Evaluation of Georeferencing

Richard Tobin
School of Informatics
University of Edinburgh
R.Tobin@ed.ac.uk

Claire Grover
School of Informatics
University of Edinburgh
C.Grover@ed.ac.uk

Kate Byrne
School of Informatics
University of Edinburgh
K.Byrne@ed.ac.uk

James Reid
EDINA
University of Edinburgh
james.reid@ed.ac.uk

Jo Walsh
EDINA
University of Edinburgh
jo.walsh@ed.ac.uk

ABSTRACT

In this paper we describe a georeferencing system which first uses Information Extraction techniques to identify place names in textual documents and which then resolves the place names against a choice of gazetteers. We have used the system to georeference three digitised historical collections and have evaluated its performance against human annotated gold standard samples from the three collections. We have also evaluated its performance on the SpatialML corpus which is a geo-annotated corpus of newspaper text. The main focus of this paper is the evaluation of georesolution and we discuss evaluation methods and issues arising from the evaluation.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis; H.3.4 [**Systems and Software**]: Performance evaluation.

Keywords

Georeferencing, Toponym Resolution, Named Entity Recognition, Evaluation.

1. INTRODUCTION

Evaluation of georeferencing systems is important but not straightforward. There is a lack of standardised resources ([8]) and comparisons are made difficult because of differences in the gazetteers that are used and the text types that are processed. In this paper we focus on evaluation of our georeferencing system, using both in-house data and the SpatialML corpus [9]. Our system combines general purpose XML-based Information Extraction technology using LT-TTT2 [5] with georeferencing-specific sub-components developed in the context of projects dealing with digitised historical collections. The diagram in Figure 1 provides an overview of the components of the georeferencer. There are two main parts, the geotagger and the georesolver. The former processes an input text and identifies the strings within it which denote place names. The latter takes the pool of recognised place names as input, looks them up in one of a number of gazetteers and determines for each place name which of the possible referents is the correct one.

In the following section we provide a brief overview of our system while a more detailed description of the georesolver can be found in Section 5. In Section 3 we describe the data sets which we have used for evaluation and in Sections 4 and 6 we provide evaluation results for the geotagger and the georesolver, respectively. In Section 7 we look at end-to-end evaluation and attempt a comparison between our system and Yahoo! Placemaker[1].

GIR'10, 18-19th Feb. 2010, Zurich, Switzerland.

2. SYSTEM OVERVIEW

The workings of the system can be illustrated with a small example. The following is a short plain text input file:

```
Some of the time savings will be remarkable: Canterbury will be an hour from London, a saving of 40 minutes; the journey from Dover will be slashed by 47 minutes and those living around Ebbsfleet near Gravesend will be just 18 minutes from St Pancras.
```

The output of the geotagger is an XML file containing linguistic mark-up, including enamex elements which wrap recognised place names. Suppressing all other mark-up, the output looks like this:

```
Some of the time savings will be remarkable:
<enamex type='location' id='1'>Canterbury</enamex>
will be an hour from
<enamex type='location' id='2'>London</enamex>,
a saving of 40 minutes; the journey from
<enamex type='location' id='3'>Dover</enamex>
will be slashed by 47 minutes and those living around
<enamex type='location' id='4'>Ebbsfleet</enamex> near
<enamex type='location' id='5'>Gravesend</enamex>
will be just 18 minutes from
<enamex type='location' id='6'>St Pancras</enamex>.
```

The enamex elements are extracted and converted into gazetteer queries, the format of which is dependent on the gazetteer being used. The system is currently configured to use one of three gazetteers:

[1] `http://developer.yahoo.com/geo/placemaker/`

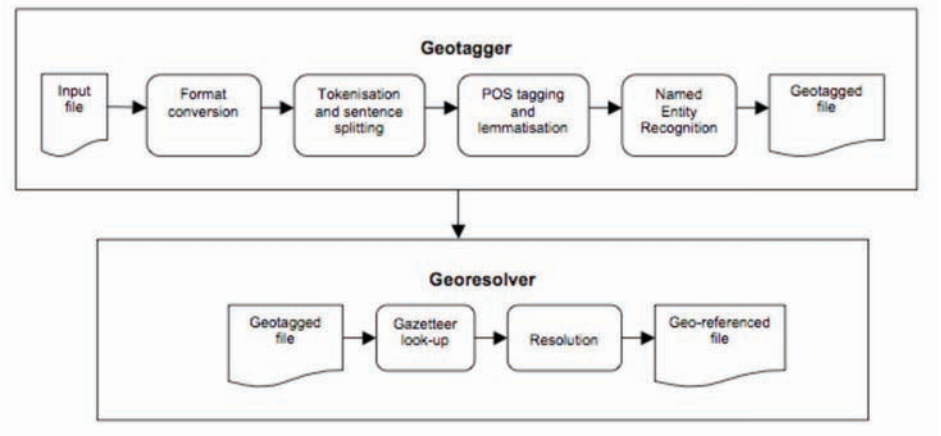

Figure 1: Overview of Georeferencing System

- geonames: the webservice for the GeoNames[2] gazetteer
- xwalk: the webservice for the GeoCrossWalk gazetteer. This has recently been replaced by the Unlock service.
- unlock: the Unlock[3] webservice for Ordnance Survey gazetteer information (currently only available to academic subscribers to the Ordnance Survey Collection).

The gazetteer server responds with a list of candidate entries which are converted to an appropriate XML format. For example, if the GeoNames gazetteer is used, the entries for "Gravesend" are converted to this:

```
<placename name='Gravesend' id='5'>
<place name='Gravesend' gazref='geonames:1000153' type='fac'
lat='-30.1666667' long='30.7333333' in-cc='ZA'/>
<place name='Gravesend' gazref='geonames:1000154' type='fac'
lat='-23.0833333' long='28.1833333' in-cc='ZA'/>
<place name='Gravesend' gazref='geonames:2164648' type='ppl'
lat='-29.5833333' long='150.3166667' in-cc='AU'/>
<place name='Gravesend' gazref='geonames:2648187' type='ppl'
lat='51.4333333' long='0.3666667' in-cc='GB'/>
<place name='Gravesend' gazref='geonames:5119167' type='ppl'
lat='40.5976048' long='-73.9651383' in-cc='US'/>
<place name='Gravesend' gazref='geonames:6690892' type='ppl'
lat='40.594663726005' long='-73.965368270874' in-cc='US'/>
</placename>
```

Here it can be seen that GeoNames has returned six possible candidates for "Gravesend", two in South Africa, one in Australia, one in Great Britain and two in the U.S. The two candidates in South Africa are both facilities (fac) while the others are populated places (ppl).

The results of gazetteer look-up for any one place will contain zero, one or more than one entry. In the case of zero, if the gazetteer has no information, the place will have to remain unresolved. In the case of one entry (e.g. for "Ebbsfleet"), the georesolver takes this to be the correct resolution. In the case of multiple entries, as with "Gravesend", the georesolver ranks the candidate entries in order of likelihood that they are correct in this context. To do this it uses a variety of information relating to the linguistic context, to the entries for all the other places in the document (i.e. the document context) and population and containing country information where available. For the current example the georesolver correctly gives the highest rank to the Great Britain candidate.

An optional gazmap component allows the results of the georesolver to be explored in a browser using Google Maps, as shown in Figure 2.

Our system is similar in architecture and functionality to Clough's geographic metadata extraction tool [2] of the SPIRIT system [11]. This tool is comprised of a geo-parser for recognising place names (the equivalent of our geotagger) and a geo-coder for grounding place names (the equivalent of our georesolver). In both cases, place name recognition is done using standard information extraction technology: the SPIRIT geo-parser uses GATE [3] while our geotagger uses LT-TTT2 [5]. Both use rule-based named entity recognition techniques for capturing information about linguistic context combined with gazetteer-based lexical look-up. There are some differences between our geotagger and Clough's geo-parser in terms of what is output: in the version of our geotagger described here we do not recognise address parts or postcodes while Clough's geoparser does. Both systems deal with the problem of person-place ambiguity ("Francis Chichester") but our system explicitly recognises and marks up person names as well as place names. Our system implicitly recognises organisation names and will not mark up a place name which is part of an organisation name ("University of Edinburgh") since this is not a direct reference to a place. Our system also attempts to compute and pass on to the georesolver information from the linguistic context which could be helpful to resolution. For example we attempt to provide some relationships between place names such as containment (e.g. "Ipswich, Suffolk"), coordination (e.g. "Chelmsford, Colchester and Ipswich") and proximity (e.g. "Manningtree near Ipswich"). We also provide possible alternate names for gazetteer look-up of some multi-word place names (e.g. for "Co. Down", "Down" is an alternate name). Space considerations preclude a more detailed discussion but further information about our geotagger can be found in [6] and [7].

Our georesolver and Clough's geo-coder use different gazetteers. Both have Ordnance Survey-based, Great Britain-specific resources, GeoCrossWalk/Unlock for us and the Ord-

[2] `http://www.geonames.org`
[3] `http://unlock.edina.ac.uk/`

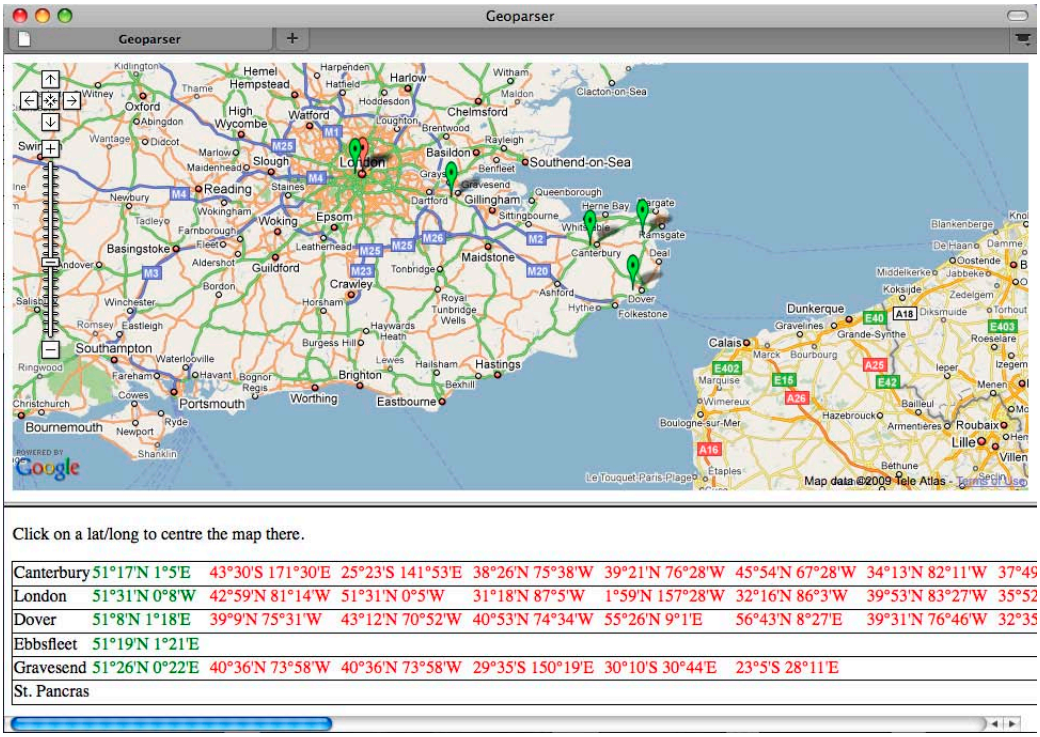

Figure 2: Google Maps View

nance Survey 1:50,000 Scale Gazetteer for the geo-coder[4]. Where we use GeoNames for world-scale information, the geo-coder uses SABE[5] and the Getty Thesaurus of Geographic Names[6]. Our system resolves place names in a document with respect to one or other gazetteer, depending on the user's selection, while the SPIRIT geo-coder combines all three of its gazetteers to resolve place names. Both systems use heuristics based on information such as population and feature type these are discussed in more detail in Section 5.

3. EVALUATION DATA

The development of our georeferencing system has most recently been supported by two JISC-funded projects, GeoDigRef[7] and Embedding GeoCrossWalk[8]. The projects aimed to enrich the metadata of digitised historical collections with georeferences and other information automatically computed using information extraction and georesolution technologies. Understanding location is a critical part of any historical research, and the nature of the collections make them an interesting case study for testing automated methodologies for extracting content. The projects have looked at how automatic georeferencing of resources might be useful in developing improved geospatial browsing and search capabilities within and across collections. The three collections georeferenced during the projects were:

- Histpop (History Data Service). The Online Historical Population Reports[9]
- Parl18c (BOPCRIS Digitisation Unit). 18C British Official Parliamentary Publications[10]
- The Stormont Papers (Arts and Humanities Data Service). Parliamentary Debates of the devolved government of Northern Ireland[11]

Samples from these data sources were annotated for evaluation purposes: first they were annotated for place and person name entities to provide test data for the geotagger and, subsequently, the place names were annotated for georeference resolution, once with respect to the GeoNames gazetteer and once with respect to GeoCrossWalk (except for the Stormont data since GeoCrossWalk does not have coverage of Northern Ireland). The georesolution annotations consist of the selection of the correct candidate entry from the pool of candidates returned by gazetteer look-up at the time of annotation and not only is the correct entry preserved but all the competing entries are retained as well. This kind of georesolution annotation, which we believe to be unique, has the advantage that it can be used for comparison of georesolution techniques independently of all other factors. For both stages, annotation guidelines were drawn up. Ideally some of the data would have been doubly

[4] http://www.ordnancesurvey.co.uk/oswebsite/products/50kgazetteer/
[5] Seamless Administrative Boundaries of Europe, http://www.eurogeographics.org/content/euroboundarymap.
[6] http://www.getty.edu/research/conducting_research/vocabularies/tgn/
[7] http://edina.ac.uk/projects/GeoDigRef_summary.html
[8] http://www.kcl.ac.uk/iss/cerch/projects/portfolio/embedding.html
[9] http://www.histpop.org
[10] http://www.parl18c.soton.ac.uk
[11] http://stormontpapers.ahds.ac.uk

annotated in order to monitor annotation quality through the calculation of inter-annotator agreement. Unfortunately however, project resources did not permit this. For geotagging we can infer some measure of the difficulty of the task by considering that in an earlier project ([4]), which dealt with same subset of BOPCRIS data as discussed here, the inter-annotator agreement for person and place names was 91.5%

As well as evaluating our system on the historical data annotated in-house, we have also evaluated on third-party data. Both can be used straightforwardly for geotagger evaluation but they require different approaches to evaluation of georesolution. The third-party data is the ACE 2005 English SpatialML Annotations[12] corpus [9] which contains marked-up place names resolved with respect to the Integrated Gazetteer Database (IGDB) [10]. The georesolution annotations provide lat/long and other information from the correct gazetteer entry but no information about alternative candidates is retained.

4. GEOTAGGER EVALUATION

4.1 In-house Data

Table 1 shows information about the three in-house test sets. The Histpop set is comprised of 500 documents, each corresponding to an OCRed page randomly selected from the complete Histpop collection. The BOPCRIS set contains 92 randomly selected pages from Volumes 14 and 50 of the Journals of the House of Lords. The Stormont set contains 12 randomly chosen documents but since each document represents a day of proceedings it can contain many pages the 12 documents contain a total of 471 pages. On average, a BOPCRIS page is twice the length of a Histpop page (1,118 tokens for BOPCRIS vs. 523 tokens for Histpop) and a Stormont document (15,459 tokens) is more than ten times the length of BOPCRIS page. Although there are differences in the number of documents per set, the difference in absolute corpus size is not so large. There are varying densities of place names with the Histpop data containing the largest proportion.

Table 1: Overview of In-house Geotagger Test Sets

Collection	Docs	Sents	Tokens	Places
Histpop	500	9,329	261,676	5,890
BOPCRIS	92	5,486	102,851	1,181
Stormont	12	7,601	185,503	1,216

The geotagger output was compared to the gold-standard annotations and performance was measured using the standard measures of precision, recall and F1-score (the harmonic mean of precision and recall). The results are shown in Table 2.

Table 2: Geotagger Evaluation for In-house Data

	Precision	Recall	F1-score
Histpop	82.09%	80.78%	81.43
BOPCRIS	55.92%	61.56%	58.61
Stormont	71.72%	74.67%	73.17

[12]Available from the Linguistic Data Consortium (LDC): `http://www.ldc.upenn.edu`

It can be seen from Table 2 that the results for Histpop and Stormont are considerably better than the results for BOPCRIS. The latter originates from an earlier period (1688-1817) than the other two sets and is more challenging for OCR.

4.2 SpatialML

The corpus with SpatialML annotations is comprised of 428 documents from newswire and usenet sources from 2003 2005. The corpus is not annotated for sentences or tokens but after processing through our system we count 14,615 sentences and 254,941 tokens. The original data contains 6,339 place name entities but some of these do not correspond to the notion of place name that our system aims to recognise. In particular, locational adjectives such as "Japanese" and unnamed place expressions such as "city" are annotated. Once these are removed (through an imperfect automatic script), 4,790 place names are left.

Table 3: Geotagger Evaluation for SpatialML Data

	Precision	Recall	F1-score
SpatialML	63.39%	75.26%	68.82

The results on the ACE 2005 English SpatialML corpus are not as good as reported for other systems on ACE data: the best overall F1-score by other systems for all entity mention types in ACE 2005 seems to be around 85[13] and [9] reports an F1-score of 78.5 on the SpatialML corpus. The difference is partly because the geotagger is a rule-based system designed to be relatively generic but customisable to new text types, with the version tested here being customised to the in-house test sets. By contrast, the systems that tend to perform best on newswire data like ACE are machine-learning based ones which are trained on large amounts of annotated newswire data.

5. THE GEORESOLUTION PROCESS

The input to the georesolver is an XML file with place names marked up, as well as some context-derived features concerning relations such as containment and proximity. Georesolution consists of two main stages, look-up of the place names in a gazetteer and resolution which ranks the resulting matches.

5.1 Gazetteer Look-up

First, the place names are extracted from the geotagged file and duplicate place names are reduced to a single representative. The result is an XML file containing a top-level <placenames> element and a <placename> child for each unique place name. The placenames are then passed to a gazetteer-specific look-up script: each gazetteer's script does the look-up in whatever way is appropriate but they all produce a gazetteer-independent output. The gazetteer-dependent actions are: generating queries in an appropriate format, sending them to the relevant server, and converting the results to a common format, in terms of both structure and vocabulary (feature type, for example). Queries are for both the name as it appears in the text and for any alternative names provided by the geotagger.

[13]`http://www.nist.gov/speech/tests/ace/ace05/doc/ace05eval_official_results_20060110.htm`

In the output from gazetteer look-up, each <placename> element contains a number of <place> elements which are the candidate places from the gazetteer. These elements have attributes as follows:

- lat (latitude)
- long (longitude)
- gazref (an id formed from the gazetteer name and the id returned by the gazetteer)
- in-cc (where available, the ISO country code of the containing country)
- type (feature type). Our set of feature types is deliberately coarse-grained so that other gazetteers' sets can be easily mapped to it. We do not need fine-grained types for georesolution. The types are: **water** (river, lake etc.); **civil** (administrative division); **civila** (top-level administrative division); **country** (country); **fac** (building, farm etc.); **mtn** (mountain or valley); **ppl** (populated place); **ppla** (capital of top-level administrative division); **pplc** (capital of a country); **rgn** (region); **road** (road, railway etc.); **other** (other).

Each gazetteer script is responsible for doing this mapping. After gazetteer look-up, duplicate elimination is done on the candidates for each place name, as the alternative names may have resulted in duplicate results from the gazetteer. The gazetteer look-up files are input to the resolution component described below.

There are a number of issues concerning the gazetteer look-up process which should be explored further. The scripts would benefit from more tailoring to the different gazetteers and, in particular, details of the matching such as case-sensitivity currently entries which differ only with respect to case (London vs. LONDON) are not treated as duplicates. However, it is difficult to do a case insensitive look-up if the gazetteer does not provide it; we would have to try all plausible case combinations and then eliminate duplicates. An alternative would be to download the gazetteer and search using a case-insensitive index, however, this might not always be an option if we are interfacing to an existing system which may have licensing restrictions or be dynamically updated.

The GeoCrossWalk and Unlock gazetteers are confined to Great Britain. Even for documents about Britain this leads to problems, since there are likely to be occasional references to other places, and the system will either return nothing or some quite irrelevant place with the same name. To mitigate this we augment GeoCrossWalk and Unlock with an additional list (derived from GeoNames) of places outside Britain with a population of more than 200,000. This produces more issues; for example the capitalisation problem mentioned above results in "LONDON" being found as the town in Canada, but not as the capital of England.

5.2 Georesolution

Once look-up has provided a list of candidate entries, the georesolver ranks them in order of likelihood. Before applying any of the heuristics described below, we first try to augment the information about each candidate place with information about population and containing country. This is done by consulting lists of large places derived from GeoNames and Wikipedia. If there is a place in the lists with the same name and similar latitude and longitude (within one degree), we assume a match. The information added is containing country (the attribute *in-cc*) if not already present and population (the attribute *pop*) if available.

In [8], Leidner analyses the heuristics for resolution used in earlier work such as [1], [2], [12], [13] and [14] and lists 17 which he labels H0-H16. In the list below, we categorise our heuristics by the properties of documents that motivate them, and note the related heuristics from Leidner's list.

- **Some references are unambiguous (Leidner s H0).** If the gazetteer has only one entry that is a candidate for the place name, we accept it. This is not implemented explicitly; a single candidate is bound to be first in the list. The frequency of unambiguous references depends greatly on the gazetteer, and is naturally higher with a local gazetteer than a worldwide one. In the Histpop data 11% of references are unambiguous with respect to GeoNames, and 21% with respect to GeoCrossWalk. BOPCRIS unambiguous references are 11% for GeoNames and 14% for GeoCrossWalk, while Stormont has 39% for GeoNames.
- **Multiple occurrences of the same place name in one document usually refer to the same place (Leidner s H4).** This is an important heuristic because it allows contextual information available for one instance of a place name to be applied to other instances of the same name. Our system always implicitly assumes that multiple instances refer to the same place.
- **References to important places are more common that to small ones. Importance can be estimated based on information from the gazetteer such as population (Leidner s H3 and H7) and feature type (Leidner s H6 and H12).** We prefer higher populations, and have a preference order for feature types (for example, we prefer populated places to "facilities" such as farms and mines). In addition, if the gazetteer has a limit on the number of results returned, it may well use population and feature type in deciding which candidates to discard. It might be useful to vary the feature type preference order for documents in different domains (for example, parliamentary records are likely to refer to administrative divisions), but we did not attempt this.
- **Containment and proximity information may be present in the text, for example London, England or Leith near Edinburgh (Leidner s H1).** We strongly favour candidates consistent with such contextual information; "Leith near Edinburgh" will favour candidates for Leith that are near candidates for Edinburgh, and vice versa.
- **Documents frequently refer to nearby places (Leidner s H5 and H9).** Minimising the bounding box has the disadvantage that a document concentrating on some locality may well also refer to a few distant places, resulting in a bounding box covering a huge area unrelated to the focus of the document. For example, a document describing a visit to Scotland may mention that the author flew from San Francisco via Amsterdam. Minimising all the pairwise distances has the same problem. We observe instead that it is common for places in a document to fall into clusters, so we attempt to favour candidates that are clustered. For each candidate for a place name, we compute its distance from the nearest candidate for

each other place name. We then find the average distance to the nearest five other places, and prefer candidates for which this is smaller.

- **The user may have external knowledge of the area referred to by the document (Leidner s H8).** The georesolver can be called with an optional "locality" parameter specifying the geographic focus of the document so that it prefers candidates within a given distance of a given latitude and longitude. We show the results of using this in the tables.

Each of these heuristics is assigned a value in the range 0-1, using logarithmic scaling for the population and clustering. The scaled values are combined to produce a single score for each candidate. We can potentially change this formula in accordance with things we know about the text: perhaps weight population higher and clustering lower for news articles, for example. We did not have enough data to investigate optimal weightings for the different heuristics or the scaling of the individual heuristic scores.

The output of the georesolver is the same list as was input except that the entries for each place have a score from which a ranking can be obtained. The entry with the highest score, i.e. the one ranked number 1, is the preferred reading.

6. GEORESOLVER EVALUATION

As mentioned above, the in-house data sets have georesolution annotation which preserves the list of candidates for a place name. This enables us to perform an evaluation of the resolution algorithm in isolation from any other aspect of the system: given exactly the same set of candidates as were available to the human annotator, we can test whether the entry that the system ranked highest is the same entry (i.e. has the same gazref id) as the entry chosen by the annotator. We refer to this is as exact evaluation'.

For the SpatialML corpus, we only have access to information about the correct interpretation of a place name and not the alternatives from which it was selected. To evaluate against this gold standard we define correctness in terms of lat/long: if the system's highest ranked entry has a lat/long within a certain distance of the lat/long in the gold mark-up then it is considered correct. We refer to this as proximity evaluation'. Note that the in-house data sets can be used in both kinds of evaluation and that, in fact, the proximity evaluation is useful when the gazetteer offers near-duplicates (e.g. Bristol as a populated place vs. Bristol as an administrative district) for many purposes either would be a useful choice but the exact evaluation will penalise failure to get the exact entry.

6.1 Exact Evaluation

The georesolver was evaluated against the in-house test sets under the following conditions:

- the input was the gold-standard entity mark-up to ensure evaluation of only the georesolver and not the end-to-end system;
- the gazetteer entries for the resolver to rank were exactly the entries that were available to the human annotators for Histpop and BOPCRIS there were two gold annotation sets, one using GeoNames and one using GeoCrossWalk while Stormont was annotated only for Geonames (since GeoCrossWalk does not cover Northern Ireland);
- the gold standard entity mark-up was automatically augmented with linguistic context features (concerning alternate names, containment, proximity etc.) so that the georesolver would have the same information that would be provided in geotagged data;
- georesolution was tested both with the user supplied locality parameter switched off and with it switched on. For Histpop and BOPCRIS the setting was "-l 55.45 -5.2 655 .3" (to cover the British Isles) and for Stormont the setting was "-l 54.6 -6.8 92 .3" (to cover Northern Ireland);

Table 4: Exact Evaluation for In-house Data

Histpop	GeoNames	GeoCrossWalk
documents	499	500
place names	5,882	5,890
no candidate	424	1,203
'none' selected	349	252
no selection	18	0
non-'none' selected	5,091	4,435
baseline	1,113 (21.9%)	1,983 (44.7%)
correct without locality	3,554 (69.8%)	2,833 (63.9%)
correct with locality	3,835 (75.3%)	2,835 (63.9%)
BOPCRIS	GeoNames	GeoCrossWalk
documents	92	92
place names	1,181	1,181
no candidate	339	462
'none' selected	80	43
no selection	27	26
non-'none' selected	735	650
baseline	156 (21.2%)	233 (35.8%)
correct without locality	494 (67.2%)	515 (79.2%)
correct with locality	565 (76.9%)	515 (79.2%)
Stormont	GeoNames	
documents	12	
place names	1,216	
no candidate	150	
'none' selected	74	
no selection	7	
non-'none' selected	985	
baseline	480 (48.7%)	
correct without locality	836 (84.9%)	
correct with locality	888 (90.2%)	

During gold annotation, a number of cases arose: (1) no gazetteer entry was found during gazetteer look-up; (2) entries were found but the human annotator considered that there was no correct entry (they selected none'); (3) entries were found but the human annotator neither chose one nor selected none' we consider these to be annotation errors; (4) entries were found and the human annotator selected one of them as correct. In Table 4 we exclude cases (1) to (3) from the evaluation though we indicate the numbers of each of the cases. We exclude the second case as the system will always choose one of the entries because it is not designed to make a none of the above' judgement. Table 4 also shows the effects of the use of the locality parameter. We have included a baseline which is the score that would be obtained by randomly selecting entries. Note that the baseline gives some indication of how ambiguous the names are and that the greater ambiguity from using GeoNames leads to a lower baseline result.

6.2 Proximity Evaluation

As discussed above, proximity evaluation can be used on any test set where place names have been associated with lat/longs. This means that we can perform proximity evaluation of our system on both the in-house data and SpatialML. Since the system can be used with more than one gazetteer, we can evaluate using any of the gazetteers that are available to us. Here we consider GeoNames and GeoCrossWalk for our in-house test sets and GeoNames and Unlock for the SpatialML corpus (since GeoCrossWalk is no longer available and couldn't be used for more recent experiments). Note that, being restricted to Great Britain, GeoCrossWalk and Unlock would not be expected to perform well on world-scale text. However they are both supplemented with large populated place entries from GeoNames (as described in Section 5) and it is interesting to investigate their performance.

In the evaluations reported in this section, the georeferencer starts from gold standard place name entities so that georeferencing is evaluated in isolation. In the experiments summarised in the tables, we have used 5km as the proximity measure: if a system lat/long is within 5km of a gold lat/long then it is considered correct. We have used the same locality options as described above for the in-house data but not for SpatialML since this deals with world news.

Table 5: Proximity Evaluation for In-house Data

	geonames	xwalk
Histpop with locality		
no. of place names	5091	4435
no. for which gaz entries found	5091	4435
correct within 5km	4177	4112
as % of total	82.0%	92.7%
BOPCRIS with locality		
no. of place names	735	650
no. for which gaz entries found	735	650
correct within 5km	598	593
as % of total	81.4%	91.2%
Stormont with locality		
no. of place names	985	
no. for which gaz entries found	985	
correct within 5km	905	
as % of total	92.1%	

Table 5 provides an alternative view of the system's performance on the in-house data: in Table 4 we saw performance measured by a strict criterion while Table 5 looks at the same system output but analysed with a less strict measurement where in all cases the percentage of correct resolutions increases significantly. This increase is a reflection of the fact that gazetteers contain multiple entries for essentially the same place.

Table 6 shows proximity results for SpatialML. There are a total of 3628 placenames in the gold annotations once we have removed certain cases: places with no lat/long assigned, non-named places (e.g. "city", "region") and adjectival forms of place names (e.g. "Iraqi"). The use of the system with GeoNames provides results which are comparable to the results on the in-house data. Unlock is less successful but we can take this to be a consequence of its being focused on Great Britain with additional knowledge of only very large places elsewhere in the world. The fact that it was unable to find gazetteer entries for 579 places seems to confirm this conclusion. Our results fall short of those reported by [9] whose disambiguator scores 93.0 F-measure. It is not clear exactly how their scoring mechanism operates so it is hard to draw conclusions from a comparison with our results. In [2], Clough reports 89% accuracy on the SPIRIT evaluation set (130 web pages, 1,864 unique place names). These were selected to contain only UK place names and were therefore less diverse than SpatialML.

Table 6: Proximity Evaluation for SpatialML

	GeoNames	Unlock
SpatialML		
no. of place names	3628	3628
no. for which gaz entries found	3538	3049
correct within 5km	2946	2143
as % of total	81.2%	59.0%

7. END-TO-END EVALUATION

So far we have evaluated georesolution in isolation by giving it gold standard entities as input. As described in Section 4, geotagger performance is imperfect and end-to-end evaluation of the system will give a more realistic view of system performance. In this section we report on an end-to-end evaluation on SpatialML of our system using the GeoNames gazetteer: Table 7 shows the results. The gold corpus contains 3628 place names and our system finds 2923 of these, which means there are 705 entities in the gold corpus which our system failed to recognise. Georesolution using GeoNames of the place names it did recognise scores 69% of the total using proximity evaluation within 5km.

Table 7: End-to-end Evaluation for SpatialML

	System	Placemaker
SpatialML		
no. of place names	3628	3628
no. for which gaz entries found	2923	2635
correct within 5km	2504	882
as % of total	69.0%	24.3%
correct within 25km	2520	1067
as % of total	69.5%	29.5%
correct within 50km	2558	1677
as % of total	70.5%	46.2%
correct within 100km	2664	2133
as % of total	73.4%	58.8%

The functionality of our system is similar to that of Yahoo! Placemaker and it is possible to compare Placemaker end-to-end with our end-to-end system: the results for Placemaker are reported in the third column of Table 7. In both cases we removed all mark-up from the SpatialML corpus and submitted the files to the respective systems. Where our system recognises 2923 of the place names, Placemaker finds 2635. A glance at a sample of the missed places for our system shows a consistent failure to recognise all lower-case place names (e.g. "afghanistan"), inability to handle spelling errors (e.g. "Ameirca") and many miscellaneous failings of the recognition rules. A glance at a sample of the missed places for Placemaker show a high-precision recognition system which might well out-perform ours except that assumptions about the boundaries of place names differ from the mark-up in SpatialML. For example, "Addison, Texas" is one place name for Placemaker and two for SpatialML and

our comparison algorithm penalises Placemaker unfairly in such cases.

We compared the georesolution results of both our system and Placemaker to the gold lat/long annotation and performed proximity evaluation with a range of distances. With 5km as the maximum distance, Placemaker performs very poorly in comparison to our system. However, we believe that much of the difference is accounted for by gazetteer differences. GeoNames and the Integrated Gazetteer Database (IGDB) which is used for the gold annotation appear to have been drawn from similar sources so that their lat/long values for large places such as countries and states are very close. Placemaker uses Yahoo! GeoPlanet's Where On Earth IDs (WOEIDs) and has lat/longs which, while correct, are further from the IGDB ones. For example, Japan has a lat/long of 36.000,138.000 in both GeoNames and the IGDB while the lat/long from Placemaker is 37.4876,139.838. Both point to Japan but only the Geonames one is within 5km of the IGDB one. For this reason we have shown proximity evaluation results for a range of distances and it can be seen that while our system results only show a slight improvement given larger distances, the Placemaker results improve quite significantly.

A large proportion of the SpatialML places are countries there are 3,568 <PLACE type= COUNTRY'> elements in the SpatialML corpus. Of these, 1,544 have lat/long values and are not location adjectives and are therefore retained in our gold version of the corpus. To explore the issue further, in a follow-on experiment we performed the same evaluation as reported in Table 7 except that we counted as correct any Placemaker country entity which matched a country entity with the same name in the gold corpus. With proximity of 5km, the number of correctly resolved place names rose sharply to 2,086 (57.5%). There are some other entity types in our gold version of the corpus, such as CIVIL' (583, used e.g. for U.S. states) and RGN' (41, used for place names such as "Middle East") which might give rise to the same evaluation issues as country names. If these were taken into account, the Placemaker results would probably improve further, though we have not attempted to test this. The experiments that we have done show that our system and Placemaker both appear to perform moderately well on the SpatialML corpus but they also serve to highlight some of the problems that arise when attempting to perform comparative evaluations.

8. CONCLUSIONS AND FURTHER WORK

In this paper we have described our georeferencing system and have conducted a range of evaluations, especially of the georesolution component. We hope to have provided some insight into the complexity of this kind of evaluation where the numbers depend on a wide range of factors. The experiment described in the previous section demonstrates that there is still much work to be done to provide a clear comparison between systems which use different gazetteers.

We have described our work with historical collections but have not discussed issues relating to historical change. It is not ideal to georeference historical collections using contemporary gazetteers because spelling changes may affect look-up success and because administrative boundaries change over time. In future work we would hope to properly address these issues.

9. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR*, 2004.

[2] P. Clough. Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of Workshop on Geographic Information Retrieval (GIR'05)*, 2005.

[3] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the Association for Computational Linguistics*, 2002.

[4] C. Grover, S. Givon, R. Tobin, and J. Ball. Named entity recognition for digitised historical texts. In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[5] C. Grover and R. Tobin. Rule-based chunking and reusability. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, 2006.

[6] C. Grover, R. Tobin, K. Byrne, and M. Woollard. Use of the Edinburgh geoparser in the GeoDigRef and Embedding GeoCrossWalk projects. Technical report, School of Informatics, University of Edinburgh, `http://www.inf.ed.ac.uk/publications/report/`, 2010.

[7] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball. Use of the Edinburgh geoparser for georeferencing digitised historical collections. *Phil. Trans. R. Soc. A*, Submitted.

[8] J. L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, School of Informatics, University of Edinburgh, 2007.

[9] I. Mani, J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner. SpatialML: Annotation scheme, corpora, and tools. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[10] S. Mardis and J. Burger. Design for an integrated gazetteer database: Technical description and user guide for a gazetteer to support natural language processing applications. Technical report, Mitre, 2005.

[11] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang. The design and implementation of SPIRIT: a spatially-aware search engine for information retrieval on the internet. *International Journal of Geographic Information Systems (IJGIS)*, 21(7), 2007.

[12] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL Workshop on Analysis of Geographic References*, 2003.

[13] F. Schilder, Y. Versley, and C. Habel. Extracting spatial information: grounding, classifying and linking spatial expressions. In *Proceedings of Workshop on Geographic Information Retrieval (GIR'04)*, 2004.

[14] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *Fifth European Conference for Digital Libraries (ECDL 2001)*, 2001.