Ranking Georeferences for Efficient Crowdsourcing of Toponym Annotations in a Historical Corpus of Alpine Texts

Janis Goldzycher
University of Zurich
janis.goldzycher@uzh.ch

Isabel Meraner
University of Zurich
isabel.meraner@wsl.ch

Martin Volk
University of Zurich
volk@cl.uzh.ch

Simon Clematide
University of Zurich
siclemat@cl.uzh.ch

Abstract

This paper presents a simple method to rank georeference candidates to optimally support the workflow of a citizen science web application for toponym annotation in historical texts. We implement the general idea of efficient crowdsourcing based on human and artificial intelligence working hand in hand. For named entity recognition, we apply recent neural pretraining-based NER tagger methods. For named entity linking to geographical knowledge bases, we report on georeference ranking experiments testing the hypothesis that textual proximity indicates geographic proximity. Simulation results with online reranking that immediately integrates user verification show further improvements.

1 Introduction

Named entity recognition (NER) in texts (Nadeau and Sekine, 2007) is an established and crucial task in Information Extraction (Tjong Kim Sang and De Meulder, 2003; Weissenbacher et al., 2019). The recognition of *toponym mentions*, i.e. the detection of names for geographical entities of interest such as cities, mountains, rivers, regions, etc. typically relies either (a) on gazetteer lookup and rule-based pattern matching techniques, which are hand-crafted by language and domain experts, or (b) on supervised machine learning methods for sequence labeling, which need annotated task-specific in-domain training material for good performance. The main problems of NER in general are insufficient coverage of gazetteers or lack of in-domain training material, geo/non-geo ambiguities and the number of entity classes that need to be distinguished.

Named entity linking (NEL) of toponyms is normally cast as a consecutive task to NER and consists in annotating each toponym mention with a unique identifier from a domain-specific knowledge base. This linking of toponyms, also known as toponym resolution (TR) (Leidner, 2007) or Geocoding[1] (Gritta et al., 2019), "grounds" the mentions in georeferences of geographic ontologies, which in turn provide points or complex polygons in a geographic coordinate system. These shapes can then be used for geovisualization of the toponyms on a map (Figure 1).

The main problems of toponym resolution are the ambiguity of toponym names (e.g. in Switzerland alone there are 12 mountains called *Schwarzhorn*), and especially in the case of historical texts, the renaming of geographical entities over time and changes in the spelling of names, which leads to insufficient coverage of name variants even in large contemporary geographical ontologies. For our corpus, additional problems arise from multilinguality, (a) because many geographical entities genuinely have more than one name due to the multilingual cultural background of Switzerland, and (b) because we are dealing with a multilingual text corpus.

The historical corpus of Alpine texts for which we aim at a complete, fine-grained and precise toponym annotation consists of the early yearbooks of the Swiss Alpine Club (SAC) published since 1864 (Göhring and Volk, 2011). Mostly written in German and French, it contains mountaineering reports, scientifically oriented contributions written for an interested lay public and club news, thus constituting a highly valuable domain-

[1] The term Geotagging corresponds to NER tagging restricted to location names.

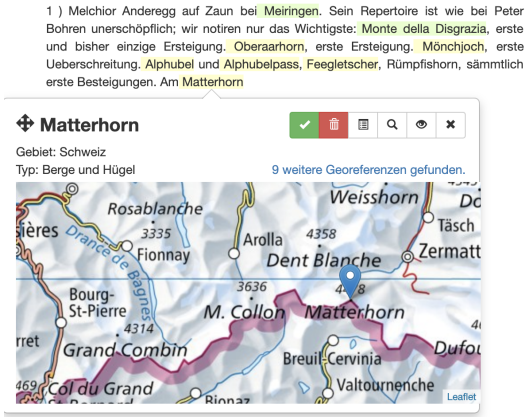

Figure 1: Efficient in-text validation of top-ranked toponym references in the crowdsourcing web interface where people can read pages from yearbooks and annotate (verify, add, delete) georeferences. A mouse click on the unverified toponym mention "Matterhorn" (yellow background) opens a map view where in the best case a second click concludes the verification.

specific resource when geographically fully indexed and publicly available.

Given the difficulties of toponym annotation in domain-specific historical texts, *automatic* toponym resolution methods are not able to achieve the desired performance. For NER using a domain-specific rule-based system, Kew et al. (2019) report a recall of 63% and a precision of 88% when evaluating on 1300 sentences sampled from the full corpus (1864-2015). For a modern neural NER approach (Akbik et al., 2018), a recall of 71% and precision of 87% is reached when using the output of the rule-based system as a silver quality training corpus and roughly 800 manual corrections. As the quality of NEL is bound by NER, providing more and better training material is crucial for achieving higher performance. In order to do so, crowd-sourcing toponym annotations seems promising given the positive experience of asking the SAC community to crowd-correct the OCR errors in this corpus (Clematide et al., 2016).

However, the task of toponym annotation is more complex and knowledge-intensive than OCR correction. Our goal is to provide an efficient workflow that ensures that automatic pre-annotation and human correction from citizen scientists profit from each other as early as possible. For NER, this means to retrain the neural NER models regularly and to update the pre-annotations without interfering with already curated material. Recent neural NER taggers (Akbik et al., 2018) with language modeling pretraining have modest requirements for task-specific training material. In the interface, we additionally adapted our original correction workflow where NER and NEL were hitherto closely intertwined, now allowing for corrections restricted to NER (mentions and toponym types) if preferred by the user.

For NEL, it means to minimize the user's efforts to identify the correct georeference of a toponym mention. Ideally, the NEL component should (a) precompute all possible georeference candidates for a mention (taking into account typical spelling variations) in order to free the user from performing time-consuming knowledge base queries on his own, and (b) rank these candidates such that the true reference appears first on the list. Figure 1 illustrates the intended setup for linking the mention "Matterhorn" to its intended georeference.[2] Verifying a suggested georeference candidate by a single click is far less time consuming than searching through a long unordered list (sometimes up to 70 candidates).

The remainder of this paper reports on simple and efficient methods to optimally rank georeference candidates for toponym annotation based on the principle of textual and geographical proximity (Buscaldi and Rosso, 2008; Buscaldi, 2011).

2 (Re)Ranking Georeference Candidates

We investigate two scenarios: (a) *ranking* candidates using as only evidence automatically computed georeference candidates, and (b) dynamically *reranking* candidates simulating a human validation process where the automatic rankings and the human corrections serve as iteratively improving evidence. The original ranking happens offline during automatic NEL. In contrast, the reranking happens online during the annotation process and can be done in the client's web browser.

The ranking algorithms for both scenarios rely on the hypothesis that textual proximity indicates geographic proximity (Buscaldi, 2011). Both scenarios make use of this hypothesis by applying a point system that rewards the target candidate (sitting at the center of a sliding window) that is geographically closest to a toponym candidate from a context position of the sliding window. Figure 2 illustrates, how georeference candidate 2 of the

[2] Our citizen science web application https://www.geokokos.ch currently features the offline ranking of georeferences.

"…*Mönchjoch*, erste Ueberschreitung. ***Alphubel*** und *Alphubelpass*, *Feegletscher*, *Rümpfishorn*, sämmtlich erste Besteigungen."

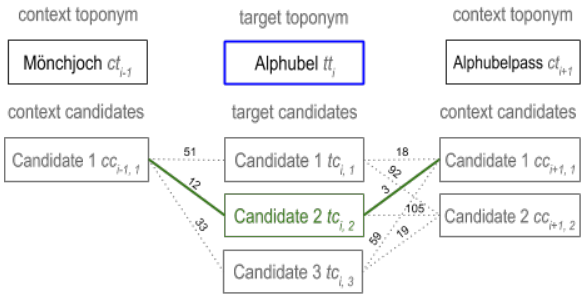

Figure 2: Candidate ranking with window size $n = 1$ of the text snippet shown above. Scores are assigned by building a set of candidate pairs for each toponym in the context of the target toponym and by rewarding target-context pairs with the smallest distance.

ambiguous toponym ***Alphubel*** is rewarded twice due to its smallest distance to all other candidates in a context window of $n = 1$ toponyms. In other words, each context toponym "votes" for the target toponym candidate with the smallest distance.

Ranking. More formally (see Algorithm 1), given a **t**arget **t**oponym tt_i at position i we determine the set of **c**ontext **t**oponyms $\{ct_{i-n}, ..., ct_{i-1}, ct_{i+1}, ..., ct_{i+n}\}$. Then, **t**arget **c**andidate tc_j indexes all admissible georeference candidates of the target toponym tt_i, and for each context toponym ct_k the **c**ontext **c**andidate cc_{kl} indexes all its admissible georeferences. The score of every target candidate tc_j is initialized with 0, and for each context position k, the score of tc_j with the smallest distance of all target/context pair (tc_j, cc_{kl}) is incremented by 1. Thus, for a given context size n containing $2n$ toponyms, $2n$ is the maximum candidate score in a window.

Aggregating all window scores over a yearbook results in a single global score for each georeference. Our final scoring normalizes the yearbook scores of each georeference into the range $[0, 1]$ and multiplies it with the score of each candidate georeference from the local window. In this way, the overall prominence of a georeference in a yearbook (in early years, each SAC yearbooks had one mountain region as a main topic) is combined with the proximity in a "local story" told within the context window. Candidates are then sorted in descending order according to the final total score in order to produce the candidate ranking.

"Wir fliegen mit dem Blick über den [Tschingelgletscher] hin und einen Moment verweilen wir bei der jähen Gneistafel des Lauterbrunner ***Breithorns***, welches, scharf in seinen breiten Gräten, uns nur kahle Platten zeigt und von dessen Fuss einige sekundäre Gletscher in's *Lötschenthal* herabhängen. Ueber [Ebene Fluh], *Grosshorn* und *Gletscherhorn* fliegen wir neuerdings hinweg, senken wieder den Blick in den grossen Ocean des [Aletschfirns] und stehen gebannt vor den scharfen Formen der *Jungfrau* […]".

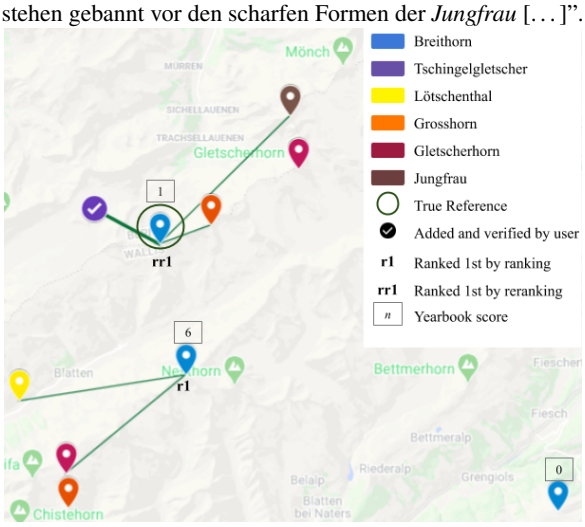

Figure 3: Candidate ranking and reranking for "Breithorn" in the text snippet shown above. Detected and linkable toponyms are in italics. Toponyms in brackets were not automatically linked with candidates and are thus not included in the ranking. But they are included in the reranking if they appear in the left window of the target toponym. Two target and four context candidates are too far away to be included in the map. Context-target candidate pairs with closest distances resulting in rewards are connected with thin green lines. The yearbook scores of the target entities are indicated above the entity's marker.

Dynamic Reranking As soon as humans correct a georeference, new information is available that can be used to update and improve existing candidate rankings on the fly and to further minimize the effort of a crowd corrector. In order to assess the expected benefit, we define the following correction simulation strategy that assumes the user to correct all toponyms in reading order of the text. Each time a user verifies a reference candidate, we update the candidate ranking of the following toponym.

The dynamic reranking is also based on window and yearbook scores and only differs in the following aspects: (a) The yearbook scores are not updated by the window scores. (b) The window score rewards by 10 points instead of 1 point if a verified candidate is involved. (c) The window score rewards by 3 points if only one candi-

Algorithm 1 Candidate Ranking

Input: pages P, window size n
Output: pages P
initialize yearbook scores ys
for each $page \in P$ **do**
  for each target toponym $tt_i \in page$ **do**
    initialize window score map ws
    $tc \leftarrow$ get_candidates(tt_i)
    $ct \leftarrow$ get_context_toponyms(tt_i, n)
    for each context toponym $ct_k \in ct$ **do**
      $cc_k \leftarrow$ get_candidates(ct_k)
      for each context candidate $cc_{kl} \in cc_k$ **do**
        for each target candidate $tc_j \in tc$ **do**
          set $tc_{\hat{j}}$ to tc_j if closest to cc_{kl}
        end for
      end for
      $ws\left[tc_{\hat{j}}\right] \mathrel{+}= 1$
    end for
    update page with ws
  end for
  increment ys by ws
end for
update P with ys
sort candidates of P by ys and ws
return P

date is involved. Figure 3 shows how candidates for **Alphubel** are ranked and reranked. Both candidates are assigned the same window score but the southern central candidate has a higher yearbook score and is thus ranked first by the ranking. When a user reads the sentence and adds and verifies **Tschingelgletscher**, which is close the correct candidate, the dynamic reranking updates the candidate positions and ranks the correct candidate on the first place.

3 Ranking and Reranking Experiments

We test the quality of our ranking method on German pages of the yearbook 1864 and 1874. Our NEL uses two different geographical ontologies, SwissNames3D[3] for toponyms within Switzerland and GeoNames[4] for all others. We made this choice in order to achieve maximal coverage in Switzerland and to avoid linking ambiguity due to multiple knowledge bases. For linking with SwissNames3D, only 23 relevant entity types out of 103 are used[5], for GeoNames 127 out of 676 entity types (feature codes) are used. Table 1 reports the number of toponym candidates and their ambiguity. Note that 30% of the toponyms cannot be resolved by the NEL, and therefore, they do not contribute to the ranking.

[3] https://shop.swisstopo.admin.ch/en/products/landscape/names3D

[4] https://www.geonames.org

[5] We exclude field names (traditional "Flurnamen" in German) due to their extensive ambiguity.

For our experiments, we randomly sampled 20 pages from the yearbook 1864 and 1874 that contain at least 4 ambiguous toponyms and manually resolved all ambiguous cases. Additionally, we invested roughly one hour per page to verify or add other toponyms on the page.

Evaluation Systematically evaluating NEL systems is still a challenging task (Rosales-Méndez, 2019). In our case, we focus on the improvement of the candidate ranking, therefore, considering only the cases where the true georeference is actually one of the proposed candidates. Deleted or newly added toponyms do not appear in our evaluation statistics.

In Table 2, we report results for 3 different ranking conditions: **Randomized** (rand.) is a baseline that shuffles the candidates arbitrarily. **Ranking** (rank.) reports the outcome of the proximity ranking algorithm. **Reranking** (rerank.) shows the results of our dynamic reranking derived from the correction simulation. Our evaluation measure reflects the overall frequency of a correct georeference being ranked first (labeled as rank1), second (rank2), third (rank3) or below rank three (rank4+).

Additionally, we report the relative improvement of ranking in comparison to random shuffling, the improvement of reranking in comparison to ranking and relative error reductions. Further, for comparability, we report the mean reciprocal rank (MRR) of the true references. For a given set of ranks of true references R, the MRR is computed as

$$\mathrm{MRR}(R) = \frac{1}{|R|} \sum_{i=0}^{|R|} \frac{1}{R_i} \tag{1}$$

with $R_i \in [1, 4]$ because we map all rank positions > 4 to 4 for consistency with the absolute ranks reported (rank1 to rank4+).

An important hyperparameter of our approach is the sliding window size n. We evaluated our ranking system with values between 1 and 10 and decided to use a size of 4, which is efficient to compute and performs as well as larger windows. Table 3 shows the rank1 and MRR results for varying window sizes.

Discussion We see that the simple ranking algorithm works pretty well in general. Especially for the yearbook sample 1864, there is a stark relative improvement over random shuffling of 267%.

	1864			1874		
	SN	GN	∑	SN	GN	∑
# topo			2496			3618
topo w/o georef			749			1078
topo w/ georef	1594	153	1747	2102	438	2540
-ambig	1362	69	1431	1835	246	2081
+ambig	232	84	316	267	192	459
+ambig (in %)	15	55	18	13	44	18

Table 1: Toponym statistics for our automatic NER and NEL in German articles of yearbooks 1864 and 1874 using the geographical databases *SwissNames3D* (SN) and *GeoNames* (GN).

	1864						1874					
	rand		rank		rerank		rand		rank		rerank	
	#	%	#	%	#	%	#	%	#	%	#	%
rank1	30	34	80	90	85	96	11	22	18	37	34	71
rank2	27	30	3	3	2	2	21	43	22	45	12	25
rank3	12	14	1	1	1	1	5	10	3	6	0	0
rank4+	20	23	5	6	1	1	12	25	6	12	3	4
total	89		89		89		49		49		49	
MRR	.59		.93		.97		.53		.64		.84	
Rel. impr. rank1			267		106				164		189	
Rel. error reduction			85		56				18		55	

Table 2: Evaluation of candidate ranking methods based on textual and geographical proximity hypothesis. The context window size for these results is 4. Relative improvement is computed on rank1 results (comparing rand to rank and rank to rerank). Relative error reduction is analogous.

Reranking then cannot improve much more on top of that. For 1874, ranking works decently, but leaves many true georeferences on second position. Reranking almost doubles the number of rank1 rankings. Reranking also reduces the number of rank3 and rank4+ cases in comparison to ranking. The poor ranking performance in 1874 is probably due to several ambiguous toponym occurrences where a lot of the surrounding named entities were not found by the NER component initially. The reranking based on incremental user corrections alleviates this problem.

It is interesting to note that by qualitatively looking at reranking errors we could detect several errors in the initial ground truth. A next step for improving the ranking is probably the inclusion of external prominence features (population size, existence of a Wikipedia page, etc.) directly available from some of our geographical knowledge bases.

4 Conclusion

We have shown that a simple ranking approach using a sliding window of 4 is an effective way to profile the intended georeferences on top positions in our two test sets. The quality of our automatic preannotation in historical texts is low enough to profit from a dynamic reranking that integrates human verification as early as possible into the georeference suggestions prominently presented to the user. In crowdsourcing, human and artificial intelligence should work hand in hand in order to efficiently produce high-quality annotations. The disambiguation of rank1 cases using a two-click verification speeds up the process and leaves more time for citizen scientists to address the difficult toponym resolution problems that need real detective work.

	1864						1874					
n	1	2	3	4	5	10	1	2	3	4	5	10
rank1	75	80	83	85	84	84	35	30	35	34	33	33
MRR	.91	.94	.96	.97	.97	.97	.85	.80	.85	.84	.83	.84

Table 3: Comparison of reranking performance with increasing values for window size n.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1638–1649, Santa Fe, NM, USA.

Davide Buscaldi. 2011. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19.

Davide Buscaldi and Paolo Rosso. 2008. Map-based vs. Knowledge-based Toponym Disambiguation. In *Proceedings of the 5th Workshop on Geographic Information Retrieval (GIR)*, pages 19–22.

Simon Clematide, Lenz Furrer, and Martin Volk. 2016. Crowdsourcing an OCR Gold Standard for a German and French Heritage Corpus. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Ann Göhring and Martin Volk. 2011. The Text+Berg Corpus: An Alpine French-German Parallel Resource. In *Proceedings of the 18th Traitement Automatique des Langues Naturelles Conference (TALN)*, Montpellier, France.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2019. A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, pages 1–30.

Tannon Kew, Anastassia Shaitarova, Isabel Meraner, Janis Goldzycher, Simon Clematide, and Martin Volk. 2019. Geotagging a diachronic corpus of

alpine texts: Comparing distinct approaches to toponym recognition. In *Proceedings of the Workshop on Language Technology for Digital Historical Archives*, pages 11–18, Varna, Bulgaria. INCOMA Ltd.

Jochen L. Leidner. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Henry Rosales-Méndez. 2019. Towards better entity linking evaluation. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 50–55, New York, NY, USA. Association for Computing Machinery.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL) at HLT-NAACL - Volume 4*, pages 142–147, Edmonton, Canada.

Davy Weissenbacher, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2019. SemEval-2019 task 12: Toponym resolution in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 907–916, Minneapolis, Minnesota, USA. Association for Computational Linguistics.