Dealing with Heterogeneous Big Data When Geoparsing Historical Corpora

C.J. Rupp, Paul Rayson, Ian Gregory, Andrew Hardie, Amelia Joulain, Daniel Hartmann
Lancaster University
Lancaster, UK
p.rayson@lancaster.ac.uk

Abstract—It has long been known that ‘variety’ is one of the key challenges and opportunities of big data. This is especially true when we consider the variety of content in historical corpora resulting from large-scale digitisation activities. Collections such as Early English Books Online (EEBO) and the British Library 19th Century Newspapers are extremely large and heterogeneous data sources containing a variety of content in terms of time, location, topic, style and quality. The range of geographical locations referenced in these corpora poses a difficult challenge for state-of-the-art geoparsing tools. In the context of our work on Spatial Humanities analyses, we present our solution for dealing with the variety and scale of these corpora.

Keywords—Text mining; Toponym Resolution; Historical Corpora; NLP Pipelines and Workflows

I. INTRODUCTION

Geoparsing is the process of automatically identifying place-name instances within a text and subsequently allocating each place-name to a coordinate [1]. Doing this allows the researcher to extract place-name instances from a text so that it can subsequently be mapped and analysed in geographical information systems (GIS). This, in turn, opens up new potential for analysing texts in ways that stress their geographies, and new sources for use in GIS which has traditionally been restricted to the quantitative and the cartographic [2], [3]. Geoparsing is, however, an error-prone process. Two potential sources of error are particularly problematic: establishing whether a proper noun is, or is not, a place-name, and then disambiguating place-names that can refer to different places. An example of the first type is that “Lancaster” can be a town, a plane (“Lancaster bomber”), a person (“Stuart Lancaster”), a nobleman or a pub (both “Duke of Lancaster”). An example of the second type is that as well as the city of Lancaster in the north-west of England, there are at least five towns called Lancaster in the USA, one in South Africa, and one in Australia, not to mention that Lancaster is also an old-fashioned name for the county of Lancashire. Geoparsers such as the Edinburgh Geoparser [1] do incorporate rules to attempt to resolve both of these types of problem, but these issues are particularly complex and in some cases the disambiguation requires world-knowledge and specific domain expertise. In one study, Tobin et al. suggested that the Edinburgh Geoparser successfully geoparsed 75% of place-names within the Histpop collection of 13 million words published by the census and the General Register Office from 1801-1937 [4]. However, different corpora are likely to give significantly different error rates depending on factors as diverse as OCR quality and the subject material or topic.

In the Spatial Humanities project, we have applied and adapted geoparsing techniques for historical texts and reported previously on small-scale studies using a corpus of Lake District travel literature [5]. When such techniques are scaled up to deal with big data in extremely large heterogeneous corpora, at least two issues arise. First, the timescale of the underlying data processing is vastly increased, although this can be dealt with by employing parallel techniques such as MapReduce. Second, the variation of format, topic and geographical focus across multiple texts dictates that either geoparsing techniques need to self-adapt or separate tailored geoparsers are applied to sub-corpora divided by time and space criteria. In this paper, we propose an alternative approach to the one-size-fits-all solution by turning the NLP pipeline on its head and geoparsing result text selected by theme rather than geoparsing the whole dataset first.

II. DATA AND PROPOSED METHOD

The two large datasets that we are analysing in our project are Early English Books Online (EEBO) and the British Library’s Nineteenth Century Newspapers collection. We are using the transcribed version of the EEBO corpus from the Text Creation Partnership (TCP) consisting of around 1.2 billion words. EEBO itself includes over 125,000 works printed between 1473 and 1700, and the TCP version incorporates just over 44,000 of those works. The British Library dataset incorporates an estimated fifty billion words published in British newspapers of the 19th century. This newspaper corpus has been automatically recovered, using OCR techniques, adding a further layer of uncertainty to the analysis of place-names. Simply geoparsing these corpora in their entirety in one pass with one system is not the best solution due to the variability within the data at the book and article level.

When geoparsing this volume of material, it is extremely difficult to establish what errors there are and subsequently correct them. Instead we propose an approach based on context-sensitive geoparsing. Our research centre round using the geoparser to answer the question: which places are associated with the theme I am interested in, where the theme might be cholera in the nineteenth century or religion in the Early Modern period.

Our proposed procedure is therefore:

1) Decide on a search term or terms.

1http://www.lancaster.ac.uk/spatialhum/
2as available in September 2013
2) Extract the first search term from the corpus with fifty
words of co-text to the left and right
3) Geoparse this concordance extract
4) Explore the results in the text and by mapping to
identify errors
5) Correct these errors to an update file
6) Re-run the geoparser (stage 3) this time incorporating
the updates

The advantage of this is that the first search-term used is
likely to need significant correction however as further search
terms are added the time taken to correct them will decline
rapidly and the user quickly builds up an effective corrections
file that is sensitive to the particular characteristics of the
corpus under study.

III. HOW TO GEOPARSE CORPUS SEARCH RESULTS

Our task is to identify place-names in the immediate con-
text of a given search term, starting with a corpus search result
from a corpus that has not (yet) been geoparsed. The retrieval
software which we use (Corpus Workbench/CQPweb [6]) allows any amount of co-text to be extracted for each con-
cordance line; as noted above, for this analysis we work with 50
tokens before and 50 tokens after each concordance hit. This
is a somewhat wider span of co-text than is usually utilised
when studying collocation effects; however, our experience
has shown that relevant discoursal associations between the
concept searched for and mentions of places can operate
over rather longer spans of text than do lexico-grammatical
collocactions. Depending on the language being studied, a ±50
word span may or may not cover a full sentence. In Early
Modern English text, sentences are frequently rather long,
consisting of long sequences of coordinated or juxtaposed
clauses, and in such cases, the full sentence may not be
captured. This is illustrated by the example shown in Table
1, a line from a concordance for nun in the EEBO corpus,
where the ±50 word co-text captures the beginning of the
containing sentence, but not the end. However, in these cases,
topic shifts that in contemporary English would occur between
sentences tend to occur within the sentence, and therefore the
±50 span still represents a reasonable cut-off for discourse
relevance. This approach to sentence breaks does however
mean that care needs to be taken when submitting to the
gеoparser the file that contains the concordance table. The file
as a whole does not contain a fully coherent text; each separate
concordance line must be treated independently from those
before and after, and, most particularly, the ending sentence
fragment of one line must not be treated as part of the same
text as the starting sentence fragment of the next line. This is,
however, easily accomplished, as the geoparser does not utilise
any operations which operate across paragraph boundaries.
Therefore, marking each line of the concordance table as a
separate paragraph forces the geoparser to treat each stretch
of co-text as an isolated unit. In theory, breaks in the flow of
the syntax resulting from truncated sentences at the edges of
the concordance line might degrade geoparser accuracy (for
example, by disrupting its ability to accurately identify proper
nouns by part-of-speech tagging). In practice we have no
evidence of any such detrimental effect arising from sentence
truncation. As the context is defined as a window of a given
number of words, the initial and final sequences may be
incomplete sentences. However, this constructed text should
be adequate to for identifying place-names and resolving them
to a specific location, by assigning coordinates, provided that
each segment is processed independently.

Table I shows an instance of the term nun with a place-
name in the immediate context. The example is encapsulated
in paragraph tags, but the leading and trailing sequences of the
context fields are incomplete sentences.

In practice, the geoparser performs a sequence of analysis
steps, including tokenization, sentence splitting, POS tagging
and the application of Named Entity Retrieval (NER) rules,
to identify place-names: geotagging. These are all restricted
to the local context and can be contained by our imposition
of paragraph boundaries. The only mechanisms that range
more widely are in the georesolution rules which use an
extensive gazetteer to identify the location the place-name
is intended to refer to. These rules do include a mechanism
that ranges over the whole text, to provide a weighting for
locations that are closer to other resolved place-names in the
text, so that clusters of locations in the same area can be
recognised. The georesolution process operates on the basis
of scores assigned by a set of factors, each of which can be
independently weighted, so that the highest scoring candidate
can be selected from the locations in the gazetteer that match
the place-name. The optionality of each resolution mechanism
is already designed in, so it is easy to eliminate the clusteriness
factor, and, via a locality parameter, to boost the scores of
locations in, or closer to, the British Isles.

Another process in the geoparsing of the concordances
is independent of the Edinburgh geoparser. We are dealing
with historical texts which are prone to spelling variation. The
spelling conventions may differ from the modern standard, and
across instances between and perhaps even within a text. This
is a known problem, as English spelling was only gradually
standardised. We make use of an existing spelling normalisa-
tion package: VARD [7]. This can provide a substitution of a
variant spelling with a normalised modern form, which is what
we need for the geoparser, as all its lexicons and gazetteers
are based on modern information.

We encode the geoparsing results as annotations on the
original left and right co-text in the concordance tables. This
implies a further processing step where the geoparsing
results on the normalised text are transferred to the original
concordance lines. This is implemented as an alignment of the
original concordance and the geoparsed concordance. The re-
solved place-name annotations — and, if necessary, a record of
the normalisations applied – are thus available in concordance
tables resulting from a keyword search on a full corpus. Such a
concordance is therefore equivalent to a concordance extracted
from an underlying geoparsed corpus, but without the need to
actually process the whole corpus in advance.
IV. APPLICATIONS OF THE STRATEGY

A straightforward application of this method could follow the example cited above. First, retrieving all instances of the word nun in the EEBO corpus via a CQPweb search, the search result yields a set of text segments, of the keyword and surrounding context. These can be sequenced as a constructed text and submitted to the geoparser pipeline, with a locality parameter set to prefer locations in the British Isles and no positive weighting for locations clustering in the text. This will result a set of georesolved instances or place-names that collocate with the term nun. Figure 1 is a map, generated using ArcGIS, to represent the distribution of those locations.

However, on the initial geoparsing, some locations may remain undetected or unresolved, or on further examination prove to be incorrectly resolved. There is some scope to correct these shortcomings by compiling a specialised gazetteer to be used in conjunction with the generic gazetteer in the georesolutions process. Such a gazetteer list can be specialised to specific period, or a specific region, in conjunction with locality parameters, or potentially to a specific topic. The main scope for intervention is in the georesolution process and the locality parameters, or potentially to a specific topic. The main

with it however is that it is unclear whether the pattern reflects a clear geography associated with the search term or whether it simply reflects the underlying geography of the corpus. Thus does our theme really cluster in London or is it simply that the corpus as a whole focuses on London? We must therefore be able to reject the null hypothesis that the pattern found is simply an artefact of the underlying distribution of place names in the corpus. This is made problematic by the fact that we do not know the underlying geography of all place names within the corpus because we do not want to geoparse the entire corpus. To avoid this problem we use Kulldorf’s [8] spatial scan statistic. This was developed for the identification of clusters of diseases against variable background populations, however the principle remains the same – we want to test whether place names that collocate with the search-term we are interested in still form the same clusters when the underlying distribution of place names is taken into account. This can be done using the Bernoulli model within Kulldorf’s statistic in which the search term under study is compared to a control search term to see where the search term clusters in comparison. In the example shown the distribution of place name instances that collocate with nun is compared with the distribution of mother to identify which places are associated with nuns more or less than would be expected given the distribution of another word associated with women. Figure 2 shows the distribution of references to nun against the background of references to mother. It suggests that there do seem to be clusters on nun instances in the West Country, West Midlands, West Yorkshire, Kent and on the south-east coast of Scotland. Other apparent clusters from figure 1, notably in London appear to be likely to be artefacts of the underlying geography of the corpus as nun[s] are mentioned less than would be expected here when compared to mother. Clearly doing this robustly requires the use of a wider range of background terms, for example, for nun we might want to use a range of terms for women, a range for religion, and a range for urban and test all of these to see whether clusters of instances of nun appear to be statistically robust. Further investigation of the underlying corpus can then be used to explain these clusters, for example, the one in south-
east Scotland is caused by a famous incident where the nuns of Coldingham cut off their own noses and lips to preserve their chastity from Viking invaders. range of different themes (search terms associated with, for example, women, religion and urban areas) to test the extent to which clusters are robust to variable controls.

Figure 3 exemplifies an application of this approach within our analysis of the British Library newspaper corpus. In earlier work [3], we scrutinised a corpus drawn from the Registrar Generals reports for terms and concepts associated with the water-borne diseases that were major causes of death in the nineteenth century, such as cholera and dysentery. While we were able to geoparse that (substantially smaller) corpus in full, for the British Library material, the concordance-oriented technique described here is essential due to its massive scope. For example, the corpus data for The Era, one weekly newspaper which was published from 1838 onwards, constitutes 377 million tokens. We are currently working on an analysis of the terms related to disease in The Era and, subsequently, other newspapers from the British Library data (dealing with each publication separately in the first instance allows us to circumvent the problem of variations in the quality of OCR across publications). Figure 3 exemplifies the output of this process for The Era, using the search term cholera. In contrast to the nun/woman analysis, in this case our point of comparison would be maps generated using the same methodology, but from the Registrar General reports or other news publications. At this stage, we have already identified differences in the distribution of the term cholera relative to the annual government reports in the Registrar General corpus. However, it is necessary to tease apart differences that are due to genuine distinctions in the perception of the geographical spread of cholera (or whatever other disease) between different source corpora, and differences that arise due to other factors. For example – perhaps surprisingly to a modern reader, a proportion of the instances of cholera in The Era occur in advertisements, rather than in news accounts of cases or outbreaks of the disease.

V. Conclusion

We have presented a new solution for extracting geographical information from extremely large heterogeneous corpora. By shifting the geoparsing step out of its standard position in the NLP pipeline, and instead using the geoparser to extract place-names from the co-text of a query result, we enhance the procedure of geographical analysis of text in several ways. As well as the obvious gain in processing speed, the smaller geoparser output allows enhanced manual correction, for drastically improved geoparsing results, tailored directly to the specific research questions and to the specific corpus. The impact of two key problems in geoparsing, especially, historical texts - spelling variation/OCR errors, and place-name ambiguity and heterogeneity - is thereby drastically reduced. Moreover it becomes straightforward to investigate whether geographical patterns identified in the data are robust, or merely an artefact of the underlying spread of place-names within the corpus. Critically, our new approach puts the user in control of the geoparsers accuracy in a way which would not be possible with the output of processing hundreds of millions or billions of words. Future work will seek to apply more adaptive approaches to improve the outcomes of full-corpus geoparsing, enabling a two-pronged approach to the identification and analysis of places in very large textual datasets.

ACKNOWLEDGMENT

The project is funded by the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013) / ERC grant Spatial Humanities: Texts, GIS, places (agreement number 283850) and is affiliated with the CASS ESRC-funded research centre (grant reference: ES/K002155/1). We would like to thank Claire Grover, the geoparser developers, and the Unlock Text team in Edinburgh for their support with the geoparser, and James Baker from the British Library for enabling data access.

REFERENCES