

Classifying Named Entities in an Alpine Heritage Corpus

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter

In the project "Text+Berg" we digitize and archive the heritage of alpine literature from various European countries. In a first step our group digitizes all yearbooks of the Swiss Alpine Club from 1864 until today. The books comprise articles in German, French and Italian, a total of around 100.000 pages. This paper describes the corpus and the project phases towards its digitalization. We then focus on the classification of named entities, in particular person names and geographic entities.

1 Introduction

In the project Text+Berg¹ we digitize the heritage of alpine literature from various European countries. This is a joint project by the German Department and the Institute of Computational Linguistics at the University of Zurich. Currently our group digitizes all yearbooks of the Swiss Alpine Club (SAC) from 1864 until today. Each yearbook consists of 300 to 600 pages and contains reports on mountain expeditions, culture of mountain peoples, as well as the flora, fauna and geology of the mountains. The corpus is thus a valuable knowledge base to study the changes in all these areas. It provides a time line of information on hot topics like climate change, sustainable use of alpine resources, and technological developments ranging from transportation to media, as well as communication and climbing equipment. But the corpus is also a resource to catch the spirit of Switzerland in cultural terms: What does language use in alpine texts show about the cultural identity of the country and its change over time?²

This paper describes the corpus and the project phases towards its digitalization. We then describe the recognition and classification of named entities, in particular person names and geographic entities.

2 The Text+Berg Corpus

The Swiss Alpine Club was founded in 1863 as a reaction to the foundation of the British Alpine Club a year before. Thus our corpus has a clear topical focus: conquering and understanding the mountains. But at the same time it covers a wide variety of text genres as for example expedition reports, (popular) scientific papers, book reviews, etc. The articles focus mostly on the Alps, but over the 144 years the books have probably covered any mountain region on the globe.

Some examples from the 1911 yearbook may illustrate the diversity. There are the typical reports on mountain expeditions: "*Klettereien in der Gruppe der Engelhörner*" (English: Climbing in the Engelhörner group) or "*Aus den Hochregionen des Kaukasus*" (English: From the high regions of the

¹ See <http://www.textberg.ch>.

² See [3] for the theoretical and methodological background of research in this domain.



Caucasus). But there are also articles on scientific topics such as topography, geology or glacierology. The 1911 book contains scientific articles on the development of caves ("*Über die Entstehung der Beaten- und Balmfluhhöhlen*") and on the periodic variations of the Swiss glaciers ("*Les variations périodiques des glaciers des Alpes suisses*").

The corpus is multilingual. Initially the yearbooks contained mostly German articles and few in French. Since 1957 the books appeared in parallel German and French versions (with some Italian articles) which allows for interesting cross-language comparisons. The parallel versions followed the same article sequence and page structure, but not all articles were translated. For example, the parallel 1958 German and French yearbooks contain 27 translated articles (German-French), plus 38 of the same articles in both yearbooks (DE:18, FR:17, IT:3).

3 Project Phases

We have already collected all books in two copies (as a result of a call for book donations by the Swiss Alpine Club). One copy was cut open so that the book can be scanned with automatic paper feed. The other copy remains as reference book.

3.1 Scanning and OCR

We use state-of-the-art OCR software to convert the images to text. This software comes with two lexicons for German

(and one for French and Italian) which match the spelling after 1901 and the new orthography after the spelling reform of the late 1990s. For the German spelling of the 19th century (e.g. old *Nachtheil* and *passiren* instead of modern *Nachteil* and *passieren*) we will have to rely on the quality of the character recognition without lexicon look-up. Fortunately the yearbooks were set in Antiqua font from the start in 1864. So we do not have to deal with old German Gothic font (Fraktur).

A group of student volunteers helps in the correction of the OCRred text. The idea is to get the text structure right and to eliminate the most obvious OCR errors. We are also experimenting with methods for automatic OCR-error correction, e.g. statistical approaches as described in [7].

3.2 Mark-up of the Text Structure

We first introduce a mark-up of the text structure. Specially developed programs annotate the text with TEI-conformant XML tags for the beginning and end of each article, the title and the author, for page numbers, footnotes and caption texts. Much of that information can be checked against the table of contents and table of figures in the front matter of the yearbooks. This increases the annotation accuracy.

3.3 Language Identification

Proper language identification is important for most of the subsequent steps of automatic text analysis, e.g. part-of-speech recognition and lemmatization. In addition, it should be possible to limit the search in the text corpus to a specific language. Of course it is also interesting to evaluate the change in language use in our corpus over the years (German vs. French vs. Italian vs. other languages).

Therefore we use a language identification program³ to determine the language for each sentence. Such a fine granularity helps us to detect quotes and direct speech in languages different from the text language (as for example English sentences in German text).

With respect to other languages we are interested in learning more about the usage of the Swiss minority language Rhaeto-Romanic, a Romance language that is still spoken today by a few 10'000 people in the canton Graubünden. When is this language used in our corpus? And to what extent does the corpus contain texts, passages and quotes of direct speech in Swiss German? Finally, what is the role of English in these books given that British mountaineers and tourists were amongst the first and most active in the 19th century? For example, we were surprised to find that the 1903 yearbook contains a German article with the English statement that Switzerland has turned into "the playground of Europe".

3.4 Archiving, Access and Distribution

In the final phase the annotated corpus will be stored in a database which can be searched via the internet. Because of our detailed annotations the search options will be more powerful and lead to more precise search results than usual search engines. For example, it will be possible to find the answer to the question "List the names of all glaciers in Austria that were mentioned before 1900." We also annotate the captions of all photos and images so that they can be included in the search indexes.

³ We use Michael Piotrowski's *Lingua-Ident*.

In addition to the query module we will provide easy access to the texts and images through a variety of intuitive and appealing graphical user interfaces. We plan to have clickable geographic maps that lead to articles dealing with certain regions or places.

4 Named Entities in our Text+Berg Corpus

Named entity recognition is an important aspect for information extraction. But it has also been recognized as an important aspect for the access of heritage data. [2] argue for named entity recognition in 19th century Swedish literature, distinguishing between 8 name types and 57 subtypes.

We have investigated methods for named entity recognition in newspaper texts [9], and in this paper we report on how these methods work on our Text+Berg corpus.

4.1 Person Names

In a previous project [9] we had built three rule-based modules for the recognition of person names, geographical names and company names respectively in a corpus of a weekly business-oriented computer newspaper. The module for person name recognition relied on a long list of person first names and the fact that person names in newspapers are usually introduced by a first name followed by a last name. Thereafter the last name can be used alone within the same newspaper article.

If the name is not mentioned for a longer period of text, then it needs to be reintroduced. We had modelled this observation with an algorithm which we called learn-apply-forget. When the person name is introduced after a first name trigger, then the last name is saved with a certain priming level (this is the learning step). This priming level is decreased for each subsequent sentence that does not contain the name. If the name does occur, then this increases the priming number. In our current research we investigate whether the same algorithm also works for our Text+Berg corpus.

One striking difference between our computer newspaper corpus and our current corpus is that many person names in the alpine yearbooks are not introduced by first name plus last name but rather by a title or a function term followed by a last name. Address forms like English *Mr.*, *Mrs.*, German *Herr*, academic titles (*Prof.*, *Dr.*) but also *Ingenieur* (engineer), *Major*, are particularly common in the early 1900s.

4.2 Geographical Names

In our work on named entity recognition in newspaper texts [9] we had only distinguished two types of geographical names: city names and country names. This was sufficient for texts that dealt mostly with facts like a company is located in a certain country or has started business in a certain city. But our Text+Berg corpus deals with much more fine-grained location information: mountains and valleys, glaciers and climbing routes, cabins and hotels, rivers and lakes. In fact the description of movements (e.g. in mountains) requires all kinds of intricate references to positions and directions in three dimensions.

Thus it is no surprise that geographers and computational linguistics alike are working on the problems of structuring

the semantics of spatial expressions. There are numerous initiatives to build geographic ontologies (e.g. [5]), and there are special workshops that deal with the analysis of geographic references in natural language text (for example the HLT-NAACL 2003 Workshop on Analysis of Geographic References).

According to the organizers of this workshop the analysis of geographic references in text involves four distinct stages:

1. geographic entity reference detection (hypothesizing that the strings *Matterhorn*, *Reuss*, *Zurich* are referring to geographical entities, i.e. a mountain, a river and a city respectively; this step includes the grouping of multiword names like *Mont Blanc*, *Col de Peuterey*, *Kleine Windgällen*, *Crans Montana*, *St. Moritz*)
2. contextual information gathering (classification and possible locations)
3. disambiguation (*Freiburg im Breisgau, Germany* vs. *Freiburg im Üechtland, Switzerland*; *Simon Bolivar* as a person name vs. as a mountain name; *Essen*, *Halle*, *Hof* as city names vs. as a regular nouns)
4. grounding (assignment of geographic coordinates; *Zurich* is on 47°22'N 8°33'E)

There have been a number of approaches on the identification of geographical references (e.g. [6], [4] and [1]). But they have focused mostly on newspaper texts. Our texts are much denser in terms of geographical references since mountain climbing is the central topic. For example, in the following paragraph we can identify the names of mountains (*Bocktschिंगel*, *Kleiner Ruchen*, *Hintere Kalkschyen*), of a glacier (*Hüfigletscher*), a cabin (*Hüfihütte*) and a snow formation on a mountain (*Bocktschingelfirn*).

*In kurzer Zeit erreichten wir sodann auf öfter beschriebenen Wege über den **Bocktschिंगel** und die **Hüfigletscher** das südliche Ufer und die **Hüfihütte**. Ganz wider Erwarten brach am 16. August ein glanzvoller Tag an. Um 4 Uhr verließen wir die Hütte und gewannen auf dem Weg, auf dem wir hergekommen waren, den **Bocktschingelfirn** um 6 Uhr, dann weiter über den **Kleinen Ruchen** den Nordgipfel der **Hintern Kalkschyen** um 8 Uhr 30 Min. Ohne Aufenthalt stiegen wir über den Südgrat ab zur berühmigten Scharte, deren Überwindung wir nach kritischer Musterung sogleich in Angriff nahmen. (SAC-Jahrbuch 1910, p.298, bold face added)*

In addition to the names there are other descriptive elements that provide for the textual coherence of the spatial description, many of those provide directions (*südliche Ufer*, *Nordgipfel*, *Südgrat*).

In our previous project we had identified geographical names based on large gazetteers for city and country names. In addition to the listed base forms our program was able to recognize genitive forms (*Frankreichs*, *Münchens*) as well as adjectival forms (*Münchner*, *Bad Homburger*). In recognizing mountain names we also need to take care of occasional plural forms (*Fergenhorn* - *die drei Fergenhörner*).

Since it is inefficient to compute all inflected forms beforehand, we use decomposing and lemmatization wherever possible. For example, the genitive form *Gornergrates* is split into *Gorner+grates* and then reduced to the base form *Gornergrat*. The corpus itself serves as dictionary source for verification of the computed lemmas. By splitting we also collect compounding elements like *Gorner* which occurs

frequently in *Gornergletscher* but rarely alone. The parallel French version *glacier de Gorner* helps to identify such compound elements.

In addition we will go beyond the pure list matching by aiming for context-based disambiguation between different name classes and between names and regular nouns.

- *Polen*, *Schweden* as country name vs. as inhabitant from that country
- *Berlin*, *Bern*, *Paris* as city name vs. as denoting a government
- *Jungfrau*, *Mönch* as mountain name vs. regular noun

4.2.1 Coreference Resolution

As a step towards extracting spatial descriptions we will identify coreference chains throughout each text. Coreference means that two words refer to the same object. Detecting coreference involves the mapping of different words to the same object, but it also requires the disambiguation of the "same" name referring to different objects. Here are some examples of coreference variants:

- name abbreviations (*Bad Homburg* - *Homburg*, *Eiger* - *Vordereiger*)
- definite noun phrases (*Sinambum* - *der heilige Berg*, *Tambohorn* - *die elegante Pyramide*)
- partial references (*Eiger* - *die Nordwand*)
- name variants (historical, dialectal, across languages: *Mons Egere* - *Eiger*, *Weisshorn* - *Wysshorn*)

We intend to solve the coreference problem with a combination of rule-based and statistical methods.

We ground the various place and direction terms used in mountain climbing on the data in geographical information systems (GIS). In other words, the textual descriptions of geographical features will be unambiguously mapped to the coordinates of the underlying real-world phenomena, as supplied by the GIS. For example, if the text talks about "the glacier on the north face of a mountain X", then this relation can be grounded and saved.

4.2.2 Mark-up of Semantic Relations

In this project we will go beyond the recognition of geographical entities towards the recognition of spatial descriptions. This means that we will automatically identify in which spatial relation two physical objects are. For example, we want to determine that "a cabin is located on a mountain" or that "a town is between a mountain and a river".

In order to reach this goal we need to analyze the texts in more detail. In other words we need to apply an automatic parser for syntax analysis. Our research group has recently developed robust and fast dependency parsers for English and German. Subsequent work has shown that the parser can be efficiently ported to other languages [8]. The parsers have a rule-based backbone, but they resolve ambiguities with statistical means. The statistical disambiguation, which is extracted from a treebank, allows the parser to prune substantially during parsing and to return likely analyses that are licensed by the grammar, ranked by their probability. Our parser has been shown to achieve state-of-the-art speed and accuracy.

5 Conclusion

We are working on the digitalization and annotation of alpine texts. In a first step we compile a corpus of 145 yearbooks from the Swiss Alpine Club. The Club has agreed to host the complete archive on its web server. Users will be able to inspect the search results per article as PDF documents so that they can enjoy the original look and feel (including all the pictures). In addition we will distribute our corpus as an XML-tagged textual resource for researchers.

If we can attract sufficient funding, we hope to have a preliminary version of the yearbooks on the web for full-text search and article-wise inspection in early 2010. We would like to have a preliminary version of the annotated corpus (as XML files) available for distribution by the summer of 2010.

6 Acknowledgments

We would like to thank the many student helpers who have contributed their time to this project.

References

- [1] Amitai Axelrod. On building a high performance gazetteer database. In *Proceedings of Workshop on the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Budapest, 2003.
- [2] Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In *Proceedings of The ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, Prague, 2007.
- [3] Noah Bubenhofer. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Number 4 in Sprache und Wissen. de Gruyter, Berlin, New York, 2009.
- [4] Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of Workshop on the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Budapest, 2003.
- [5] Guillermo Nudelman Hess, Cirano Iochpe, Alfio Ferrara, and Silvana Castano. Towards effective geographic ontology matching. In *Proceedings of the Second International Conference on GeoSpatial Semantics (Mexico City)*, Lecture Notes in Computer Science, pages 51–65. Springer, November 2007.
- [6] Erik Rauch, Michael Bukatin, and Kenneth Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of Workshop on the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Budapest, 2003.
- [7] Martin Reynaert. Non-interactive OCR post-correction for giga-scale digitization projects. In A. Gelbukh, editor, *Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference, CICLing 2008*, Lecture Notes in Computer Science, pages 617–630, Berlin, 2008. Springer.
- [8] Rico Sennrich, Gerold Schneider, and Martin Volk. A new hybrid dependency parser for German. In *Proceedings of GSCL-Conference*, Potsdam, 2009.
- [9] Martin Volk and Simon Clematide. Learn-filter-apply-forget. Mixed approaches to named entity recognition. In Ana M. Moreno and Reind P. van de Riet, editors, *Applications of Natural Language for Information Systems. Proc. of 6th International Workshop NLDB'01*, volume P-3 of *Lecture Notes in Informatics (LNI) - Proceedings*, pages 153–163, Madrid, 2001.

Contact

Prof. Dr. Martin Volk
 Universität Zürich, Institut für Computerlinguistik
 Binzmühlestr. 14, CH-8050 Zürich
 volk@cl.uzh.ch



The Text+Berg project at the University of Zurich is a collaborative effort. The group is headed by **Martin Volk**, a Professor of Computational Linguistics, and linguist **Noah Bubenhofer** then at the UZH German Department now at IdS Mannheim. Both are interested in multilingual corpus linguistics and its applications. Martin Volk is the co-founder of TextShuttle GmbH, a company focusing on language technology for the media industry, and Noah Bubenhofer is a project leader of "semtracks", the Laboratory for Computer Based Meaning Research. They are assisted by **Adrian Althaus**, a student of Philosophy and Computational Linguistics at the University of Zurich who is the main coordinator of the Text+Berg project. **Maya Bangerter**, a dedicated mountaineer, complements the Text+Berg management team. She is a PhD student in Computational Linguistics and coordinator of the Master program in Multilingual Text Analysis.