Creating a Novel Geolocation Corpus from Historical Texts

Grant DeLozier[*,**], **Ben Wing**[*,**], **Jason Baldridge**[*], **Scott Nesbit**[†]
University of Texas at Austin[*]
University of Georgia[†]
grantdelozier@gmail.com, ben@benwing.com
jasonbaldridge@gmail.com, snesbit@uga.edu
[*] The first two authors contributed equally to the content of the paper.

Abstract

This paper describes the process of annotating a historical US civil war corpus with geographic reference. Reference annotations are given at two different textual scales: individual place names and documents. This is the first published corpus of its kind in document-level geolocation, and it has over 10,000 disambiguated toponyms, double the amount of any prior toponym corpus. We outline many challenges and considerations in creating such a corpus, and we evaluate baseline and benchmark toponym resolution and document geolocation systems on it. Aspects of the corpus suggest several recommendations for proper annotation procedure for the tasks.

1 Introduction

Geographic information is an important component of a number of areas including information retrieval (Daoud and Huang, 2013), social media analysis, and historical research (Nesbit, 2013; Grover et al., 2010; Smith and Crane, 2001). To date however, very few corpora exist for text geolocation tasks, and those which do exist have flaws or are very small in size. This is particularly true for tasks seeking to do geolocation work with historical texts. In the realm of document geolocation, there exist no historical corpora whatsoever; in the realm of toponym resolution historical corpora exist, but are flawed in important respects (Speriosu and Baldridge, 2013; DeLozier et al., 2015).

This paper describes the process of annotating a set of American Civil War archives commonly known as the *Official Records of the War of the Rebellion* (officially titled *The War of the Rebellion: a Compilation of the Official Records of the Union and Confederate Armies* and henceforth abbreviated as WOTR), arguably the most important and widely used corpus in this area of historical study[1].

	Docgeo subset	Topo subset	Full data
Total tokens	1,743,331	447,703	57,557,037
# volumes	118	15	126
# documents	7,533	1,644	254,744
Avg. tokens/document	231.43	272.32	225.94

Table 1: Statistics on WOTR, annotated subset and full data (using documents predicted based on a sequence model derived from the annotated data, as described in §3).

Document geolocation and toponym resolution enable work on the specific content of individual documents and themes contained within this corpus, revealing the ways in which content is distributed in the corpus over time and space (Ayers and Nesbit, 2011; Thomas III, 2011). Themes in this corpus pertinent to the study of Civil War literature include the rise of irregular warfare, the end of slavery in Confederate and Union states, the use of railroads by United States and Confederate armies in the war, and the destruction of the war-making capacity of the Confederate states. The annotation process and geolocation tools also enable historians to reexamine the process by which the archive was produced, an area which has recently seen growing interest (Sternhell, 2016).

We develop geolocation corpora for two related but separate tasks: document geolocation (docgeo) and toponym resolution (TR). Statistics on the full WOTR corpus and the annotated document geolocation and toponym subsets are shown in Table 1 and Table 2.

Geographic summaries of the annotations are given in Figure 1 (documents) and Figure 2 (toponyms). The docgeo annotations are concen-

[1] http://ehistory.osu.edu/books/official-records

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 188–198,
Berlin, Germany, August 11, 2016.

	Docgeo Subset
Documents	8,121
Documents with geometries	5,035 (62%)
Documents with only points	4,811 (59%)
Documents with polygons	224 (3%)

	Topo Subset
Avg. toponyms/document	7.17
Toponyms	11,795
Toponyms with geometries	10,380 (88%)
Toponyms with points	8,130 (69%)
Toponyms with polygons	2,296 (19%)
People	7,994
Organizations	2,591

Table 2: Statistics on WOTR, annotated subset (using documents predicted based on a sequence model derived from the annotated data, as described in §3).

trated in a number of areas that saw heavy fighting, such as in Virginia, South Carolina and Northern Georgia. The toponym annotations are more concentrated around the western theater of the Civil War. In both corpora, almost all US states are represented by at least some references. The toponym annotations contain more full-state polygons, while the docgeo annotations are primarily points, leading to the differing appearances of the two maps.

2 Geolocation tasks

Both toponym resolution and document geolocation involve assigning geographic reference, usually latitude-longitude coordinates, to spans of text, but differ as to the size of the span. Toponym resolution involves assigning such reference to individual, potentially ambiguous toponyms (e.g. *Springfield* or *Dallas*), while document geolocation assigns geographic reference to larger spans of text (documents, broadly construed).

Among the key difficulties associated with both tasks are ambiguity of reference, fluidity in the definition of the tasks, and lack of sufficient and/or appropriate training material. As an example of the issues surrounding ambiguity, consider the toponym *Springfield*. Dominant place name gazetteers indicate at least 236 unique senses of the term (and these underestimate the true total), with possible references spanning the globe. TR systems must choose referents in these highly ambiguous scenarios, even when correct referents are not listed in gazetteers. In document geolocation, the problem is even more acute, as a document can potentially be assigned a location anywhere on the globe.

Another issue affecting both domains is fluidity in how one defines the task itself. In toponym resolution, metonymy—the ability of a place name to refer to something closely related to a place (e.g. a government)—and demonymy—names for the people who inhabit an area (e.g. Americans)—are properties that must be considered. All existing TR corpora include metonymic uses of place names. The Local Global Lexicon (LGL) corpus (Lieberman and Samet, 2012) includes demonyms as toponyms and georeferences them, while all other corpora do not. An additional issue pertains to the range of entity types a system is expected to resolve. Many corpora limit their expectations to larger entities—e.g. TR-CoNLL (Leidner, 2008) is limited to cities, states, and countries), while others focus more on highly local entities (e.g. bus stops) (Matsuda et al., 2015). A final issue relates to whether systems ought to resolve places which are embedded inside other named entities. For example, the LGL corpus expects *New York* in the expression *New York Times* to be resolved to the state of New York. Many of the characteristics of existing TR corpora are summarized in Table 3.

In document geolocation, different researchers have interpreted the task differently, depending on the corpus: typically as either as the location of the document's author when the document was created, or as the geographic *theme* (i.e. topic) of the content of the document. The former interpretation has usually been used when working with social-media corpora such as Twitter (Han et al., 2014; Schulz et al., 2013) and Flickr (O'Hare and Murdock, 2013; Bolettieri et al., 2009), and the latter with encyclopedic corpora such as Wikipedia (van Laere et al., 2014) and historical corpora such as the unpublished Beadle Corpus (Wing, 2015). Another difficulty with using the geographic-theme interpretation is that this reference may not be easily identifiable for some texts. (For example, only about 10% of the articles in the English Wikipedia have document-level annotations assigned to them.)

An additional issue related to the definition of both tasks is the scope of the geographic reference. Smaller geographic entities, such as cities and neighborhoods, can be reasonably approximated as a point in latitude-longitude space, while

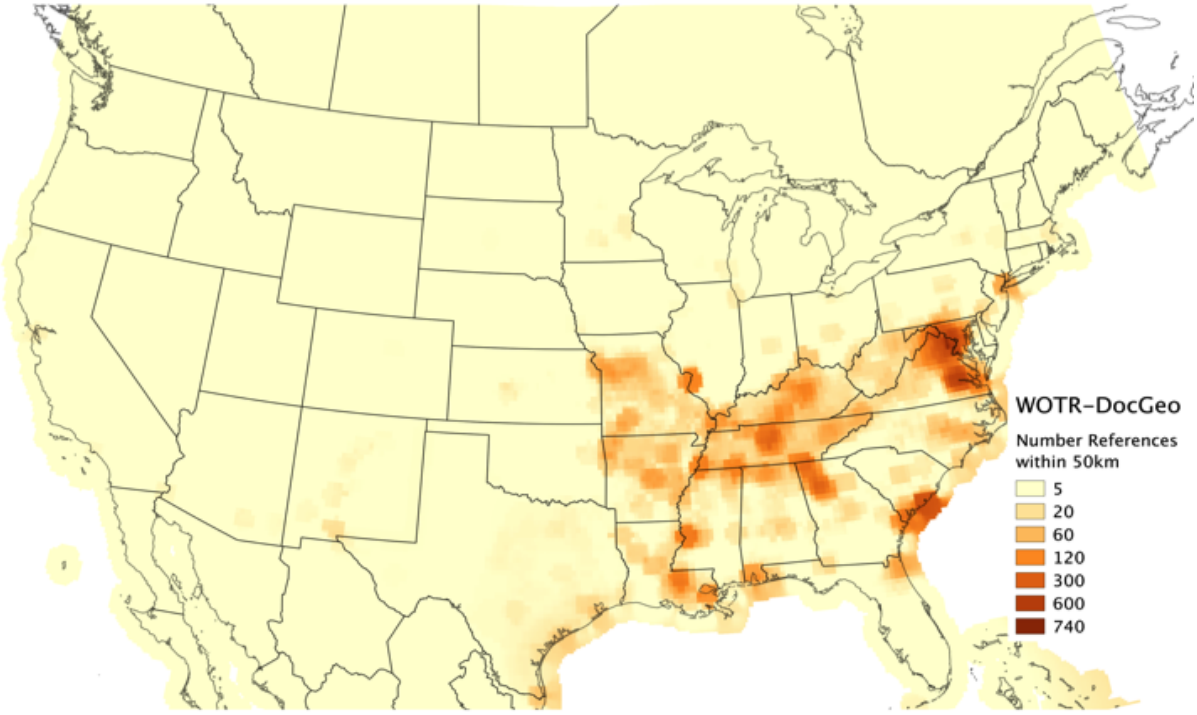

Figure 1: Distribution of References in WoTR-DocGeo

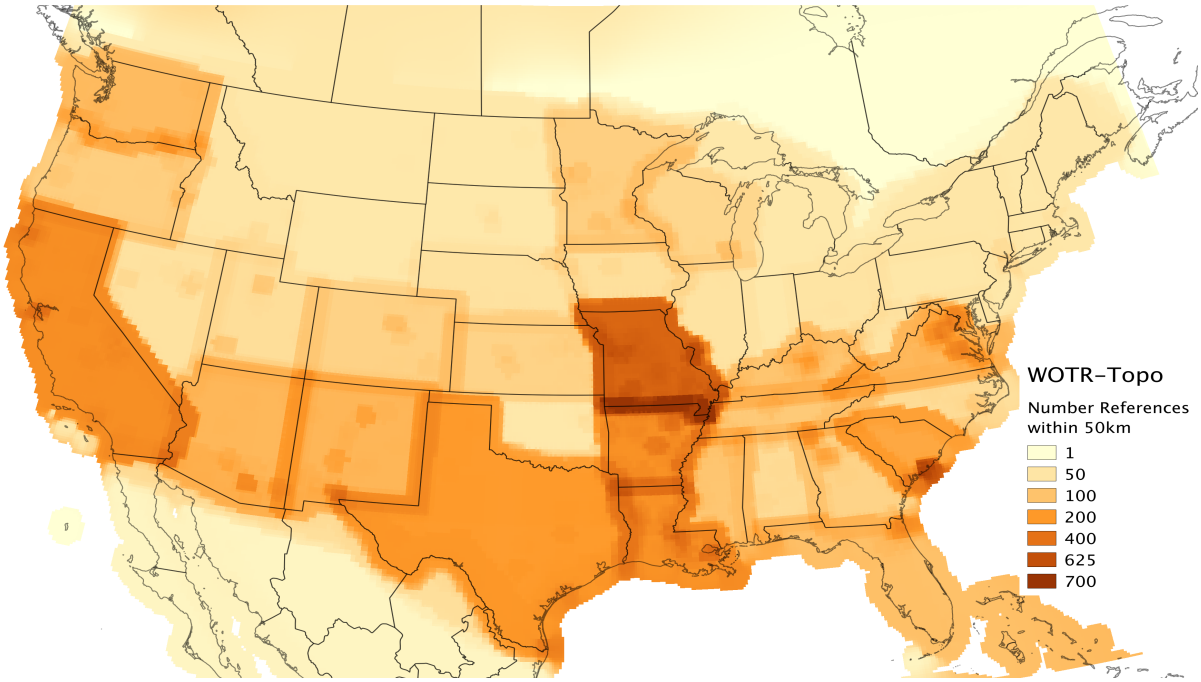

Figure 2: Distribution of Toponyms in WOTR-Topo

it is more difficult to do so for larger entities such as states or countries. Various solutions have been used for this problem, depending on the corpus. Wikipedia and most gazetteers take the simplest approach of assigning a point to all entities, regardless of size. However, for large entities such as countries, this necessitates choosing a single representative point (e.g. the geographic centroid or the capital city), which leads to many problems (e.g. the geographic centroid of the UK is a point in the Irish Sea).

Toponym resolution, especially, currently suffers from a lack of sufficient training material. Existing training corpora fixate around very narrow ranges of geographic entities. One major corpus used in toponym resolution, TR-CoNLL, has only 800 unique strings and 6259 toponyms, while gazetteers such as GeoNames list over 8 million unique places (which still greatly underestimates the true number of toponyms). Such mismatches do more than underscore the need for larger and more domain-diverse corpora; they point to fundamental issues associated with learning to resolve geographies from language. Geographic entities, like all named entities, are fiat objects; naming them dictates their existence (Kripke, 1980). Many systems have attempted to alleviate paucity problems by splicing corpora with latent annotations inferred from a more general resource like Wikipedia (Speriosu and Baldridge, 2013; Santos et al., 2014; DeLozier et al., 2015).

In document geolocation, the amount of training material available is crucially tied in with how the task is defined (as described above). Abundant training material is available from the various language-specific versions of Wikipedia and from social-media sites such as Twitter and Flickr, but the variations in language and task definition make the corpora highly domain-specific. This means that cross-corpus generalization is fraught with difficulty, particularly in domains where no previously-published corpora exist, such as historical documents. Nonetheless, researchers have achieved some success from docgeo domain adaptation, using Wikipedia as out-of-domain training material for historical documents under a co-training setup (Wing, 2015) and Flickr as a source of language-model data for geolocation of Wikipedia (De Rouck et al., 2011).

3 Data preparation

The source data available was in the form of text scanned directly from the published books using OCR (optical character recognition), and then hand-corrected. The digital form of the collection we accessed included page breaks which sometimes occur in the middle of a word, footnotes and headers undifferentiated from body text, and no formal delimiting of where particular records began and ended. Figure 3 is an example of part of the source text of a volume in the collection, after preprocessing to stitch up page breaks and remove footnotes, headers, footers, etc., but before splitting into individual documents.

To alleviate some of these issues in working with this form of the text, the following steps were taken to improve our annotated version of the corpus:

1. Remove page breaks and stitch up paragraphs divided across the breaks.

2. Create a GUI annotation tool to allow annotators to quickly note the extent of documents (which we term *spans*) and indicate the document locations on a map.

3. Create a sequence model to automatically split up the continuous text into documents, training it on the documents manually marked up by the annotators.

Stitching up page breaks As mentioned above, the source text is in the form of individual pages scanned from the published books, with page breaks, footnotes, stray headers, etc. often interrupting a paragraph in the middle of a word, frequently in an inconsistent fashion. A program was written that used various heuristics to do the majority of work, although several more steps and a good deal of hand editing were required to achieve satisfactory results.

Automatically locating document spans There is no indication in the source text where one document ends and another one begins. In a letter, for example, sometimes the destinee appears near the beginning of the letter, following a heading describing the location and date, while in other cases the destinee appears at the very end, after the salutation. Both examples can be seen in the text box in the annotation tool screen shot in Figure 4,

Table 3: Toponym Corpora

Corpus	Domain	Entity Types	Reference Types	Metonyms	Demonyms	Nested NE	Toponyms
TR-CoNLL	Contemporary International News	Cities,States, Countries	Point only	Yes	No	Most Encompassing NE	6259
LGL	Contemporary Local Newspapers	Few Locales, cities, states, countries	Point only	Yes	Yes	Annotates Embedded Places	5088
LRE	Tweets from Japan	Highly local 'facilities' and above	Point only	?	No	?	951
WOTR	US Civil War Letters + Reports	Locales, Cities, and States	Point and Polygon	No	No	Most Encompassing NE	10380

along with the way that successive documents directly abut each other. Because the unit of analysis is a single document, it is necessary to locate the beginning and end of each document, and this must be done automatically since only a fraction of the text was manually annotated.

To do this, a CRF (conditional random field) sequence model was created using MALLET (McCallum, 2002). Each successive paragraph was considered a unit in the sequence labeling task, and labeled with one of the following: *B* (beginning), *I* (inside), *L* (last), or *O* (outside), similar to how named entity recognition (NER) sequence labeling is normally handled. CRF's have the advantage over HMM's (hidden Markov models) that they can be conditioned on arbitrary features of the visible stream of paragraphs, including the neighbors of the actual paragraph being labeled. This allowed for various features to be engineered, such as (1) the presence of a date at the end of a line, possibly followed by a time; (2) the presence of certain place-related terms typically indicating a header line, such as *HEADQUARTERS*, *HDQRS* or *FORT*; (3) the presence of a rank-indicating word (e.g. *Brigadier*, *General* or *Commanding*) at the beginning of or within a line; (4) the presence of a line beginning with a string of capital letters, typically indicating a header line; (5) the presence of certain words (e.g. *obedient servant*) that typically indicate a salutation; (6) the combination of the above features with certain punctuation at the end of the line (comma, period, or colon); (7) the length of a line; (8) all of the above features for the actual paragraph in question as well as the previous, second-previous, next, second-next, and combinations thereof; and (9) the first and last words of the paragraph, after stripping out punctuation.

The resulting model performed well, but did not consistently handle the cases where the destinee is at the end of the letter, and so a postprocessing step was added to adjust the spans whenever such a situation was detected.

4 Annotation process

4.1 Annotation tool

A GUI annotation tool was written that allows document spans to be selected in a text box and points or polygons added on a map. Figure 4 shows a screen shot of the tool at work. Spans of text are indicated with inward-pointing red arrows at their edges and are colored yellow (a marked span without geometry), green (a span with geometry) or cyan (currently selected span for adding or changing the geometry). Points and polygons can be added by drawing directly on the map, by using the list of recent locations below the map, or (in the case of points) by entering a latitude/longitude coordinate into the text box and clicking **Set Lat/Long**.

The annotation tool is written in HTML and JavaScript using the OpenLayers[2] and Rangy libraries[3], with data stored using Parse, a *backend-as-a-service* which allows for free data storage within certain storage and bandwidth limits.

4.2 Document geolocation annotation

The docgeo annotation process took 280 hours over two months. Five annotators were hired, although in practice most of the work was done by a single annotator. 25-page subsections of 118 of 126 volumes were annotated with geographies. A few of the volumes had an additional 75 pages annotated.

4.2.1 Document annotator guidelines

Annotators were hired to note the individual documents within the archives and attach document-level geometries to them, which are intended to encode the geographic *theme* of the content of the

[2] http://openlayers.org/

[3] https://github.com/timdown/rangy

...

2. While congratulating the troops on their glorious success, the commanding general desires to impress upon all officers as well as men the necessity of greater discipline and order. These are as essential to the success as to the victorious; but with them we can march forward to new fields of honor and glory, till this wicked rebellion is completely crushed out and peace restored to our country.

3. Major-Generals Grant and Buell will retain the immediate command of their respective armies in the field.

By command of Major-General Halleck:

N. H. McLEAN,

Assistant Adjutant-General.

HEADQUARTERS DEPARTMENT OF THE MISSISSIPPI,
Pittsburg, Tenn., April 14, 1862.

Major General U. S. GRANT,

Commanding District and Army in the Field:

Immediate and active measures must be taken to put your command in condition to resist another attack by the enemy. Fractions of batteries will be united temporarily under competent officers, supplied with ammunition, and placed in position for service. Divisions and brigades should, where necessary, be reorganized and put in position, and all stragglers returned to their companies and regiments. Your army is not now in condition to resist an attack. It must be made so without delay. Staff officers must be sent out to obtain returns from division commanders and assist in supplying all deficiencies.

H. W. HALLECK,

Major-General.

NEW MADRID, April 14, 1862.

J. C. KELTON:

General Pope received message about Van Dorn and Price. Do you want his army to join General Halleck's on the Tennessee? His men are all afloat. He can be at Pittsburg Landing in five days. Fort Pillow strongly fortified. Enemy will make a decided stand. May require two weeks to turn position and reduce the works. Answer immediately. I wait for reply.

THOMAS A. SCOTT,

Assistant Secretary of War.

SPECIAL ORDERS, HDQRS. DIST. OF WEST TENNESSEE,
No. 54. Pittsburg, Tenn., April 14, 1862.

II. Brigadier General Thomas A. Davies, having reported for duty to Major-General Grant, is hereby assigned to the command of the Second Division of the army in the field.

By order of Major-General Grant:

[JNumbers A. RAWLINS,]

Assistant Adjutant-General.

CAIRO, ILL., April 14, 1862.

H. A. WISE, Navy Department:

...

Figure 3: Example of WOTR source text, after stitching up text across page breaks, removing extraneous headers/footers/footnotes, etc.

document. The theme of a document is the primary location or locations that the document concerns. For example, if the document describes a battle, skirmish or other military action, the location of that action is the document's geography. Most correspondence is headed by the location at which it was written, which is often the same as the geographic theme, depending on what the content of the correspondence says. Annotators were allowed to mark multiple locations or to draw a polygon around an area of the map, which is useful when for example the geographic theme is logically a body of water or a section of a state rather than a single point. However, in the interests of achieving as many annotations as possible, annotators were encouraged to not overly make use of polygons or multiple points, preferring a single point when possible. In particular, the mere mention of a place name in a document is not sufficient for it to be included in the geographic theme; it must be of primary relevance to the subject of the document.

Annotators were encouraged to look up toponyms found within the text to retrieve their latitude/longitude coordinates, with helpful relevant keywords attached as necessary, such as *Civil War* or the region or commander mentioned in the larger document context. Annotators were shown how to retrieve the geocoordinate from Wikipedia pages, which was by far the most-frequently used resource, although Google Maps and niche US Civil War websites were used as well.

4.2.2 Document annotation challenges

Geographically diverse documents A large fraction of documents mention multiple places, and our annotators frequently struggled with determining the geographic theme of these documents, preferring to mark multiple points in questionable cases. These cases are common, with an average of 1.84 points per annotated document. The systems whose results are described in Table 5 are designed to work with documents annotated with a single point; to handle multiple-point documents, the centroid of the points was taken.

Difficult to geolocate documents The geographic theme of many documents is difficult to determine because they don't mention any easily identifiable locations. Some documents contain only ad-hoc names (e.g. *McCullan's Store* or temporary army camps named after individual com-

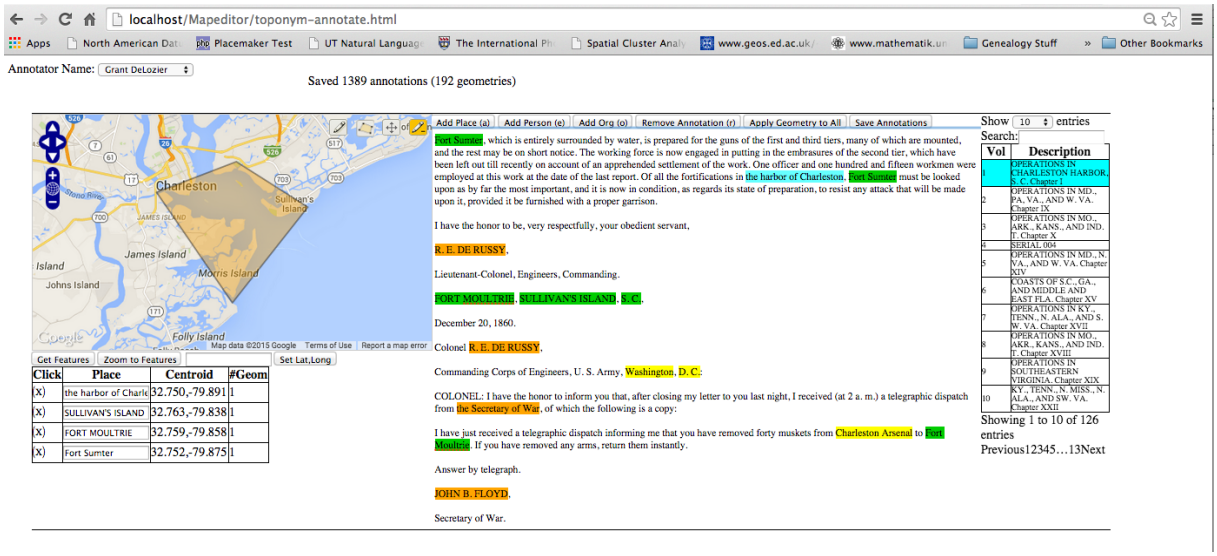

Figure 4: Screen shot of the toponym annotation tool. Place names highlighted in yellow, place names with geoemetries in green.

manders). Many documents mention only a location relative to a previously-specified location in a different document, making the theme discoverable only by looking at the whole series of correspondence. In some cases no clear geographic theme exists at all. In all such cases (amounting to about 38% of the total), the annotators assigned no geometry to the documents.

4.3 Toponym annotation

To begin the toponym annotation procedure, we identified a subset of the volumes which had been annotated with document geolocations (subsections of 15 volumes, selected in part for geographic and topic diversity). Stanford's Named Entity Recognizer (NER) was then run on the collection of documents, using the standard MUC, CoNLL trained models (Finkel et al., 2005). The place annotations that Stanford NER produced were used as a pre-annotated set, which annotators were then asked to correct and add geographic reference to.

The toponym annotation process, which spanned 4 months and occupied 290 hours, resulted in the annotation of 11,795 toponyms (10,389 with geometries) spanning 1,644 annotated documents across 100 page subsections of 15 volumes. Originally all toponym annotations were done by a single annotator. After this process all of the original annotations were reviewed by a second team of three annotators. These annotators were asked to correct a number of problems with the annotations that were not realized until after the initial annotation process had finalized. Corrections to the original annotation mostly focused on building consistent approaches to the challenges outlined in §4.3.2.

4.3.1 Toponym annotator guidelines

Annotators were asked to quickly scan the documents and look for place names. Place names which were not detected by Stanford NER should be added, and other entities incorrectly classified as places should be deleted. We directed annotators to include point, multi-point, polygon, and multi-polygon geometries where appropriate.

They key guidelines annotators were given for the task concerned three aspects of toponyms: metonymy, demonymy, and nested named entities. Annotators were asked to exclude metonymic and demonymic names from annotation. Named Entity Classification researchers have typically adopted the stance of annotating the most encompassing named entity (Finkel and Manning, 2009), though there are exceptions to this trend as is the case in the LGL corpus. Following the majority of related work, we ask annotators to only mark toponyms which constitute the most encompassing named entity (e.g. *44th Virginia Cavalry* is marked as an organization, and in this case the word *Virginia* would not be marked). Not included among nested named entities are toponym hierarchies, or disambiguators such as in the phrase *Richmond, VA, CSA*. In these cases each toponym is annotated with separate reference. To find the reference of places, annotators were allowed access to Internet search. As with document geolocation, annotators were encouraged to look up

troublesome toponyms on the Internet, and mostly made use of Wikipedia.

4.3.2 Toponym annotation challenges

Conjunctive toponyms (toponyms that are joined by conjunctions) are a problem when they are in the form of *Varnell's and Lovejoy's Stations*. Here we assumed two toponyms should be added. However, due to how our GeoAnnotate tool worked, we could not annotate overlapping, discontinuous spanning place names. In these cases we asked annotators to mark *Varnell's* as a place separate from *Lovejoy's Stations*, including the *Stations* term only with the second toponym.

Possessive toponyms (toponyms partially consisting of a person's name) appeared in the corpus, e.g. *Widow Harrow's house*. Originally, we asked annotators to avoid annotating these as toponyms. We later amended our guidelines to ask annotators to mark these as toponyms only when the possessed entity was capitalized (e.g. Varnell's Station would be annotated).

Difficult Toponyms (toponyms that could not be geographically referenced) made up about 12% of the overall toponyms in Wotr-Topo. This was typical of toponyms that described the locations of ferries, bridges, railroads, and mills. These features usually no longer exist, so discovering their exact reference even with access to Google is very difficult.

Rivers, and physical features are difficult to reference geographically because their geometric definitions are often highly complex, vague, and poorly defined in gazetteers. Rather than ask annotators to annotate the full extent of rivers, we asked them to mark a point on the river that they felt was most relevant to the context. Annotators tended however to opt for whichever point the river's Wikipedia page indicated, though this was not always the case.

Geographically vague toponym regions appear in the texts. Some of the common examples appearing in the text are *the North*, *the South*, *the West*, and *Northern Mississippi*. We asked annotators to mark these as toponyms, and attempt to draw their reference given the context.

Referring Expressions (e.g. *the stone bridge*) are common. We originally asked annotators not to annotate them, yet we failed to anticipate referring expressions which were partially constituted of place names (e.g. *the Dalton road*). Given that these expressions contain proper place names, and are places themselves, we decided to ask annotators to try and reference the whole expression (i.e. the location of the road). Unfortunately though, discovering the georeference of such roads is very difficult, and annotators tended to mark the location as a point near one of the embedded city toponyms.

Embedded Named Entities : We gave our annotators a rule to only annotate the entity type of the *most-encompassing* named entity. Using this rule expressions like *44th Virginia Cavalry* became annotated as one single organization, rather than a place inside an organization. We did not anticipate however the range of semantically equivalent expressions such as *44th Cavalry of Virginia* or *44th Cavalry from Virginia*. The former form we tended to mark as an organization, while the latter we marked as an organization *44th Cavalry* plus a toponym *Virginia*.

5 Baseline and benchmark system evaluation

In order to gain an understanding of the difficulties of the corpus and encourage its adoption, we evaluate the performance of a number of baseline and benchmark systems on the dataset.

For docgeo, two methods are used for constructing grid cells: **Uniform** and adaptive (**KD**), which adjusts cell sizes to equalize the number of documents in each cell (Roller et al., 2012). **LR** uses flat logistic regression while **Hier** constructs a coarse-to-fine hierarchy of grids with a beam search (Wing and Baldridge, 2014)[4].

For TR, **Population** selects a matching gazetteer referent with the highest population. **WISTR** is a bag of words multinomial logistic regression model trained on Wikipedia (Speriosu and Baldridge, 2013). **SPIDER** is a weighted distance minimization approach that prefers selecting gazetteer referents that occupy minimal area (Speriosu and Baldridge, 2013). **TopoCluster** uses a geographic density estimation of the toponym and context words; **TopoClusterGaz**[5] additionally 'snaps' to the nearest gazetteer referent (DeLozier et al., 2015). All TR systems were

[4] https://github.com/utcompling/textgrounder

[5] https://github.com/grantdelozier/TopoCluster

Table 4: WoTR Toponym Resolution Results

System	A@161	Mean	P	R	F-1
Random	22.2	2216	14.8	6.4	8.9
Population	63.1	1483	42.2	18.2	25.4
SPIDER	67.1	482	37.8	16.3	22.7
WISTR	65.5	895	**54.9**	15.6	24.4
WISTR+SPIDER	67.0	489	37.9	16.4	22.9
TopoCluster	57.0	604	31.8	25.9	28.6
TopoClusterGaz	**71.5**	**468**	37.7	**30.7**	**33.8**

Table 5: Doc Geolocation Results

System	Acc@161km	Median	Mean
Random/Uniform	3.4	1009.5	865.6
Random/KD	8.3	828.8	753.2
NaiveBayes/Uniform	74.8	194.7	53.1
NaiveBayes/KD	72.2	204.4	80.2
LR/Uniform	**77.2**	189.8	53.6
LR/KD	74.4	182.1	59.8
Hier/Uniform	76.8	185.5	49.6
Hier/KD	76.2	**171.8**	**47.2**

trained using out of domain resources, but some weights and parameters (e.g. context window size) were optimized using the WOTR dev set.

Table 5 shows the results of a number of current text-only document geolocation systems (Wing, 2015) on WOTR. Compared with Naive Bayes, both flat (LR) and hierarchical logistic regression (Hier) produce additional benefits. Hier produces the best mean and median despite the fact that it is designed primarily for larger corpora than WOTR. Uniform grids do slightly better overall, a result we have seen before in similar-sized corpora, but adaptive (KD) grids do better with Hier, which is able to compensate somewhat for the larger adaptive grid cells found in low-density areas through its use of multiple grid levels.

Table 4 shows the resolution results of many state-of-the-art toponym resolution systems on the test split of WOTR. As can be seen, TopoClusterGaz outperforms all resolvers on all metrics when oracle NER is used, and outperforms others on Recall and F-1 Score when predictive NER is included in the evaluation. Key to the TopoClusterGaz's success is the ability to predict on both non-gazetteer and gazetteer matched entities, directly boosting Recall and F-1 Score by large margins. When evaluating on a development set of the data, we observed that most differences in system performance could be sourced to how the respective systems dealt with place names that do not have specific GeoNames entries, or are spelled differently than their GeoNames entry (e.g. *Camp Lapwai*, *Colo. Terr.*). TopoCluster often produced correct predictions on these entities, while the gazetteer dependent systems like Population, WISTR, and SPIDER were unable to make predictions. NER inclusive scores (P, R, F-1) are generally much lower for WoTR-Topo than other datasets because the NER systems utilized (Stanford-NER and openNLP-NER) are trained on very different domains. Nevertheless, strongly superior recall on the gazetteer-independent TopoCluster systems leads to higher F-1 scores on the dataset.

6 Conclusion

The War of the Rebellion corpus represents a unique domain for geolocation research. From the perspective of toponym resolution, the corpus is innovative in many respects: richness of geometric annotation (annotations with multi-point, polygon geometries), corpus size (with roughly twice the toponyms of other corpora), and place names not in gazetteers. Baseline system resolution results indicate that the corpus is the most difficult of the corpora surveyed, with A@161 km scores–and especially NER-inclusive scores–being significantly lower than the next most difficult corpus, LGL (DeLozier et al., 2015). The corpus is the first published document-geolocation corpus focusing on historical texts, the first based on running text, the first that was annotated specifically for the task of theme-based document geolocation, and the first annotated with multi-point and polygon geometries. Finally, the availability of text marked both with toponym and docgeo annotations presents new opportunities for joint inference.

7 Corpus availability

The corpus is freely available at our github page[6] under an MIT License. We hope others may expand and improve on the annotations.

8 Acknowledgements

We would like to thank David Staley and Ohio State University's Department of History for access to their high quality version of the War of the Rebellion corpus. This research was supported by a grant from the Morris Memorial Trust Fund of the New York Community Trust.

[6] https://github.com/utcompling/WarOfTheRebellion

References

Edward L. Ayers and Scott Nesbit. 2011. Seeing emancipation: Scale and freedom in the american south. *Journal of the Civil War Era*, 1(1):3–24.

Paolo Bolettieri, Andrea Esuli, Fabrizio Falchi, Claudio Lucchese, Raffaele Perego, Tommaso Piccioli, and Fausto Rabitti. 2009. CoPhIR: a test collection for content-based image retrieval. *CoRR*, abs/0905.4627.

Mariam Daoud and Jimmy Xiangji Huang. 2013. Mining query-driven contexts for geographic and temporal search. *International Journal of Geographical Information Science*, 27(8):1530–1549.

Chris De Rouck, Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. 2011. Georeferencing wikipedia pages using language models from flickr. In *Semantic Web, 10th International conference, Proceedings*, page 8.

Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 141–150. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889.

Bo Han, Paul Cook, and Tim Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49(1):451–500.

Saul A. Kripke. 1980. *Naming and Necessity*. Harvard University Press.

Jochen L Leidner. 2008. *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal Press, Boca Raton, FL, USA.

Michael D Lieberman and Hanan Samet. 2012. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 731–740. ACM.

Koji Matsuda, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. 2015. Annotating geographical entities on microblog text. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 85.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Scott Nesbit. 2013. Visualizing emancipation: Mapping the end of slavery in the american civil war. In Justyna Zander and Pieter J. Mosterman, editors, *Computation for Humanity: Information Technology to Advance Society*, pages 427–435. New York: Taylor & Francis.

Neil O'Hare and Vanessa Murdock. 2013. Modeling locations with social media. *Information Retrieval*, 16(1):30–62.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1500–1510, Stroudsburg, PA, USA. Association for Computational Linguistics.

João Santos, Ivo Anastácio, and Bruno Martins. 2014. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, pages 1–18.

Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. 2013. A multi-indicator approach for geolocalization of tweets. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM'13: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

David A Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, pages 127–136. Springer.

Michael Speriosu and Jason Baldridge. 2013. Text-driven toponym resolution using indirect supervision. In *ACL (1)*, pages 1466–1476.

Yael A. Sternhell. 2016. Afterlives of a confederate archive: Civil war documents and the making of sectional reconciliation. *Journal of American History*, 102(4):1025–1050.

William G. Thomas III. 2011. *The Iron Way: Railroads, the Civil War, and the Making of Modern America*. Yale University Press.

Olivier van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt, and Christopher B. Jones. 2014. Georeferencing wikipedia documents using data from social media sources. *ACM Trans. Inf. Syst.*, 32(3):12:1–12:32, July.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 336–348, Doha, Qatar, October. Association for Computational Linguistics.

Benjamin Wing. 2015. *Text-Based Document Geolocation and its Application to the Digital Humanities*. Ph.D. thesis, University of Texas at Austin.