Proceedings of he 8th Workshop on Challenges in the Management of Large Corpora, pages 24–30
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Geoparsing the Historical Gazetteers of Scotland: Accurately Computing Location in Mass Digitised Texts

Rosa Filgueira[1], Claire Grover[2], Melissa Terras[3] and Beatrice Alex[2,3]
[1] Edinburgh Parallel Compute Centre
[2] School of Informatics
[3] Edinburgh Futures Institute
University of Edinburgh
r.filgueira@epcc.ed.ac.uk, grover@inf.ed.ac.uk, M.Terras@ed.ac.uk, balex@ed.ac.uk

Abstract

This paper describes work in progress on devising automatic and parallel methods for geoparsing large digital historical textual data by combining the strengths of three natural language processing (NLP) tools, the Edinburgh Geoparser, spaCy and defoe, and employing different tokenisation and named entity recognition (NER) techniques. We apply these tools to a large collection of nineteenth century Scottish geographical dictionaries, and describe preliminary results obtained when processing this data.

Keywords: text mining, geoparsing, historical text, Gazetteers of Scotland, distributed queries, Apache Spark, digital tools

1. Introduction

Ongoing efforts towards the mass digitisation of historical collections mean that digitised historical texts are increasingly being made available at scale for research. This paper describes work in progress on devising automatic and parallel methods for geoparsing large digital historical textual data. Geoparsing means automatically tagging place names in text and resolving them to their correct latitude and longitude coordinates or gazetteer entry. We combine the strengths of three natural language processing (NLP) tools, the Edinburgh Geoparser (Grover et al., 2010)[1], spaCy[2], and defoe (Filgueira et al., 2019)[3], and employing different tokenisation and named entity recognition (NER) techniques. We apply these tools to the Gazetteers of Scotland, a large collection of nineteenth century Scottish geographical dictionaries, and describe preliminary results obtained when processing this data. Our end goals are to develop more accurate geoparsing for such historical text collections but also to make such data accessible to users, in particular scholars who may not have the necessary technical skills to build tools to analyse the text themselves.

2. Background and Related Work

Text mining large historical text collections, and making that text available for others to analyse, has been an activity much pursued at the juncture of Digital Humanities and library and archive digitisation. For example, Clifford et al. (2016) focused on analysing text with respect to commodity trading in the British Empire during the 19th century. Currently, there is a similar effort to develop and apply NLP tools to historical newspapers as part of a variety of projects including Living with Machines[4], The Viral Texts Project[5] and Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914.[6]

In terms of geoparsing historical text, this area of research is relatively specialised, which means that there is limited related work. The Edinburgh Geoparser, one of the tools used for this work, has been previously adapted to work with historical and literary English text (Alex et al., 2015; Alex et al., 2019) and has been further modified or applied to a number of different text datasets (Grover and Tobin, 2014; Rupp et al., 2013; Rayson et al., 2017; Porter et al., 2018) Similar tools have applied geoparsing to historical text in other languages, e.g. historical French literary text (Moncla et al., 2017) and Swedish literary text (Borin et al., 2014).

In the context of Scotland, there is not one comprehensive historical gazetteer available for research as a downloadable resource. There is an online resource called the Gazetteer for Scotland[7] which allows users to search for and find out about places in Scotland but this data is limited to online search access only.

Our challenge here is then threefold: how can we compute spatial characteristics within historical texts? How can we be assured of the accuracy of our approaches? And how can we build our historical gazetteer of Scotland, to provide information and data for others to reuse in research and teaching?

3. The Gazetteers of Scotland

For evaluating our work, we are applying our tools to The Gazetteers of Scotland (see Table 1), a collection of twenty volumes of the most popular descriptive historical gazetteers of Scotland in the nineteenth century.[8] They are considered to be geographical dictionaries and include an alphabetic list of principal places in Scotland, including towns, counties, castles, glens, antiquities and parishes. This dataset was recently made available by the National Library of Scotland on its Data Foundry[9] which makes a

[1] https://www.ltg.ed.ac.uk/software/geoparser/
[2] https://spacy.io/
[3] https://github.com/alan-turing-institute/defoe
[4] https://livingwithmachines.ac.uk/
[5] https://viraltexts.org
[6] https://oceanicexchanges.org
[7] https://www.scottish-places.info/
[8] https://data.nls.uk/data/digitised-collections/gazetteers-of-scotland/
[9] https://data.nls.uk

series of its digitised collections publicly available.

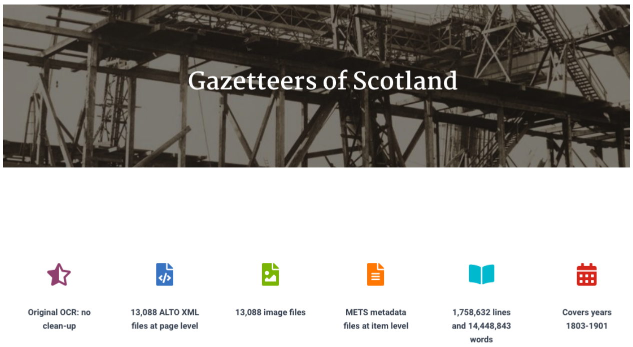

Gazetteers of Scotland
Original OCR: no clean-up
13,088 ALTO XML files at page level
13,088 image files
METS metadata files at item level
1,758,632 lines and 14,448,843 words
Covers years 1803-1901

Figure 1: The Gazetteers of Scotland data on the NLS Data Foundry.

The Gazetteers of Scotland are comprised of over 13,000 page images, their OCRed text in ALTO-XML format and corresponding METS-XML format for describing the metadata for each item in the collection (see Figure 1). In total, the OCRed text amounts to almost 14.5 million words and collectively these gazetteers provide a comprehensive geographical encyclopaedia of Scotland in the nineteenth century. While this is a valuable resource, it is too time-consuming to geoparse this data manually due to its size.

Year	Title	Volumes
1803	*Gazetteer of Scotland*	1
1806	*Gazetteer of Scotland: containing a particular and concise description of the counties, parishes, islands, cities with maps*	1
1825	*Gazetteer of Scotland: arranged under the various descriptions of counties, parishes, islands*	1
1828	*Descriptive account of the principal towns in Scotland to accompany Wood's town atlas*	1
1838	*Gazetteer of Scotland with plates and maps*	2
1842	*Topographical, statistical, and historical gazetteer of Scotland*	2
1846	*Topographical dictionary of Scotland*	2
1848	*Topographical, statistical, and historical gazetteer of Scotland*	1
1868	*Imperial gazetteer of Scotland; or Dictionary of Scottish topography, compiled from the most recent authorities, and forming a complete body of Scottish geography, physical, statistical, and historical*	2
1882	*Gazetteer of Scotland*	1
1883	*Ordnance gazetteer of Scotland*	6
1901	*Ordnance gazetteer of Scotland*	1

Table 1: Gazetteers of Scotland, 1803-1901. The first column shows the publication year, the second the title and the third the number of volumes per gazetteer.

4. NLP Tools

4.1. The Edinburgh Geoparser

The Edinburgh Geoparser is a language processing tool designed to detect place name references in English text and ground them against an authoritative gazetteer so that they can be plotted on a map. The geoparser is implemented as a pipeline with two main steps (see Figure 2). The first step is geotagging, in which place name entities are identified. The second step is georesolution, which grounds place name entities against locations contained in a gazetteer. Typically, there are multiple candidates for a given place name entity, and the georesolver ranks candidates in order using various contextual clues. The georesolver allows the user to control which gazetteer to use, the main ones being GeoNames[10] or open Ordnance Survey resources, both of which we access using a service hosted by University of Edinburgh Information Services. The best choice of gazetteer will depend on the document that is being processed: if its content is global then Geonames is usually the most appropriate gazetteer but if the content is limited to Great Britain, Ordnance Survey gazetteers may help to limit the potential for ambiguity. One of the main heuristics in the georesolution step is to prefer the candidate with the largest population, but only GeoNames reliably provides this information; for this reason we have used GeoNames in this project. However, there is a way to reflect the fact that the content of the Gazetteers of Scotland is by its nature concerned primarily with Scotland by biasing disambiguation in favour of the correct Scottish places (e.g. prefer Perth, Scotland to Perth, Australia). We do this by supplying the bounding box which covers Scotland to the georesolver, which then tends to prefer candidates within the bounding box even if they have smaller populations. However, for the experiments shown in Section 5 we have not yet supplied the bounding box, but in the future we plan to do it so, so will be able compare results with and without bounding box. It is by monitoring these type of pipeline choices that we will be able to ascertain both accuracy and efficiency of our algorithmic georeferencing approaches.

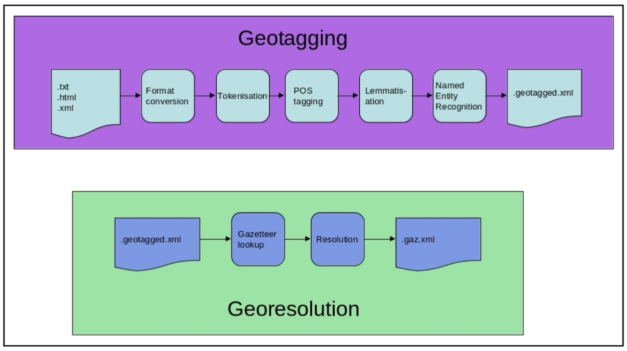


Figure 2: The Edinburgh Geoparser pipeline.

4.2. spaCy

spaCy is an open-source library for advanced Natural Language Processing in Python. It is designed specifically for production use and helps build applications that process

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

large volumes of text. Some of the features provided by spaCy are- Tokenization, Parts-of-Speech (PoS) Tagging, Text Classification and Named Entity Recognition (NER). While some of spaCy's features work independently, others require statistical models to be loaded, which enable spaCy to predict linguistic annotations. spaCy comes with two types pretrained statistical models and word vectors:

- Core models: General-purpose pretrained models to predict named entities, part-of-speech tags and syntactic dependencies.

- Starter models: Transfer learning starter packs with pretrained weights to be used as base model when training users' model. These models do not include components for specific tasks like NER or text classification.

Since the Edinburgh Geoparser gives us the flexibility to switch components, we are currently exploring the feasibility of using spaCy as one of the techniques for tokenisation and named entity recognition (NER). We have started focusing on the core models available for English [11]:

- `en_core_web_sm`: English multi-task CNN trained on OntoNotes. Assigns context-specific token vectors, POS tags, dependency parse and named entities. Small size model (11MB).

- `en_core_web_md`: English multi-task CNN trained on OntoNotes, with GloVe vectors trained on Common Crawl. Assigns word vectors, context-specific token vectors, POS tags, dependency parse and named entities. Medium size model (91MB).

- `en_core_web_lg`: English multi-task CNN trained on OntoNotes, with GloVe vectors trained on Common Crawl. Assigns word vectors, context-specific token vectors, POS tags, dependency parse and named entities. Large size model (789 MB).

To decide which spaCy model to use in our experiments, we performed an initial evaluation of the smaller and larger core models using the *Descriptive account of the principal towns in Scotland, 1828* gazetteer [12]. In this evaluation, we focused on quantifying the number of location entities identified by each model and visualising the differences between them. The `en_core_web_sm` identified 1124 locations, while `en_core_web_lg` identified 1455. Therefore, we have selected `en_core_web_lg`, since it gives us a more accurate overall results.

4.3. defoe

defoe is a scalable and portable digital toolbox for storing, processing, querying and analysing digital historical English textual data. It allows for extracting knowledge from historical text by running analyses in parallel via the Apache Spark big data framework and storing the preprocessed data (for further queries) in several storage solutions, such as an HDFS file system, an ElasticSearch distributed engine or a PostgreSQL database (see Figure 3). defoe is able to extract, transform and load (ETL) collections that comprise several XML schemas and physical representations. It offers a rich set of text mining queries to search across large-scale datasets and returns results for further analysis and interpretation. It also includes preprocessing techniques to mitigate against optical character recognition (OCR) errors and other issues (such as long-S and line-break hyphenation) and to standardise the text.

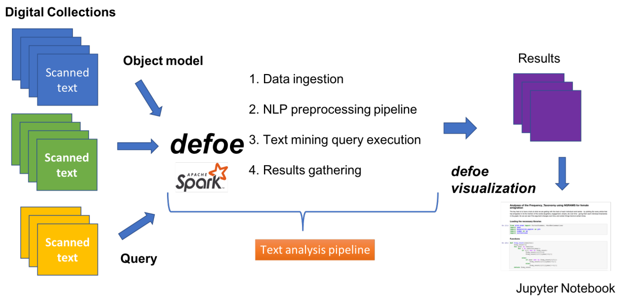


Figure 3: The defoe architecture.

defoe enables us to configure any query/queries to be submitted for an entire corpus or dataset processed, including the tokeniser and entity recogniser to use, which currently are those originally distributed within the Edinburgh Geoparser, and spaCy `en_core_web_lg` core model.

4.4. Combination of Methods

Since defoe already supports the XML schemas of the Gazetteers of Scotland, we have used it to create a new query that geoparses this collection automatically and in parallel using different geotagger options (Original geotagger from the Edinburgh Geoparser vs spaCy Name Entity) and combining them with the georesolution step of the Edinburgh Geoparser. The combined system performs the following tasks:

- Ingests the pages of all books belonging to the Gazetteers of Scotland data,
- Cleans the text to fix OCR errors caused by long-s characters and broken word tokens as a result of end-of-line hyphenation. Both steps are conducted using methods proposed and tested in (Alex et al., 2012),
- Identifies entities by employing the tokenisation and NER technique specified in the configuration file of the query,
- Applies georesolution to place name entities, and
- Groups the results by year and technique and provides them in combination with metadata associated with each book.

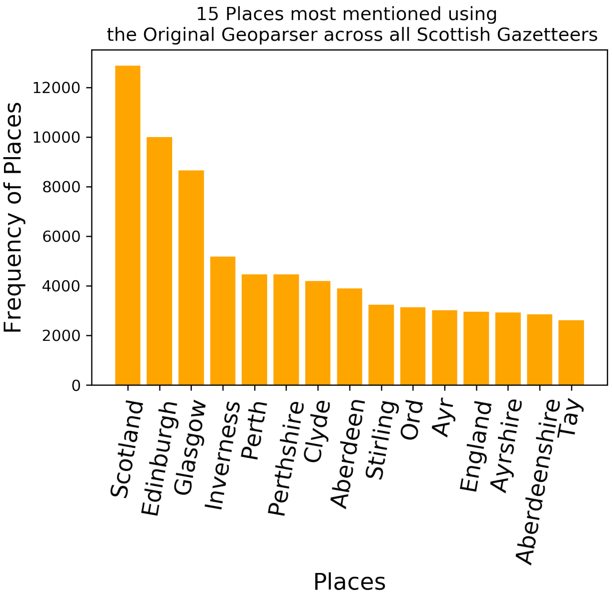
The first two steps of this query can be omitted if we apply the desired geoparser process to data that has been previously read, cleaned and stored using ElasticSearch. The parallelisation of the processing allows much faster turnarounds for obtaining and testing results. This is particularly useful during the method development process.

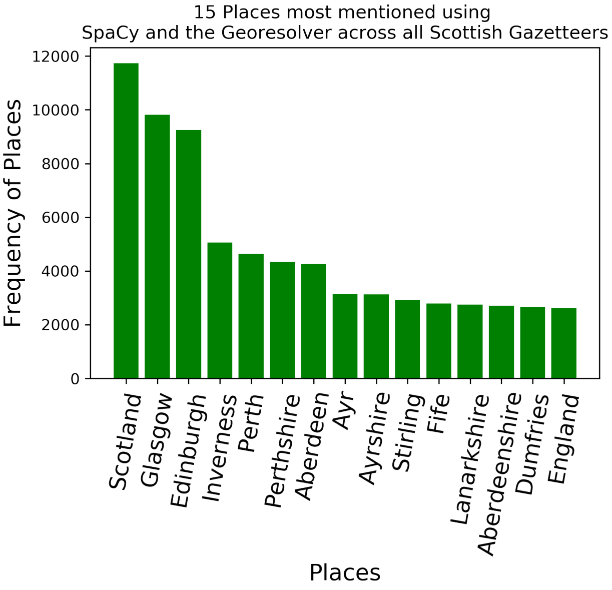
[11] `https://spacy.io/models/en`

[12] `https://github.com/alan-turing-institute/defoe_visualization/blob/master/Scottish_Gazetteer/Comparing_spacy_lang_models.ipynb`

5. Preliminary Results

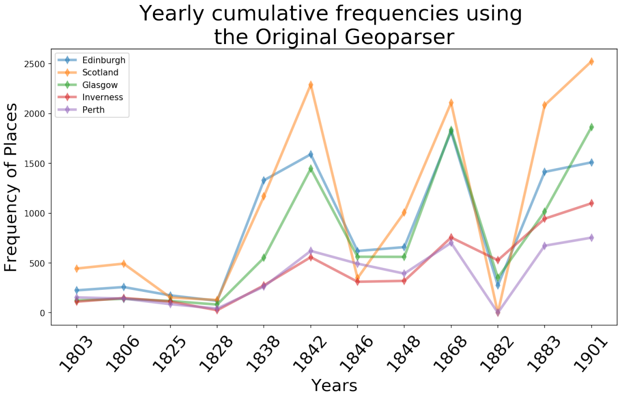
We compare different settings in defoe for the named entity recognition step, either the one from the Edinburgh Geoparser or spaCy and in both cases use the Edinburgh Geoparser's resolution step to disambiguate the place names. The georesolved output for running defoe's geoparser query using the original geotagger technique[13] or spaCy[14] is available for download. To visualise these results, we have created a collection of Jupyter Notebooks[15] where we load them into Pandas Dataframes and compare the locations that we obtain with each technique. Figures 4 and 5 show the most frequent georesolved place names across the entire gazetteers collection.

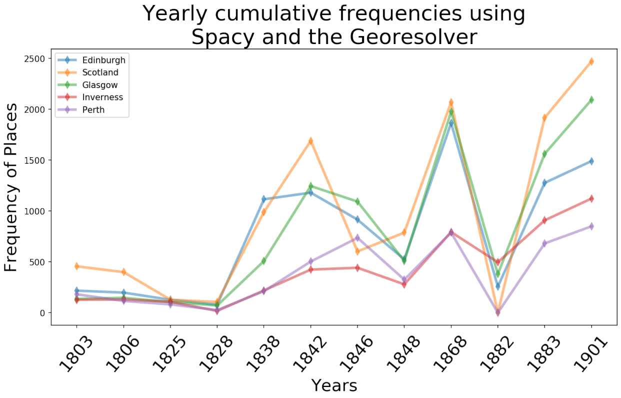




Figures 4 and 5: Most frequent georesolved locations using the Edinburgh Geoparser (above) or spaCy (below) NER.

Notice that the five most frequent locations mentioned among both techniques are *Edinburgh, Scotland, Glasgow, Inverness* and *Perth*.





Figures 6 and 7: Cumulative frequency of the five most mentioned locations using the Edinburgh Geoparser (above) or spaCy (below) NER over the years across the full Scottish Gazetteers collection.

Figures 6 and 7 show the yearly cumulative frequencies of these five places to analyse the evolution of how often they are mentioned with each technique. For reference, Figure 8 shows the normalized frequency of words for each year, obtained using a different defoe query.

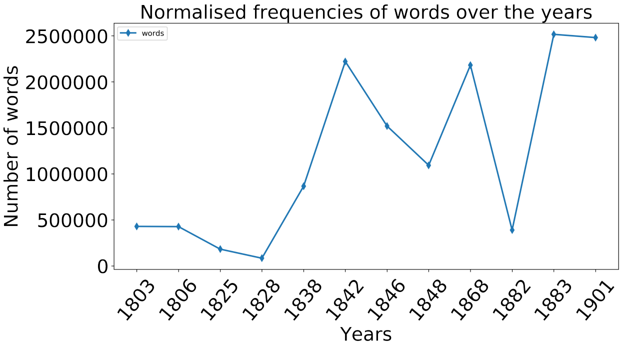


Figure 8: Normalized frequencies of words across the full Scottish Gazetteers collection.

Figures 9 and 10 show a more detailed study of the variation of locations' frequencies over the years. Both display the frequencies of the 15 most mentioned and georesolved places per year and technique.

[13] https://drive.google.com/open?id=1T26YHz5pFAEeJal0KHe77TGoKhkxYv_S

[14] https://drive.google.com/open?id=1f7iD-ng6jVurG9BtqrV_ZYYJGq9o93co

[15] https://github.com/alan-turing-institute/defoe_visualization/blob/master/Scottish_Gazetteer

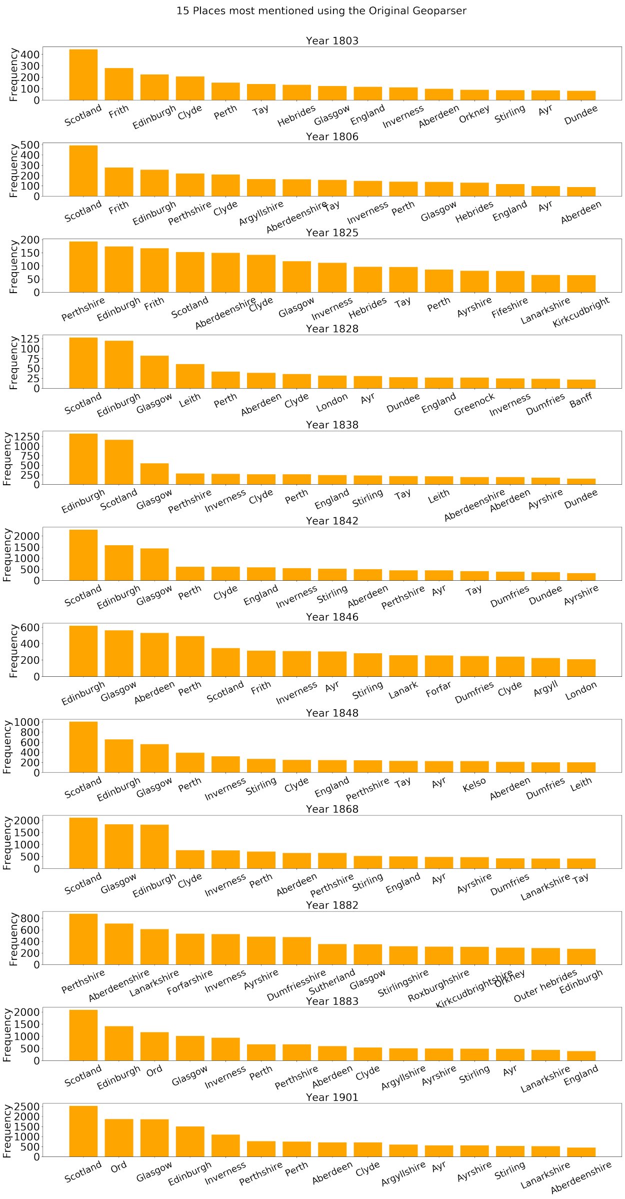


Figure 9: Most frequent georesolved locations using the Edinburgh Geoparser NER gathering the results by publication years.

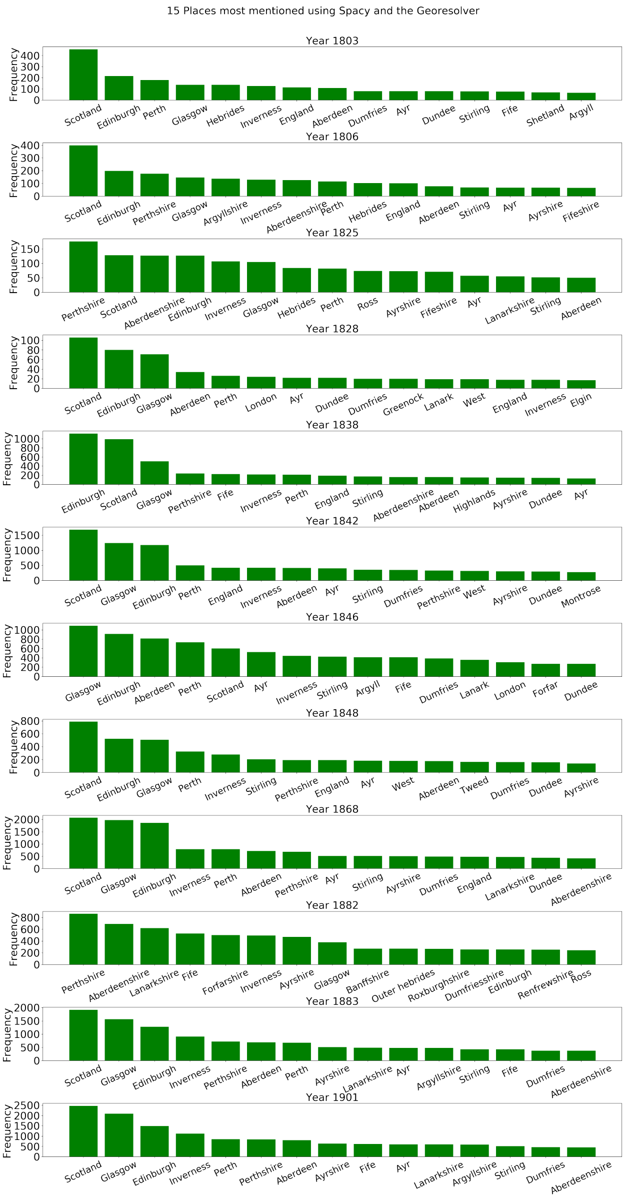


Figure 10: Most frequent georesolved locations using the spaCy NER gathering the results by publication years.

All these graphs show that the Edinburgh Geoparser is able to recognise several locations more frequently for equivalent place names.

Finally, we also explore which are the most frequent places names that have been identified but not resolved using the Edinburgh Geoparser (see Figure 11), the top four place names being Scottish shires.

We have yet to conduct a formal evaluation of the geotagging and georesolution steps on this data to see how both methods compare quantitatively and to find out where further work is needed to improve performance overall. Over the summer 2020 we plan to annotate a random subset of excerpts from the gazetteers to create a gold standard and compare it against system output. Such formal evaluation is essential to provide transparency about the accuracy of geoparsing and text mining methods developed to analyse mass digitised content automatically. We will fully document our code, and make our training set available for others, to encourage open science approaches to data analysis.

We expect that geoparsing performance on this type of data is likely to be affected by the quality of the OCR, the use of historical place name variants or spelling variation and the use of Gaelic place names. The collection contains volumes published over the course of the 19th century during which type and quality of printing and use of language changed. This is undoubtedly going to be affected by OCR quality and consistency of spellings across the volumes. Previous work showed that OCRed text has a negative cascading effect on natural language processing tasks (Alex et al., 2012; Kolak and Resnik, 2005; Lopresti, 2005; Lopresti, 2008b; Alex et al., 2019) or information retrieval (Gotscharek et al., 2011; Hauser et al., 2007; Lopresti, 2008a; Reynaert, 2008) and those using NLP approaches to historical texts, in particular, have to take care regarding how the error rate of OCR can affect analysis (Ryan Cordell, 2017).This means

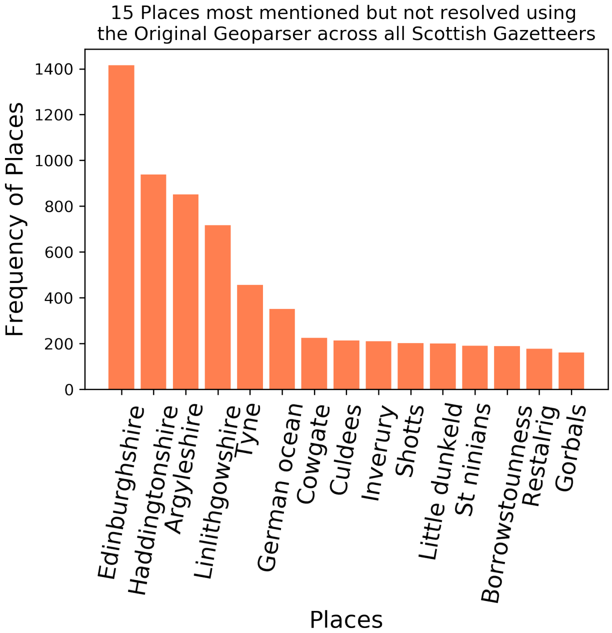


Figure 11: Most frequently identified locations which cannot be resolved with the Edinburgh Geoparser.

that during the evaluation step, we will need to carefully sample from different volumes across the collection to get a balanced view of performance overall. This could involve estimating the quality of the OCR (Alex and Burns, 2014), for example by pages, and selecting samples with different levels of quality.

The Gazetteers of Scotland are descriptive gazetteers with locations often listed alphabetically as opposed to pure alphabetical lists without descriptions. While one would expect that the latter would be easy to tag correctly throughout, for the former type, structure of descriptions can nevertheless be exploited to identify the main names of each entry, especially if the font face or type changes and information is preserved in the OCR. However, this is not the case for place names appearing inside a description as they can often be ambiguous and can overlap with people's names, for example.

Encouraging other scholars to reuse our data will require training in and understanding of these nuances, and it is likely that we will need to run workshops or bespoke support to understand how best to engage with the research communities that this could support (Terras et al., 2017).

6. Summary

We have described our investigations into the flexible deployment of NLP components for automatic and parallel processing of historical text, focusing on the geoparsing of the National Library of Scotland's Gazetteers of Scotland Collection. Our work so far has already made these texts easily searchable both by keyword and by place name grounded to latitude/longitude, but there are several extensions to this work that we wish to take forward. The first is to run the same experiments supplying a bounding box for Scotland to compare results with and without a bounding box. Then, we plan to create a representative annotated test set not only to formally evaluate the performance of various configurations of components but also to determine where improvements to the processing can most fruitfully be made. When complete, this test set can be shared with other research groups who want to evaluate their own geoparsing tools on it. A third strand of future work will be to develop map-based and other data visualisations and to consider how best to provide interfaces to a variety of potential users working within the data carpentries framework, and with the digital humanities community, to establish best practice in data sharing, training, and support structures. Our ultimate goal is to create a digital Scotland-focused historical gazetteer which can be used to drive accurate geotagging and georesolution of other Scottish historical text collections, which we aim to publish openly, for others to use. This would mean that researchers working with Scottish historical text would have the means to interrogate their data by place name and be provided with automatic links to the relevant entries in the Scottish Gazetteers. We are also developing a Text and Data Mining Library Carpentries course to teach researchers how to run different types of text analysis and how to visualise the output.[16]

7. Acknowledgements

This work was funded by the Data Driven Innovation Programme as part of the Edinburgh and South East Scotland City Region Deal. The authors wish to thank Mike Bennett at the University of Edinburgh Library, who runs and supports the Unlock service, and Sarah Ames at the National Library Scotland, who has provided us with valuable insights on the Scottish Gazetteers.

References

Alex, B. and Burns, J. (2014). Estimating and rating the quality of optically character recognised text. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 97–102.

Alex, B., Grover, C., Klein, E., and Tobin, R. (2012). Digitised Historical Text: Does it have to be mediOCRe? In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 401–409.

Alex, B., Byrne, K., Grover, C., and Tobin, R. (2015). Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal for Humanities and Arts Computing*, 9(1):15–35.

Alex, B., Grover, C., Tobin, R., and Oberlander, J. (2019). Geoparsing historical and contemporary literary text set in the City of Edinburgh. *Language Resources and Evaluation*, 53(4):651–675.

Borin, L., Dannélls, D., and Olsson, L.-J. (2014). Geographic visualization of place names in Swedish literary texts. *Literary and Linguistic Computing*, 29(3):400–404, 05.

Clifford, J., Alex, B., Coates, C., Klein, E., and Watson, A. (2016). Geoparsing history: Locating commodities in ten million pages of nineteenth-century sources. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 49(3):115–131.

[16] http://librarycarpentry.org/lc-tdm/

Filgueira, R., Jackson, M., Terras, M., Beavan, D., Roubickov, A., Hobson, T., Ardanuy, M., Colavizza, G., Krause, A., Hetherington, J., Hauswedell, T., Nyhan, J., and Ahnert, R. (2019). defoe: A spark-based toolbox for analysing digital historical textual data. In *2019 IEEE 15th International Conference on e-Science (e-Science)*, 09.

Gotscharek, A., Reffle, U., Ringlstetter, C., Schulz, K. U., and Neumann, A. (2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *IJDAR*, 14(2):159–171.

Grover, C. and Tobin, R. (2014). A gazetteer and georeferencing for historical english documents. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 119–127.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., and Ball, J. (2010). Use of the Edinburgh Geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A*, 368(1925):3875–3889.

Hauser, A., Heller, M., Leiss, E., Schulz, K. U., and Wanzeck, C. (2007). Information access to historical documents from the Early New High German period. In L. Burnard, et al., editors, *Digital Historical Corpora-Architecture, Annotation, and Retrieval*, Dagstuhl, Germany.

Kolak, O. and Resnik, P. (2005). OCR post-processing for low density languages. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 867–874.

Lopresti, D. (2005). Performance evaluation for text processing of noisy inputs. In *Proceedings of the Symposium on Applied Computing*, pages 759–763.

Lopresti, D. (2008a). Measuring the impact of character recognition errors on downstream text analysis. In B. A. Yanikoglu et al., editors, *Document Recognition and Retrieval*, volume 6815. SPIE.

Lopresti, D. (2008b). Optical character recognition errors and their effects on natural language processing. In *Proceedings of the second workshop on Analytics for Noisy Unstructured Text Data*, pages 9–16.

Moncla, L., Gaio, M., Joliveau, T., and Le Lay, Y.-F. (2017). Automated geoparsing of Paris street names in 19th century novels. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*.

Porter, C., Atkinson, P., and Gregory, I. N. (2018). Space and time in 100 million words: Health and disease in a nineteenth-century newspaper. *International Journal of Humanities and Arts Computing*, 12(2):196–216.

Rayson, P., Reinhold, A., Butler, J., Donaldson, C., Gregory, I., and Taylor, J. (2017). A deeply annotated testbed for geographical text analysis: the corpus of Lake District writing. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, pages 9–15. ACM.

Reynaert, M. (2008). Non-interactive OCR post-correction for giga-scale digitization projects. In *Proceedings of the 9th international conference on Computational Linguistics and Intelligent Text Processing*, pages 617–630.

Rupp, C., Rayson, P., Baron, A., Donaldson, C., Gregory, I., Hardie, A., and Murrieta-Flores, P. (2013). Customising geoparsing and georeferencing for historical texts. In *2013 IEEE International Conference on Big Data*, pages 59–62. IEEE.

Ryan Cordell. (2017). Q i-jtb the Raven': Taking Dirty OCR Seriously. `http://ryancordell.org/research/qijtb-the-raven/`. Book History 20 (2017), 188-225.

Terras, M., Baker, J., Hetherington, J., Beavan, D., Zaltz Austwick, M., Welsh, A., O'Neill, H., Finley, W., Duke-Williams, O., and Farquhar, A. (2017). Enabling complex analysis of large-scale digital collections: humanities research, high-performance computing, and transforming access to British Library digital collections. *Digital Scholarship in the Humanities*, 33(2):456–466, 05.