Customising Geoparsing and Georeferencing for Historical Texts

C.J. Rupp, Paul Rayson, Alistair Baron, Christopher Donaldson,
Ian Gregory, Andrew Hardie, Patricia Murrieta-Flores
Lancaster University
Lancaster, UK
crupp@lancaster.ac.uk

***Abstract*—In order to better support the text mining of historical texts, we propose a combination of complementary techniques from Geographical Information Systems, computational and corpus linguistics. In previous work, we have described this as 'visual gisting' to extract important themes from text and locate those themes on a map representing geographical information contained in the text. Here, we describe the steps that were found necessary to apply standard analysis and resolution tools to identify place names in a specific corpus of historical texts. This task is seen as an initial and prerequisite step for further analysis and comparison by combining the information we extract from a corpus with information from other sources, including other text corpora. The process is intended to support close reading of historical texts on a much larger scale by highlighting using exploratory and data-driven approaches which parts of the corpus warrant further close analysis. Our case study presented here is from a corpus of Lake District travel literature. We discuss the customisations that we have to make to existing tools to extract placename information and visualise it on a map.**

***Keywords*-Text mining; Toponym Resolution; Historical Corpora**

I. Introduction

Humanities disciplines place the study of texts at their centre. However, with an exponential rise in 'born digital' material such as email, the World Wide Web and social media, as well as vast digitisation activities for historical paper sources, there is a question over whether textual study can be scaled up to make use of such material. Models of text mining large data sets used elsewhere for summarisation, information extraction and retrieval purposes do not match well the requirements of the humanities scholar to whom the importance of close reading is key to understand the subtleties and nuances within the text. The research described in this paper fits into a larger project which is exploiting techniques from computational and corpus linguistics to extract spatial information from textual sources and analyse and visualise it using Geographical Information Systems (GIS), e.g. [1]. The combination of these complementary techniques is aimed to utilise the spatial information contained in the corpora, which would be difficult, if not impossible, to read in their entirety. This process will also allow us to determine which parts of the corpora we should read closely and how these parts relate to the remaining unread material. The research questions that are of interest in the overall project tend to revolve around three areas: (i) where is the corpus talking about? (ii) what is the corpus saying about these places? (iii) what is the corpus saying about specific themes e.g. health and disease in proximity to these places? In the work described in this short paper, we investigate the possibilities for applying text analysis techniques to large historical collections. We would like to be able to draw on the archive of a medieval chancellery or a 19th Century newspaper in the same way as we would the archive of a Twitter feed. A significant barrier to this ideal is immediately apparent in the diversity of the medium in which historical data is typically presented. Rather than a predefined protocol, we are presented with a progression of conventions in publishing, formatting, and even the language itself and its transcription. Suppose we regard this jungle of factors as individual hurdles: this paper is focussed on how well we can overcome them and how much additional effort must we invest?

We report on ongoing work in information extraction from a specific corpus of historical texts about the Lake District. We report the tasks that were necessary in geoparsing this corpus, identifying where specific locations are mentioned in the text and determining which geographical location is referred to in each case [2]. This process combines Named Entity Recognition for toponyms, geotagging, with toponym resolution, georesolution. The intended result is a Geographical Information System (GIS), a form of database extended with coordinate information. This represents the extraction of information that is grounded in an external reality, via the coordinate system, and independent of the textual form of its origin. Further analysis can be based on just the GIS, or its use in conjunction with other corpus analysis. Indeed, a GIS can be used as a basis draw together distinct corpora into a uniform whole. At this level, the textual peculiarities of individual subcorpora would no longer be visible, but we can only get there by developing strategies to smooth over the actual properties of collections of texts throughout the analysis processes, in effect, allowing standard processing tools to be applied to a wider range of historical texts. The geographical visualisations and presentations through the corpus-based methods are intended as ways back in to the text to support and direct further close reading activity.

978-1-4799-1293-3/13/$31.00 ©2013 IEEE

II. Geoparsing Historical Text

We focus on the customisation of geoparsing processes to historical texts, because some geoparsing tools are already in use. Indeed, the first phase of geoparsing, the identification of the use of placenames, or geotagging, has been a feature of Named Entity Recognition (NER) systems from their inception. NER was initially developed on news service texts where location was a prominent and crucial type of information. A more complete analysis including a second phase of toponym resolution, or georesolution, using gazetteers encoding coordinate information, was a natural development, while still a fairly challenging task. At least, the existence of a coordinate system places a limit on the resolution requirement for location information, in contrast to, e.g. person names.

Our starting point is the Edinburgh Unlock Geoparser [3], which has previously been applied to historical texts, though mainly from the 19th Century. This is provided as a webservice with a REST interface, and, as such, must provide a general purpose service acceptable to any user anywhere on the internet. The resolution strategy, includes criteria based on current, or at least recent, population figures. These may not be applicable for a historical corpus, as absolute and relative population size varies over time. We clearly have to be prepared to customise both the geotagging and the georesolution phases of the geoparser. A general purpose global gazetteer may even contain places that were not named at the time when a particular text was written.

If we regard the geoparser webservice as a black box, then our scope for action primarily addresses the input texts and the result files. However, the standard result format includes quite a lot of information, including linguistic markup of the text at the word level and candidate lists for each resolved placename. The georesolution result is the preferred candidate within each list, but the list entries provide enough information for revision of that selection, provided a result can be found within the candidates. This would in turn depend on the geotagging and its connection to the gazetteer.

Many of the most immediate obstacles to geoparsing historical text must be dealt with before any standard geoparser process can be applied. Some level of noise in the text is almost inevitable, as the text will, typically, have been digitised either by hand or using OCR (Optical Character Reader) processes. Problems with the quality of OCR results are known, but even the best human transcriptions will generate some errors. An accurate digital form of the text may contain variation in spelling and language use, perhaps even non-standard characters. When we add placenames to the mix things become just a bit more complex, as placenames may not feature in OCR language models or dictionaries used for spelling normalisation. In addition, placenames will have their own variants, which may not even be phonologically related to the modern name.

Table I
Edinburgh Geoparser Geotagging with Standard and Historical Gazetteer Information

Documents	Unlock		DEEP	
	Precision	Recall	Precision	Recall
Overall	91.58	63.68	86.06	74.40
1700-99	87.22	48.97	81.32	57.98
1700-69	93.15	57.30	85.92	66.85
1770-99	86.16	47.62	80.55	56.62
1800-99	94.70	72.25	88.70	72.25
1800-29	96.03	73.63	89.83	84.70
1830-99	92.75	70.23	87.07	83.26

The Edinburgh geoparser relies on a gazetteer for both geotagging and georesolution. The Unlock Text webservice offers two standard gazetteers and we have experimented with both options. We should also consider the possibility of variations or extensions to the standard gazetteer, although this goes beyond what the webservice currently offers. Similarly, a localisation parameter for georesolution is reported, but that options is not supported by Unlock Text. If we refer to the georesolution strategies described in [3], we may need to reimplement some of these working directly on the candidate lists, in order to optimise georesolution for a historical corpus centred on the English Lake District.

III. Geoparsing the Lake District Corpus

Our corpus consists of 81 texts ranging from 1622 to 1900, but predominantly from 18th and 19th centuries, comprising around 1.5M words. The corpus is a mixture of canonical and non-canonical literature about the Lake District and includes guidebooks, travelogues, novels, poems, journals and private letters. These texts have been transcribed from the original, thus avoiding inaccuracies that would tend to arise from OCR processes. While there will still be some minor editing errors introduced during transcription and annotation with XML document markup. The transcribed text also reflects the spelling conventions of the original text and some use of alternative characters, such as a long s. In addition to the transciption of all 81 texts, we were able to annotate place names and person names in a subcorpus of 28 texts (c. 250k words), forming a Gold Standard evaluation set. The (small) number of annotators meant that we did not require inter-annotator agreement statistics.

We initially made use of the REST service of the Edinburgh Unlock parser, which currently supports an HTML and a plain text interface. For simplicity, we prefer the plain text interface. In the conversion script from XML document markup to plain text we also include some normalisation of unusual characters and the more blatant editing errors. We submitted the Gold Standard Corpus via the Unlock Text interface to get an evaluation of baseline performance of the geoparser on the Lake District corpus.

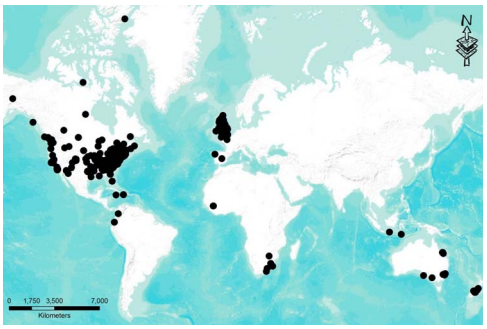

Figure 1. Georesolution Results

IV. Initial Evaluation

An initial evaluation was conducted for placename recognition, optimizing the results by counting any token overlap in the tags of the geoparser result and the hand-coded reference annotations. This reveals an adequate precision, but disappointing recall, see Table I Unlock columns. Although the recall results are not uniform accross the reference corpus. Our suspicion was that variation in spelling and in the actual placenames would be the root cause of this.

We break down the baseline precision and recall results over the 18th and 19th century texts, and then again grouping comparable numbers of documents. The indication is that the 18th century documents give inferior results, but have a much higher proportion of spelling variations. Although at least one of the earlier texts was transcribed from a later edition, so the pattern is not totally systematic.

At the outset we were not able to evaluate the georesolution in detail. However, we generated a simple distribution map (Figure 1) using the resulting GIS. This map reflected the toponym resolution based on a current and general gazetteer, producing a worldwide distribution, rather than the focus on the North West of the United Kingdom, as we expected. On the basis of this, we also prepared a reference table of coordinates for the annotated placenames, which could in turn be used as a specific extension to the online gazetteer. Figure 2 shows the map corresponding to this hand-corrected table.

V. Processing Variation

It seems clear that we will need to account for:

- Spelling variation
- Variant forms of placenames
- Placenames that are specific to the region of the English Lake District

We plan to make use of the VARD system [4] to normalise English spelling. VARD is a variant detector, based on a dictionary of standard current English forms. It generally requires some training for a specific corpus and has most successfully been used on Early Modern English. The prevalence of placenames in our corpus leads us to defer application of VARD until we can account for a substantial range of placenames and their observed variants. While we could train VARD on a subset of our corpus, in much the same way as we have prepared Gold Standard geotagging and georesolution data, we would need to ensure that enough of the placenames were included in that training set to avoid accidentally normalising them as regular spellings.

The normalisation of spelling variants in general will aid the linguistic analysis underlying the geoparser's NER component, but it will not help and may even disrupt the recognition of placename variants or names that are unknown to the current gazetteer. In order to estimate the degree to which recall is disrupted by variant or unknown placenames, we collected several lists of names specific to the Cumbrian region, including the Ordinance Survey and a historic Cumbrian gazetteer. Using this list we tried substituting the closest matching known local placename for the annotated forms in the Gold Standard Corpus. We used a simple partial mapping based on string distance and the metaphone package. This created a version of the texts with numerous errors, but with a proportion of the annotated names replaced with forms more likely to be in the default gazetteer. Geoparsing this version did produce an effect, but this strategy is only really applicable in the evaluation of an NER system, as it relies on knowing which terms are placenames.

Comparing the terms found by the geotagger, the annotated placenames and additional gazetteer lists for the specific region showed that, not only were unseen variants being missed, but also some of the documented names from our external gazetteer lists. This suggests that at the very least we will require a specialised component in the geoparser's gazetteer, augmenting the information it uses by default. We would envisage this as being analogous to the specialised spelling lists used to personalise spelling checkers. In fact, the correction list we prepared as Gold Standard data for the georesolution appears to contain most of the information we would want to add to the gazetteer, including annotated placenames and their preferred coordinates for this corpus.

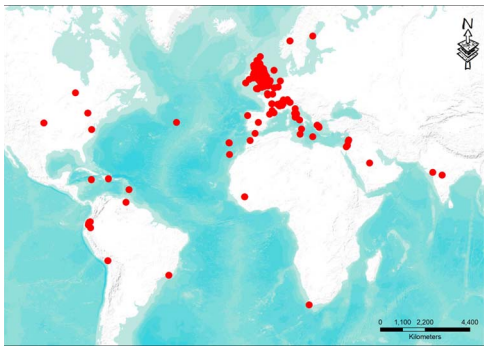

Figure 2. Hand-Corrected Georesolution Results

VI. Customisation

In order to ensure that both the recognition of placenames (geotagging) and their resolution to actual locations with coordinates (georesolution) can be improved over the baseline performance of the Unlock Text webservice, we would envisage several customisation options in the way the geoparser is used. In cooperation with Claire Grover and the Edinburgh group, we have recently gained access to a standalone version of the geoparser for further experimentation. However, some of basic customisation options can be applied to the geoparser treated as a black box.

In particular, any normalisation of the input text can be applied before submitting the text. This would include character normalisation and normalisation of spellings for common words. We would employ appropriately trained version of VARD for this spelling normalisation step, assuming that we can train both a range or 17th and 18th century spellings and a full set of Lake District toponyms and their historical variants.

The other main area for customisation would be the gazetteer that is used both in the NER component and in georesolution, as it contains both the known placenames and their coordinates. We intially saw this mainly in terms of a regionally specific extension to the standard gazetteer, which we have compiled via our Gold Standard reference set and the use of external lists from other gazetteers. More recently, we have benefitted from further cooperation with the geoparser developers in Edinburgh and the DEEP (the Digital Exposure of English Place-Names) project[1]. This has provided access to some of the data underlying a historical gazetteer of English placenames, specifically the Cumberland and Westmorland sections. This provides the additional possibility of Customising the geoparser with a historical gazetteer for our specific region. This work is still in progress, but we have an initial result for geotagging, see DEEP columns in Table I, which show considerable improvement in recall (c. 17%), but some decline in precision (c.7%), suggesting that some of the recognition conditions need to be made more specific.

We have not yet balanced out the use of the possible additional and variant gazetteer lists. It may be most useful to give priority to the historical gazetteer. On the other hand, we may find that we still need the gazetteer list built up from the Gold Standard set on our own corpus. While the focus of our corpus is on a specific area, a wide range of places in various parts of the world are actually mentioned, so that whatever priorities we customise we will still need a full gazetteer list. In a similar way, our resolution strategy would prefer a result within the Cumbrian region, but still allow for the mention of Andean peaks, colonial conurbations and the centres of ancient culture. There is an option in the geoparser which promotes results within a given region, specified as coordinate pair and radius distance. We have not yet fully explored this option, as it is not yet part of the Unlock Text interface, but this appears to be an appropriate customisation of the georesolver for our purposes.

[1] http://www.placenames.org.uk

VII. Conclusion

We have described work in progress to enhance the recognition of placenames in a corpus of Lake District literature and the resolutions of those names to specific locations with coordinates. Beyond specific details of our work we have attempted to present this as an example of how the extension and modification of standard processing tools, via the provision of task specific information, can start to smooth over the peculiarities of historical corpora. This would allow specific text corpora to participate in larger data collections, exploiting the grounding in the real world, via geographical coordinates, that an associated GIS provides. In future work, we intend to move towards a workflow framework where we can experiment with multiple knowledge sources (gazetteers and variant lists) and software components (NER and geoparser) to evaluate their accuracy when dealing with multiple historical corpora.

Acknowledgment

The research leading to these results has received funding from the European Research Council (ERC) under the European Union Seventh Framework Programme (FP7/2007-2013) / ERC grant Spatial Humanities: Texts, GIS, places (agreement number 283850).

We would also like to thank Claire Grover, the geoparser developers and the Unlock Text support team in Edingburgh, Jayne Carroll and the DEEP project, and M. and J. Norgate for putting their work online as: http://www.geog.port.ac.uk/webmap/thelakes/html/lgaz/lgazfram.htm

References

[1] M. Yuan, "Mapping text," in *The Spatial Humanities: GIS and the future of humanities scholarship*, D. J. Bodenhamer, J. Corrigan, and T. M. Harris, Eds. Bloomington: Indiana University Press, 2010, pp. 109–23.

[2] J. L. Leidner and M. D. Lieberman, "Detecting geographical references in the form of place names and associated spatial natural language," *SIGSPATIAL Special*, vol. 3, no. 2, p. 511, July 2011.

[3] C. Grover, R. Tobin, K. Byrne, M. Woollard, S. D. James Reid, and J. Ball, "Use of the edinburgh geoparser for georeferencing digitised historical collections," *Philosophical Transactions of the Royal Society A*, vol. 368, no. 1925, pp. 3875–3889, 2010.

[4] A. Baron and P. Rayson, "Vard 2: A tool for dealing with spelling variation in historical corpora," in *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham, UK, 2008.