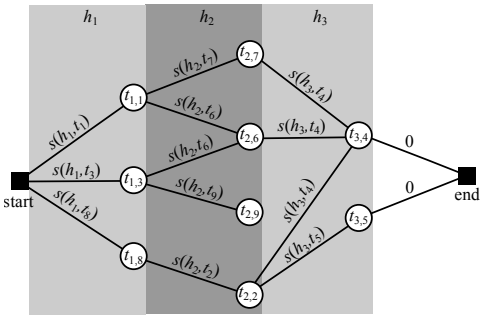

Figure 1: Decision graph.

Table 1: Result table.

input h_i	result	groundtruth	dist
Franckhfurt	Frankel	Frankfurt	66.2
Aschenburg	Scheringen	Aschaffenburg	57.3
Mildtenburg	Kaltenbrunn	Miltenberg	9.1
Bischoffhaim	*Tauberbischofsheim*	Tauberbischofsheim	0
Würtzburg	*Würzburg*	Würzburg	0
Detelbach	*Dettelbach*	Dettelbach	0
Haßfurt	Westheim bei Haßfurt	Haßfurt	6.5
Bamberg	Bischberg	Bamberg	5.1
Liechtenfelß	Steinfeld	Lichtenfels	19.9
Khulmbach	*Kulmbach*	Kulmbach	0

description (here: h_1 till h_3). For every stop h_i we calculate a string distance $s(h_i, t_j)$ to all toponyms t_j in a gazetteer. If $s(h_i, t_j) \leq \arg\min_{t_j} s(h_i, t_j) + \delta$, the node $t_{i,j}$ is considered further. δ is a predefined string distance threshold.

While creating the graph, we apply additional filters on the toponyms from the gazetteer in addition to the string distance. To find destinations t'_{i+1} from a source t_i, spatial range queries are performed searching for all toponyms more than $d_{i+1} \cdot r_{in}$ and less than $d_{i+1} \cdot r_{out}$ meters away from t_i. The parameters r_{in} and r_{out} are influenced by historic length units, detour factors which are characteristic for a certain time period and region, etc. In addition, we check if the azimuth change at a location is bigger than a threshold α (in degrees). Otherwise, the toponym is rejected; α represents the extend of directionality of an itinerary and usually the directionality is quite strong in historic route descriptions which often focus on getting from one location to another as directly as possible. Finding the best matching toponyms in our case boils down to finding the shortest path(s) in terms of the string distance from the start to the end node.

4. EVALUATION

We plan to evaluate our approach on various itineraries, starting with travel routes taken from an itinerary published in 1563 by Jörg Gail [4]. His work is said to be the first autonomously printed travel itinerary in the German language [4, *p.* 1]. We did some initial tests with route 1 of the itinerary. The route part we consider has $n = 10$ stops (see table 1). In a first step, we use the Jaro distance [7]. We apply a detour factor of 1.2 which is typical for road networks. By setting $r_{in} = 6000$ and $r_{out} = 8000$ we assume a medieval mile in Germany between 7200 and 9600 meters inspired by *https://de.wikipedia.org/wiki/Meile*. The azimuth angle is set to 90 degrees. $\delta = 0.24$ is roughly tuned by hand, all other parameters were initially set and remained unchanged. Table 1 shows the *result* given at rank 1. The toponyms in italics, $4/10 = 40\%$, are correctly identified. We split toponyms found in the gazetteer (German toponyms from *www.geonames.org* including alternate names) because e.g. Frankfurt, the starting point of our route, is contained as "Frankfurt am Main" in the gazetteer. Partial matches are not discounted because "Frankfurt am Main" should be considered equally important as "Frankfurt" which is also contained in the gazetteer as a small village far away from the route. As a consequence, a 5th match is prevented by "Westheim bei Haßfurt" which is prefered over "Haßfurt". It can also be seen that some toponyms lie close to the perfect match (*dist* denotes the geospatial distance between the *result* and the *groundtruth* location in km). This makes us believe that our geographic filter is adequate in principal. Furthermore, it is hard to find the correct start location. Using not all toponyms of the gazetteer but only populated places of a certain size seems promising, although this might also lead in a wrong direction since some medieval spots lost their importance. Note that only two out of the ten *input* locations are uniquely found in the gazetteer ("Haßfurt" and "Bamberg"). Incorporating this fact into our procedure is another aspect of future work. "Bischoffhaim" is correctly resolved to "Tauberbischofsheim" although there are three occurrences of "Bischofsheim" in the gazetteer. This makes us believe that the current approach combining various filters as well as geospatial and string distances is promising.

5. FUTURE WORK

We will examine in future work different gazetteers, string and phonetic distance measures, filter criteria, and in particular an algorithm for determining δ. Future analysis will focus on various route descriptions from the Gail itinerary and other itineraries with more exotic spelling variants.

6. REFERENCES

[1] M. D. Adelfio and H. Samet. Itinerary retrieval: Travelers, like traveling salesmen, prefer efficient routes. In *GIR*, pages 1:1–1:8, Dallas, TX, USA, 2014.

[2] M. L. Berman et al. Historical Gazetteer System Integration: CHGIS, Regnum Francorum, and GeoNames. *www.fas.harvard.edu/~chgis/gazetteer/AAG_GazIntegration_23feb2012.pdf*. Version: 23.02.12.

[3] A. Ernst-Gerlach and N. Fuhr. Retrieval in text collections with historic spelling using linguistic and spelling variants. In *JCDL*, Vancouver, BC, 2007.

[4] H. Krüger. *Das älteste deutsche Routenhandbuch. Jörg Gails Raißbüchlin.* 1974. Akad. Druck- und Verlagsanstalt, Graz.

[5] L. Moncla et al. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *GIS*, Dallas, TX, 2014.

[6] G. Recchia and M. Louwerse. A comparison of string similarity measures for toponym matching. In *Computational Models of Place*, Orlando, 2013.

[7] W. E. Winkler. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Section on Survey Research Methods*, pages 354–359, 1990.

[8] X. Zhang et al. *Disambiguating road names in text route descriptions using Exact-All-Hop Shortest Path algorithm*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 876–881. 2012.