Toponym disambiguation of landscape features using geomorphometric characteristics

C. Derungs[1], R. S. Purves[1], B. Waldvogel[2]

[1]University of Zürich - Irchel, Winterthurerstr. 190, 8057 Zürich, Switzerland
Email: curdin.derungs, ross.purves@geo.uzh.ch

[2]Swiss Federal Research Institute WSL, Zürcherstr. 111, 8903 Birmensdorf, Switzerland
Email: bettina.waldvogel@wsl.ch

1. Introduction

Landscape descriptions in natural language are a primary source of what Egenhofer and Mark (1995) call naïve geographical knowledge. Naïve geographical knowledge, however, differs for different people from different cultures and backgrounds (Mark and Turk 2003). For example a description of Uluru in Australia might be very different if given by Dutch tourist in comparison to one given an indigenous inhabitant.

Geoparsing, in particular considering toponym ambiguity is a key task in linking language to space through the assignment of geographic scopes to documents (Clough 2005). Leidner (2007) states that almost all research in geoparsing has focused on *populated places*. 'Population' furnishes toponyms with a priori knowledge that is used by state of the art disambiguation approaches (e.g. Purves et al. 2007) using the most populated place as the default toponym in disambiguation.

Landscape descriptions, however, typically contain references to unpopulated places, implying other approaches must be adopted to disambiguate.

Here we generate missing knowledge about toponyms using geomorphometric characteristics, in our case for a landscape feature known as a Hochmoor[1]. The toponym knowledge thus created is used for referent disambiguation (i.e. is London, England or London, Ontario relevant) - to our knowledge the first example of *geomorphometric disambiguation*. Our method shows considerable improvement in performance over a baseline disambiguation method. Disambiguation is the first important step towards opening up extensive sources of naïve geographical knowledge in the form of landscape descriptions in natural language which are likely to contain many ambiguous toponyms, which in turn will make such documents more accessible for a wide range of geographically rooted research.

2. Data Center Nature and Landscape

In our investigation we use documents describing Hochmoor in natural language. The documents are part of the Data Center Nature and Landscape (DNL). The DNL was established according to the specifications of the Swiss Nature and Cultural Heritage Protection to manage all Swiss data regarding protected areas of national importance. Information on the condition, composition and location of more than 500 *Hochmoor* in

[1] We use the German term *Hochmoor* which is a geographic object closely related to a high moor or a bog, to avoid semantic confusion through translation.

Switzerland has been collected in a corpus and recorded in separate datasheets (Bauer-Messmer et al. 2009). The datasheets are written in three national languages, French, Italian or German and we investigate German datasheets here (n=370). A simple gazetteer lookup performed on the documents using SwissNames[2] recognizes 600, mostly ambiguous, toponyms that can be referenced to more than 2500 locations in Switzerland.

3. Geomorphometric knowledge for toponym disambiguation

We assume that locations of toponyms used to georeference Hochmoor have a Hochmoor-like topography. Therefore a geomorphometric measurement for Hochmoor is deduced from topography. This measurement is further used as the missing knowledge in disambiguation (Figure 1).

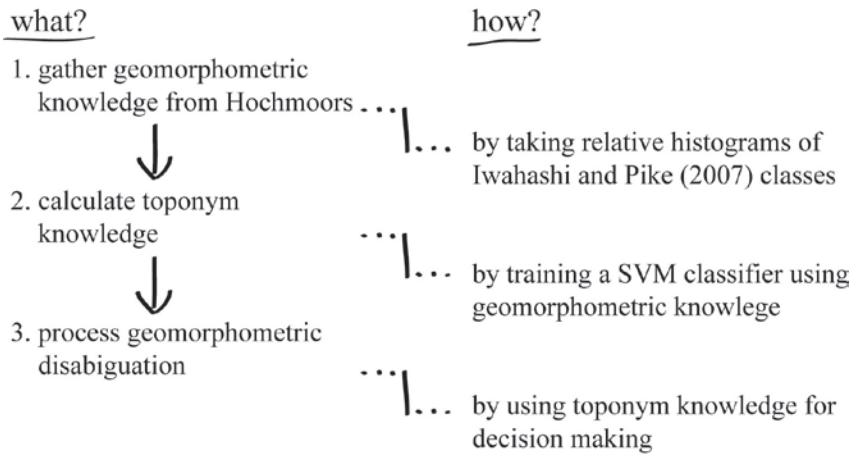

Figure 1. Workflow to process disambiguation with geomorphometric knowledge.

In a first step real Hochmoor locations (n=100) are used to infer geomorphometric knowledge. Thus, relative histograms for the 16 geomorphological classes introduced by Iwahashi and Pike (2007) are calculated for two windows of 0.25km and 5km centered on Hochmoor locations. The same is done for 1000 random locations within Switzerland (Figure 2).

What we term geomorphometric knowledge has become a vector with 32 dimensions, one vector for each Hochmoor and random location (16 classes for the 0.25km and 5km window respectively). The geomorphometric knowledge can be summarised as follows: In close proximity to Hochmoor centers (0.25km) topography is characterised by fine textures and gentle slopes (classes 9, 11, 13, 15). Steep slopes and coarse textures become more frequent if we widen the scale to the neighborhood of a Hochmoor (5km; classes 6, 8). This conforms to our notion of Hochmoor being plains in a mountainous environment, a secondary effect of the process of Hochmoor evolution.

The generated geomophometric knowledge, in terms of location-vectors with 32 dimensions, is used to train a probabilistic SVM classifier (Burges 1998) to distinguish Hochmoor from random locations (probability is equal to the distance between vector and

[2] http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html

hyperplane). The classifier can be used to quantify geomorphometric Hochmoor probability for each designated set of coordinates. In our case we are interested in geomorphometric Hochmoor probabilities for all 2500 referent locations from the datasheets. At this stage geomorphometric Hochmoor probability has become what we term toponym knowledge.

In a last step we disambiguate toponyms using the generated toponym knowledge. In a most basic disambiguation scenario each of the 600 toponyms are disambiguated with the referent location of maximum geomorphometric Hochmoor probability.

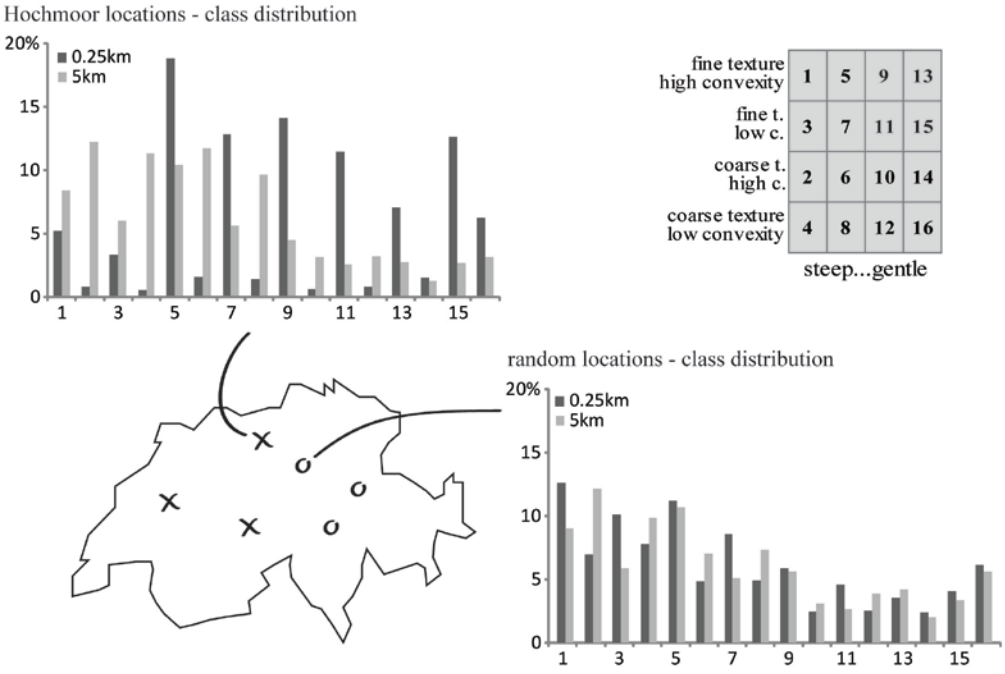

Figure 2. The 16 Iwahashi & Pike classes (upper right) and two typical relative histograms for a Hochmoor and a random location.

4. Geomorphometric disambiguation results

Here we focus on referent disambiguation of datasheets containing a single ambiguous toponym. All toponyms were manually semantic disambiguated in a previous step (e.g. removing instances of Bath where it is a place to wash and not a town).

There are 50 such single toponyms with 330 referent locations covering 20% of all datasheets. Single toponyms are the most complex case of toponym ambiguity, since knowledge gained from other, unambiguous toponyms, in a datasheet cannot be used to aid the process.

As is shown in the previous section only the referent location with the highest geomorphometric Hochmoor probability is resolved. In Figure 3 the Hochmoor probabilities for all 330 referent locations are plotted against the distance to the corresponding Hochmoor.

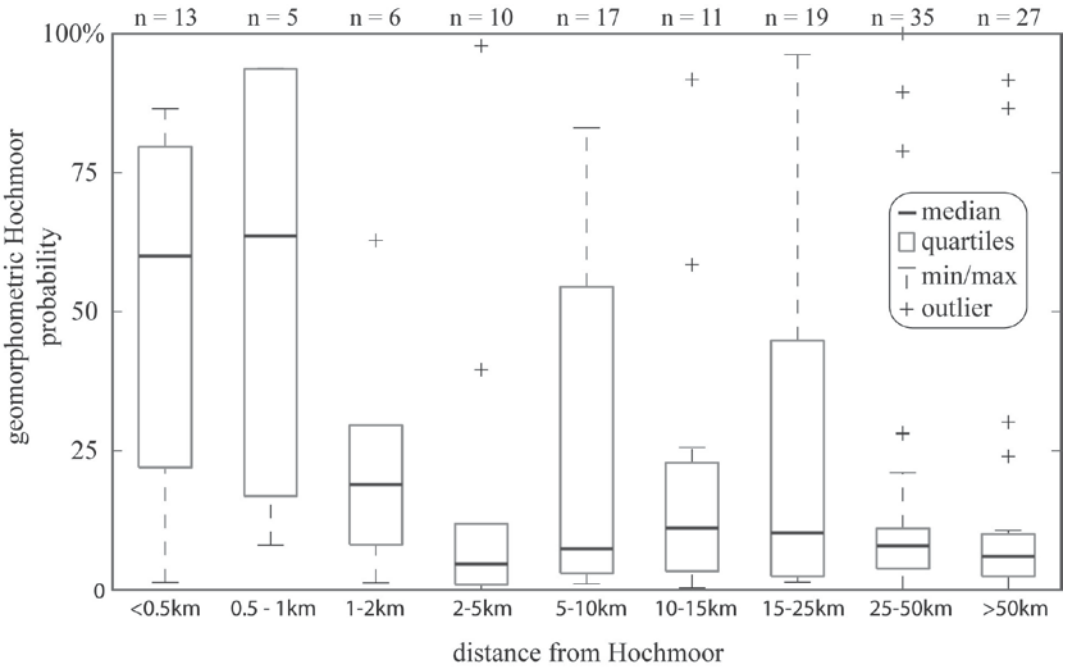

Figure 3. Boxplot of geomorphometric Hochmoor probability and distance to Hochmoor for 330 referent locations.

Figure 3 shows that geomorphometric Hochmoor probability is high for close referent locations and vice versa. In a nutshell, geomorphometric disambiguation allows us to resolve some 58% of the 330 referent locations. The baseline for disambiguation, i.e. the mean probability of successfully disambiguating toponyms by making a random decision, given no other information, is only 23%.

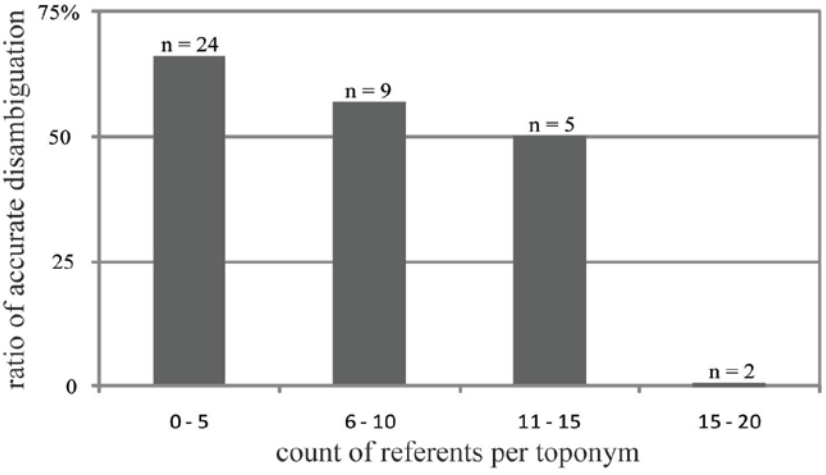

Figure 4. Disambiguation accuracy compared with count of reference locations of toponyms.

In Figure 4 the relationship between disambiguation accuracy and count of potential referents per toponym is visualized. The accuracy drops as the count of referent locations of toponyms increases.

5. Conclusions

Using the knowledge generated from geomorphometric characteristics of Hochmoor makes disambiguation more than twice as precise as the baseline (58% vs. 23%). Topography supplies substitute knowledge for cases where no a priori knowledge is available.

We used a rather basic approach to gather Hochmoor probability from topography. However, the same approach could be applied to all kinds of geographic objects (e.g hills, mountains or lakes).

Disambiguation with many referent locations is still inaccurate (Figure 4). Sometimes topographic Hochmoor probability is considerably higher for locations being far from the actual Hochmoor (Figure 3, outliers >25km). This may be due to false positive classifications, however, our inventory describes Hochmoor as classified at the present time, whilst geomorphometric characteristics describe locations with the affordance of being a Hochmoor, which may have been drained or otherwise altered in the last 200 years, which applies for some 85% of all original Hochmoor (Klaus 2007).

Many referent locations that are close to Hochmoor have rather small geomorphometric Hochmoor probabilities (Figure 3, minimas >1km). The assumption of spatial referents to Hochmoor always having a Hochmoor like topography is therefore clearly not always true.

In further work we will concentrate on resolving semantic ambiguity in landscape descriptions. We will face a very similar problem. Again there is no a priori knowledge that could serve for disambiguation. The general aim is to explicitly link landscape descriptions with space. This is the first important step to make naïve geographical knowledge in landscape descriptions useable.

6. Acknowledgements

The research reported in this paper is funded by the SNF Project 200021-100054.

7. References

Bauer-Messmer B, Grütter R, Hägli M and five others, 2009, Service Oriented Architecture, Metadata Standards and Semantic Technologies in an Environmental Information System. In: Wohlgemuth V et al. (eds), *Proceedings of the 23rd EnviroInfo*, Berlin, Germany, 101-112.

Burges CJC, 1998, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2(2):121-167.

Clough P, 2005, Extracting metadata for spatially-aware information retrieval on the internet. In: Jones C and Purves RS (eds), *Proceedings of the 2005 workshop on GIR*, Bremen, Germany, 25–30.

Egenhofer M and Mark D, 1995, Naive Geography. In: Frank AU and Kuhn W (eds), *Spatial Information Theory A Theoretical Basis for GIS*, Berlin, Germany, 988:1-15.

Iwahashi J and Pike RJ, 2007, Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*, 86(3): 409–440.

Klaus G, 2007, Zustand und Entwicklung der Moore in der Schweiz, BAFU, Bern, Schweiz.

Leidner JL, 2008, *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. PhD thesis, School of Informatics, University of Edinburgh

Mark D and Turk A, 2003, Landscape categories in Yindjibarndi: Ontology, environment, and language. *Spatial Information Theory*, 2825:28–45.

Purves RS, Clough P, Jones CB and eight others, "The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet," *International Journal of Geographical Information Science* 21, no. 7 (1, 2007): 717-745.