

Extracting Imprecise Geographical and Temporal References from Journey Narratives

Ignatius Ezeani^{1,*}, Paul Rayson¹ and Ian Gregory²

¹UCREL, School of Computing and Communications, InfoLab21, Lancaster University, Lancaster, LA1 4WA, UK

²Department of History, Lancaster University, Lancaster, LA1 4YT, UK

Abstract

Previous approaches to understanding geographies in textual sources tend to focus on geoparsing to automatically identify place names and allocate them to coordinates. Such methods are highly quantitative and are limited to named places for which coordinates can be found, and have little concept of time. Yet, as narratives of journeys make abundantly clear, human experiences of geography are often subjective and more suited to qualitative representation. In these cases, “geography” is not limited to named places; rather, it incorporates the vague, imprecise, and ambiguous, with references to, for example, “the camp”, or “the hills in the distance”, and includes the relative locations using terms such as “near to”, “on the left”, “north of” or “a few hours’ journey from”. In this demo paper, we describe our research prototype to extract and analyse qualitative and quantitative references to place and time in two corpora of English Lake District travel writing and Holocaust survivor testimonies.

Keywords

Named Entity Recognition, Imprecise locations, English Lake District corpus, Holocaust survivor corpus

1. Introduction

Human experiences which were recorded and communicated as historical text are increasingly available as digital corpora. A major challenge for researchers in the social sciences, humanities and computer sciences is how to use these texts in interdisciplinary settings to develop cohesive understandings of the historical experiences described. Understanding geographies in historical sources has received a significant amount of research interest in recent years across fields as diverse as geographical information science (GISc), corpus linguistics, natural language processing (NLP), human geography, literary studies, and digital humanities. The current state-of-the-art involves using geoparsing to automatically identify the place names in texts and allocate them to a coordinate [1]. Once georeferenced in this way, place names can be read into a geographical information system for mapping and spatial analysis. Analysis can

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story’23 Workshop, Dublin (Republic of Ireland), 2-April-2023

*Corresponding author.

✉ i.ezeani@lancaster.ac.uk (I. Ezeani); p.rayson@lancaster.ac.uk (P. Rayson); i.gregory@lancaster.ac.uk (I. Gregory)


🌐 <https://www.lancaster.ac.uk/scc/about-us/people/ignatius-ezeani> (I. Ezeani);

<https://www.lancaster.ac.uk/scc/about-us/people/paul-rayson> (P. Rayson);

<https://www.lancaster.ac.uk/staff/gregoryi/> (I. Gregory)

🆔 0000-0001-8286-9997 (I. Ezeani); 0000-0002-1257-2191 (P. Rayson); 0000-0001-8745-2242 (I. Gregory)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

also be conducted using techniques from corpus linguistics and NLP to see what words or themes are associated with the place name such as the place being associated with emotional responses such as being beautiful or inspiring fear. This combination of approaches is known as geographical text analysis (GTA) [2].

While GTA provides a useful starting point for understanding the geographies within a corpus, it is highly quantitative, is limited to named places for which coordinates can be found, and has little concept of time. Journey narratives describing human experiences are typically subjective, and places are framed not just in terms of named places, but rather in imprecise terms describing the setting or landscape e.g. “the camp”, “the majestic mountains”, and feature relative terms e.g. “a quick detour along the lake”, “turn left after the inn”. These non-mappable qualitative representations can be important features of the narratives but cannot be managed within geospatial technologies such as GIS. To understand the ways in which humans describe and relate to the world around them, we need to be able to visually represent and interpret the geographies authors and interviewees describe in ways that combine the qualitative nature of described spatial experiences with methods that render them quantitatively analysable.

In this demo paper, we will illustrate how we are extending current GTA techniques and have applied them to analyses of two large corpora: one a corpus of travel writing about the English Lake District, predominantly written in the 18th and 19th centuries; the other, a corpus of Holocaust survivor testimonies. Although based on very different types of journeys, leisure travel, and forced migration respectively, both corpora represent a collection of unique voices that coalesce to generate complex cultural and experiential geographies. The NLP research described here is part of a larger project which also incorporates methods from corpus linguistics, Qualitative Spatio-Temporal Reasoning (QSTR), GISc, and visual analytics which can help us understand how authors themselves represented the geographies that surrounded them and explore the individual and aggregate representation of the sense and experience of place that these texts contain.

The overall aim of our project is to develop techniques to learn more about the spatial and temporal information contained in a piece of writing without any prior knowledge of the geography of the places mentioned in the text. A starting point will be to automatically identify references to toponyms (‘Penrith’, ‘Pooley Bridge’, ‘River Lowther’) or geographical feature nouns (‘the town’, ‘a hill’, ‘the road’). We also want to extract interesting relationships between places (‘Pooley Bridge is about *six miles* away from Penrith’) as well as some sense of the place (‘The scenery around this lake is *tame*, but *pleasing*’). The key objective of the NLP methods shown in this demo is to identify and extract all quantitative and qualitative references to place, time, and their relationships, as shown in Figure 1.

2. Dataset and Methods

2.1. Corpus of Lake District Writing

The dataset used for this work is the Corpus of Lake District Writing (CLDW)¹ which comprises eighty texts and around 1.5 million words that describe the Lake District [4]. The earliest texts

¹CLDW and the gold standard dataset [3] are available here: <https://github.com/UCREL/LakeDistrictCorpus>

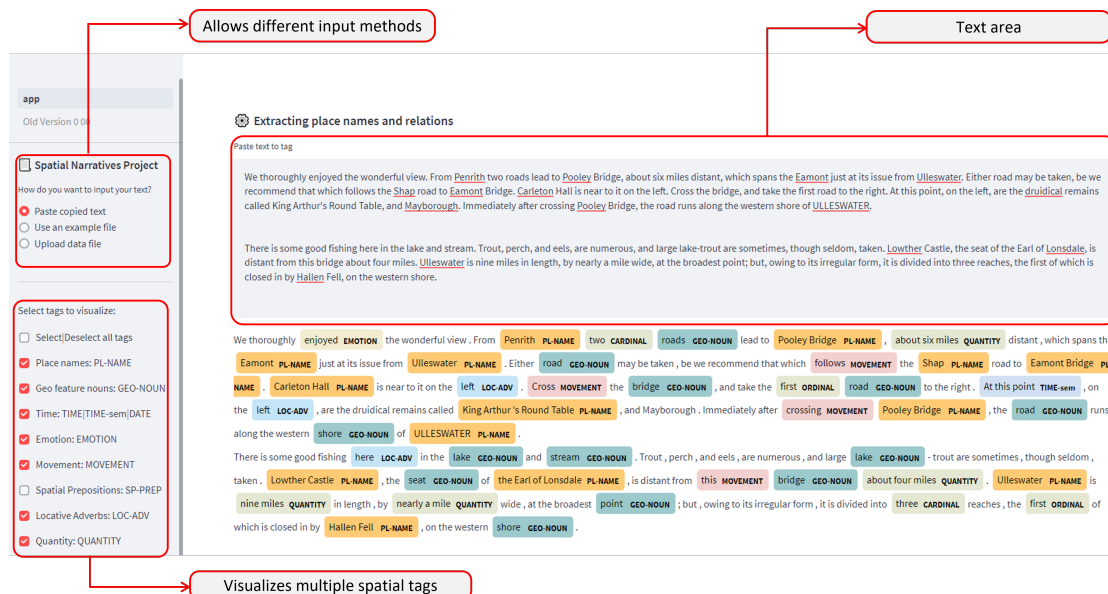


Figure 1: A screenshot of the demo interface describing the key functionalities of the current version of the demo. The app allows three input methods - pasting copied text, using an example file within the demo, and uploading a file from your local machine. You can also select all or some of the featured spatial tags - *PL-NAME*, *GEO-NOUN*, *TIME*, *DATE*, *EMOTION*, *MOVEMENT*, *SP-PREP*, *LOCADV*, and *QUANTITY*.

are from the seventeenth century and run through to the early twentieth century. It includes works by well-known Lake Poets such as William Wordsworth and Samuel Taylor Coleridge. There are also accounts of visits to the Lake District by prominent writers such as Daniel Defoe and Celia Fiennes and other less well-known writers. There are also a number of tourist guides stretching from Thomas West's (1778) "A Guide to the Lakes" to Black's (1900) "Shilling Guide to the English Lakes" [5]. While drawn from a variety of styles and genres, the majority of the corpus comprises tourist guides and travel narratives.

2.2. Extraction Pipeline

This work adopts an extraction pipeline, illustrated in figure 2, that searches for spatial elements (toponyms, geographical feature nouns, distances, and others) in the text by combining three text processing techniques in a systematic manner. The first approach is a basic rule-based method that uses hand-crafted regular expression (regex)² rules and a list of spatial items of interest to extract and tag them in text. A list of 4,227 Lake District place names from the CLDW project was used to extract the toponyms while a compilation of 138³ geographical feature nouns were used. Spatial prepositions and locative adverbs (*'aboard'*, *'between'*, *'beyond'*, *'down-town'*, *'northbound'*, *'overseas'*, etc) were also extracted and tagged.

²A regular expression [6] is a sequence of characters that specifies a search pattern in the text. See https://en.wikipedia.org/wiki/Regular_expression

³Inflections and lemmas (i.e. *road* and *roads*) extended the list to 262



Figure 2: A description of the extraction pipeline containing the rule-based method based on regular expression, named entity recognition with *spaCy* NLP library and semantic tagging with UCREL’s PyMUSAS.

However, this rule-based method has limitations. It leaves out some known place names. This is because the names were not on the list, wrongly spelt, or inconsistently capitalised. Furthermore, we needed to extract temporal references as well which was not possible with the regex method. *spaCy*’s⁴ named entity recognition (NER) feature was applied to mitigate these challenges. An NER tool identifies entities based on their context and does not require a gazetteer. It is also able to capture temporal references as well as other entities in text. The application of named entity recognition in spatial analysis of unstructured text is quite common. Amine et al [8] compared the performance of different named entity recognition models on the task of identifying spatial nominal entities (e.g. village, hut, church) from manually Wikipedia articles. Mehtab Alam Syed et al., [9] demonstrated that NER can be successfully applied to relative spatial information extraction (e.g. south of Paris, 80km from Rome). Lucie Cadorel et al., [10] also achieved over 95.7% accuracy on spatial relations extraction using an NER based extraction model.

One way to get a sense of the place is to identify elements that indicate sentiments or emotions expressed and activities mentioned in discussions about a place and this needed further development beyond the rule-based and NER methods. Hence, UCREL’s⁵ USAS semantic tagger [11] was included in the pipeline specifically to identify elements that are semantically tagged **E**: ‘*emotion*’, **M**: ‘*movement, location, travel and transport*’ and **T**: ‘*time*’.

3. Evaluation

The performance of the extraction tool was evaluated (currently only for toponyms) with the gold standard subset of the CLDW which contains 28 texts that were carefully selected to be representative of the corpus. It contains 242,000-word tokens i.e. about one-sixth of the entire corpus. The placenames included the names of a variety of different regional, national and international locations, landmarks and geographical formations marked up with a customised tag <cdplace>. The evaluation results on the gold-standard data show that our placename extraction method has a precision score of **100%** with an overall average recall score of **93.95%** thereby recording an F1 score of **96.88%**.

⁴*spaCy* is a free and open-source python library for general NLP [7].

⁵UCREL is the University Centre for Computer Corpus Research on Language at Lancaster University. See <https://ucrel.lancs.ac.uk/usas/> for the description of the top-level tags. Here we used PyMUSAS, an open-source Python implementation of the semantic tagger for English and other languages: <https://pypi.org/project/pymusas/>

4. Conclusion

In this demo paper, we have illustrated our prototype NLP pipeline for extracting imprecise geographical and temporal references from journey narratives, focussing on travel writing about the English Lake District. We have made our demo available open source as a series of Python Notebooks on our project’s GitHub repository.⁶

In future work, we will extend the gold standard corpus to include geographical feature nouns, spatial prepositions, and other non-mappable terms, in order to evaluate our pipeline’s ability to identify them. We will also evaluate our pipeline on the Holocaust dataset, which we expect to be more challenging as it consists of interview transcripts and therefore potentially less amenable to existing methods.

Also, language understanding in general faces the challenge of establishing the relationship between the vertices of the *semantic triangle* - the *language* (symbol), the *object* (or referent) in the real world that it describes, and the human *thought* (or reference) [12]. However, for spatial language representation, Stock et al. [13] proposed an extension of the semantic triangle to a *semantic pyramid* with the digital face and vertices for knowledge and structured language representation. Our future work will also explore and compare different approaches to spatial semantic modelling in the annotation of our datasets.

Acknowledgments

We thank the anonymous reviewers for their comments on our paper submission. The project is funded in the UK from 2022 to 2025 by ESRC, project reference: ES/W003473/1. We also acknowledge the input and advice from the other members of the project team in generating requirements for our research presented here. More details of the project can be found on the website: <https://spacetime narratives.github.io/>

References

- [1] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, J. Ball, Use of the Edinburgh geoparser for georeferencing digitized historical collections, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368 (2010) 3875–3889.
- [2] C. Porter, P. Atkinson, I. Gregory, Geographical text analysis: A new approach to understanding nineteenth-century mortality, *Health Place* 36 (2015) 25–34. doi:10.1016/j.healthplace.2015.08.010.
- [3] P. Rayson, A. Reinhold, J. Butler, C. Donaldson, I. Gregory, J. Taylor, A deeply annotated testbed for geographical text analysis: The corpus of lake district writing, in: *GeoHumanities’17 Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities, Association for Computing Machinery (ACM), 2017*, pp. 9–15. doi:10.1145/3149858.3149865.
- [4] J. E. Taylor, I. N. Gregory, *Deep Mapping the Literary Lake District: A Geographical Text Analysis*, Rutgers University Press, 2022.

⁶<https://github.com/SpaceTimeNarratives>

- [5] I. Gregory, C. Donaldson, A. Hardie, P. Rayson, Modeling space in historical texts, in: *The Shape of Data in the Digital Humanities*, Routledge, 2018, pp. 133–149.
- [6] J. Goyvaerts, Regular Expressions Tutorial, <https://web.archive.org/web/20161101212501/http://www.regular-expressions.info/tutorial.html>, 2016. Accessed: 2022-10-10.
- [7] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017.
- [8] A. Medad, M. Gaio, L. Moncla, S. Mustière, Y. Le Nir, Comparing supervised learning algorithms for spatial nominal entity recognition, *AGILE: GIScience Series 1* (2020) 15. URL: <https://agile-giss.copernicus.org/articles/1/15/2020/>. doi:10.5194/agile-giss-1-15-2020.
- [9] M. A. Syed, E. Arsevska, M. Roche, M. Teisseire, Geotag: Relative spatial information extraction and tagging of unstructured text, *AGILE: GIScience Series 3* (2022) 16. URL: <https://agile-giss.copernicus.org/articles/3/16/2022/>. doi:10.5194/agile-giss-3-16-2022.
- [10] L. Cadorel, A. Blanchi, A. G. Tettamanzi, Geospatial knowledge in housing advertisements: Capturing and extracting spatial information from text, in: *Proceedings of the 11th on Knowledge Capture Conference*, 2021, pp. 41–48.
- [11] P. Rayson, D. Archer, S. Piao, T. McEnery, The UCREL Semantic Analysis System, in: *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 2004, pp. 7–12.
- [12] C. Ogden, I. Richards, *The Meaning of Meaning—A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Magdalene College, University of Cambridge, 1923.
- [13] K. Stock, C. B. Jones, T. Tenbrink, Speaking of location: a review of spatial language research, *Spatial Cognition & Computation* 22 (2022) 185–224. doi:10.1080/13875868.2022.2095275.