Towards mapping of alpine route descriptions

Piotrowski, Michael ; Läubli, Samuel ; Volk, Martin

DOI: https://doi.org/10.1145/1722080.1722083

Towards Mapping of Alpine Route Descriptions

Michael Piotrowski
mxp@cl.uzh.ch

Samuel Läubli
samuel@access.uzh.ch

Martin Volk
volk@cl.uzh.ch

Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14, 8050 Zurich, Switzerland

ABSTRACT

We describe a corpus of historic mountaineering accounts and ongoing work on geocoding toponyms and route descriptions in these accounts. Mountaineering accounts contain a wealth of geographic information but its extraction for purposes of geographic information retrieval poses specific challenges, in particular the distinction between toponyms pertinent to route descriptions and those mentioned in descriptions of panoramas. We describe some preliminary considerations for natural language cues to distinguish between these two types of occurrences.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; I.2.7 [Natural Language Processing]: Text analysis

General Terms: Experimentation.

Keywords: Geographic information retrieval, toponym resolution, route extraction, cultural heritage data, mountaineering accounts.

1. INTRODUCTION

Since 1864, the Swiss Alpine Club (SAC) has published yearbooks (*Jahrbuch des Schweizer Alpenclub*); in 1925 they were replaced by a journal (*Die Alpen/Les Alpes/Le Alpi*). We are currently working on the digitization of all yearbooks and journal issues from 1864 until today. At the time of this writing, the volumes from 1864 until 1964 have been scanned and OCR-processed, resulting in a corpus of about 20 million word forms (6300 articles). Each yearbook or journal volume consists of about 300 to 600 pages and contains articles on various aspects of mountaineering and related topics, such as expedition reports, folklore and ethnography, flora, fauna, and geology; most articles are related to the Swiss Alps. At the beginning, the yearbooks contained mostly articles in German, and some in French. Since 1957, "Die Alpen" are published in parallel German and French versions (with occasional articles in Italian). This corpus is thus a valuable resource for cultural, geographic, historical, and linguistic studies.

A large portion of the corpus consists of *alpine literature*, a literary genre of its own, which includes both fiction (poetry and prose) and non-fiction work on mountaineering and general alpine topics. Most of the non-fiction works are *mountaineering accounts*, i.e., reports of ascents or expeditions; these texts are the ones that are most interesting for researchers, as they reflect the reality of the time and its contemporary perception. While the accounts are factual reports, their style is nevertheless frequently more literary and narrative than expository. For example, interspersed between the descriptions of the legs of the itinerary are often passionate digressions on the majesty of the mountains, the beauty of nature, and the value of friendship. The literary style is also evident in the use of analepses, recounting previous expeditions or events.

In the rest of this paper, we will describe some of the special properties of the texts with respect to geoparsing and geocoding, and we will outline our ongoing work towards geographic information retrieval for mountaineering accounts.

GIR'10, 18–19th Feb. 2010, Zurich, Switzerland

2. EXTRACTING ROUTES FROM TEXT

Mountaineering accounts primarily focus on routes, i.e., they are not about a peak (as a location), but about *how* that peak was reached. Thus, routes and their descriptions are of particular interest for research. We therefore aim to automatically extract and map route descriptions from mountaineering reports. We are currently focusing on German-language reports about mountaineering in Switzerland between 1900 and 1950.

2.1 Toponym Detection and Resolution

As a first step towards route extraction, it is necessary to identify and geocode the toponyms mentioned in the texts (mountains, glaciers, lakes, creeks, etc.). This is commonly being done for other types of texts, in particular news articles and Web pages (e.g., [5]). At first sight, mountaineering accounts pose the same problems for geoparsing and geocoding as other text types, especially toponym resolution, i.e., disambiguation of ambiguous toponyms.

Brunner and Purves [2] have noted that ambiguous toponyms seem to be spatially autocorrelated, i.e., that ambiguous toponyms are on average spatially closer than unambiguous names. They also noted that the scope of news articles can thus often be larger than the distance between two referents of an ambiguous toponym. We have also observed this in the mountaineering accounts: Many of the toponyms in the corpus are ambiguous, even within the small areas covered by a single text. For example, the SwissNames25 gazetteer contains 23 mountain peaks with the name *Schwarzhorn* in an area of about 180×100 km^2; in some cases, the two referents are less than 3 km apart.

Due to the peculiarities of mountaineering accounts, there are some special challenges; in particular, most of the accounts cover only a small area and thus mention many "small" features (e.g., mountain huts, foot passes, ridges) not commonly contained in gazetteers. Despite the SwissNames25 gazetteer's detailed coverage

Um 5½ Uhr verließen wir Champex und stiegen über schmale Fußwege, die nur ein Führer kennen kann, auf dem geradesten Wege ins Val Ferret hinab, das wir etwas unterhalb Ville d'Issert erreichten. Bei Praz de Fort bogen wir rechts ab, dem Fuße des Glacier de Saleinaz entgegen. [...] Über die Wälder und Gletscher emporragend, wird vor uns der steile Clocher de Portalet sichtbar; [...] In gewaltigen Felswänden erheben sich zu unserer Linken die beiden Clochers de Planereuse, [...]

Figure 1: Text sample from the corpus. Double underlining: Toponyms relevant for the route description; single underlining: Toponyms only mentioned in description of panorama.

of Switzerland, first experiments on the recognition of mountain names using a naïve approach have yielded surprisingly low recall (< 40%). There are various reasons for this. First of all, Switzerland is a multilingual country and many places have separate names in the different languages. The official maps, on which SwissNames25 is based, only give the name in the local official language. In the texts, however, toponyms are generally in the language of the text. We also have to cope with historical spellings and OCR errors. A special challenge for geoparsing is that many texts try to avoid repetitions of toponyms for stylistic reasons and employ a wide variety of coreference devices, including pronouns, shortened names (e.g., *Schreckhorn* vs. *Grosses Schreckhorn*), hypernyms (e.g., *Hütte* vs. *Schwarzegghütte*), or non-standard names (e.g., *Andersongrat* vs. *Nordwestgrat*). Also frequently used are partial coordinations such as *Glarner-, Urner- und Berner Alpen* (for *Glarner Alpen, Urner Alpen und Berner Alpen*).

2.2 Route Extraction

Once the toponyms have been geocoded, route extraction is basically the task of connecting the points identified in the first step in the correct order. Most of the work in this area has been done for directions, i.e., instructions for wayfinding. For example, Arikawa and Noaki [1] have processed textual walking directions and visualized the resulting route using a sidewalk network database.

Unlike directions, however, which usually mention only toponyms relevant to the route, the mountaineering accounts contain numerous toponyms in descriptions of panoramas, which are not directly pertinent to the route. These toponyms often outnumber the relevant ones, and the referents can be far away from the route (see fig. 1). The main challenge is thus to distinguish between points relevant for the route and irrelevant points. Due to the high number of irrelevant points, standard toponym resolution methods based on minimum bounding rectangles and centroids (such as those summarized by Leidner [3]), are likely to miscalculate the focus of these texts.

However, there are a number of textual and linguistic indicators that may help with this task. For example, massings of toponyms in a paragraph often indicate a description of a panorama—especially if the distances between the mentioned points are relatively large—, whereas an isolated occurrence of a toponym increases the likelihood that it is relevant for the route description.

On the sentence level, we have noted that sentences containing relevant toponyms often have the authors as the subject, the toponym is the object or occurs in a prepositional phrase, and the verb indicates movement or position; for example: *Um 7 Uhr 30 setzten wir uns gegen den Anderson-Grat in Bewegung* 'At 7:30 we started to move towards Anderson Ridge.' In panorama descriptions, on the other hand, the toponym tends to be the subject of the sentence, e.g., *Weit im Süden streckte der Mont Blanc sein riesiges Haupt in den blauen Äther hinein* 'Far to the south, Mont Blanc raised its giant head into the blue ether.'

Verbs of vision require further analysis, as they can occur in both cases. For example, in the sentence *[...] nun sahen wir unser heutiges Ziel, die Capanna Caecilia* 'now we saw today's destination, the Capanna Caecilia,' the toponym (a mountain hut) is a waypoint, whereas *Man sah bis hinaus zu Schwarzwald und Vogesen* 'One could see as far as the Black Forest and the Vosges' is a description of a view and not related to the route.

Lesbegueries et al. [4] have defined a number of criteria for associating text units to to one of three "spatial patterns" (point of view, itinerary, and area comparison), which may also be useful.

Making use of the linguistic information requires at least shallow parsing to identify subjects, objects, verbs and prepositional phrases and a lexical-semantic net such as WordNet or GermaNet to be able to work with abstractions such as verbs of movement, of vision, etc. Since route descriptions also have a temporal dimension, recognizing temporal expressions or *timexes* is also required; this can either be integrated with the syntactic analysis or implemented as a separate component such as a timex tagger, which detects, annotates, and normalizes temporal expressions [6]. We are currently evaluating and selecting NLP components for use in the route extraction task.

3. CONCLUSIONS

Mountaineering accounts are a special text genre, which contains numerous geographic references. Both the style of the texts and the properties of the toponyms pose specific challenges to geocoding, in particular the distinction between toponyms pertinent to a route description and those mentioned in panorama descriptions. Preliminary experiments suggest that spatial information alone is not sufficient for this task. We thus explore the use of natural language cues for distinguishing between these two types of references.

4. REFERENCES

[1] M. Arikawa and K. Noaki. Geocoding Japanese walking directions using sidewalk network databases. In G. Gartner, W. Cartwright, and M. P. Peterson, editors, *Location Based Services and TeleCartography*, Lecture Notes in Geoinformation and Cartography, pages 217–229. Springer, Berlin/Heidelberg, 2007. doi: 10.1007/978-3-540-36728-4_17.

[2] T. J. Brunner and R. S. Purves. Spatial autocorrelation and toponym ambiguity. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 25–26, New York, NY, USA, 2008. ACM. doi: 10.1145/1460007.1460013.

[3] J. L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal, Boca Raton, FL, USA, 2008.

[4] J. Lesbegueries, C. Sallaberry, and M. Gaio. Associating spatial patterns to text-units for summarizing geographic information. In R. Purves and C. Jones, editors, *Proceedings of the 3rd Workshop on Geographic Information Retrieval (GIR 2006)*, 2006.

[5] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: a new view on news. In *GIS '08: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pages 1–10, New York, NY, USA, 2008. ACM. doi: 10.1145/1463434.1463458.

[6] M. Verhagen and J. L. Moszkowicz. Temporal annotation and representation. *Language and Linguistics Compass*, 3(2):517–536, 2009. doi: 10.1111/j.1749-818X.2008.00116.x.