The Dissertation committee for Benjamin Patai Wing

Certifies that this is the approved version of the following dissertation:

# Text-Based Document Geolocation and its Application to the Digital Humanities

**APPROVED BY**

**Dissertation Committee:**

**Supervisor:** ______________________________
Jason Baldridge

______________________________
Katrin Erk

______________________________
David Beaver

______________________________
Ray Mooney

______________________________
Matt Lease

# Text-Based Document Geolocation and its Application to the Digital Humanities

by

**Benjamin Patai Wing, A.B.; B.A.; M.A.; M.S.C.S.**

**Dissertation**

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

December 2015

To my sister Jessica Grace Wing (1971-2003), may she live on in our hearts forever.

# Acknowledgments

This dissertation could not have been completed without the generous help of my advisor, Jason Baldridge, as well as the constant encouragement and support of Jennifer Schneider, William Wing, Jacqueline Sharkey, Tamra Zehner, Diandra Ford and Grant DeLozier. Heartfelt thanks also go to the remaining committee members (David Beaver, Katrin Erk, Matt Lease, Ray Mooney) and to Nora England and Ben Rapstine of the Department of Linguistics.

BENJAMIN PATAI WING

*The University of Texas at Austin*

*December 2015*

# Text-Based Document Geolocation and its Application to the Digital Humanities

by

Publication No. ______________________

Benjamin Patai Wing, Ph.D.
The University of Texas at Austin, 2015

SUPERVISOR: Jason Baldridge

This dissertation investigates automatic *geolocation* of documents (i.e. identification of their location, expressed as latitude/longitude coordinates), based on the text of those documents rather than metadata. I assert that such geolocation can be performed using text alone, at a sufficient accuracy for use in real-world applications. Although in some corpora metadata is found in abundance (e.g. home location, time zone, friends, followers, etc. in Twitter), it is lacking in others, such as many corpora of primary-source documents in the digital humanities, an area to which document geolocation has hardly been applied. To this end, I first develop methods for accurate text-based geolocation and then apply them to newly-annotated corpora in the digital humanities. The geolocation methods I develop use both uniform and adaptive (*k*-d tree) grids over the Earth's surface, culminating in a hierarchical logistic-regression-based technique that achieves state of the art results on well-known corpora (Twitter user feeds, Wikipedia articles and Flickr image tags).

In the second part of the dissertation I develop a new NLP task, text-based geolocation of historical corpora. Because there are no existing corpora to test on, I create and annotate two new corpora of significantly different natures (a 19th-century travel log and a large set of Civil War archives). I show how my methods produce good geolocation accuracy even given the relatively small amount of annotated data available, which can be further improved using domain adaptation. I then use the predictions on the much larger unannotated portion of the Civil War archives to generate and analyze geographic topic models, showing how they can be mined to produce interesting revelations concerning various Civil War-related subjects. Finally, I develop a new geolocation technique for text-only corpora involving co-training between document-geolocation and toponym-resolution models, using a gazetteer to inject additional information into the training process. To evaluate this technique I develop a new metric, the closest toponym error distance, on which I show improvements compared with a baseline geolocator.

# Contents

# Chapter 1

# Introduction

Georeferencing, the relation between information and geographic location, is an important component of textual understanding (Hill, 2006). Georeferencing is extensively used in knowledge organization, and geographic references are found throughout general conversion and writing (Buchel and Hill, 2009). It has been estimated that 50-80% of textual documents contain geographic references (Petras, 2004; Ridley et al., 2005). Automated georeferencing has become increasingly important in day-to-day life through the ubiquity of location-based services in many components of smart phones.

This dissertation focuses on one georeferencing task, that of *document geolocation*, which locates an abstract document (a stretch of text, which may in reality be a paragraph, article, chapter, etc.) in geographic space. In other words, it identifies the location that forms the primary focus of each document—for example, identifying an article whose focus is on Austin, Texas with a latitude/longitude coordinate that represents the city (perhaps its geographic center or downtown area). This can be thought of as one way to summarize the geographic content of the document.

Figure 1.1 shows the kind of summary possible using document geolocation, plotting on a per-chapter level the paths followed in John Beadle's *Western Wilds, and the Men Who Redeem Them*, published in 1878 (§2.3.1). Each location corresponds to a paragraph and is labeled by the chapter in which it occurs (using a Roman numeral). Lines are drawn connecting adjacent para-

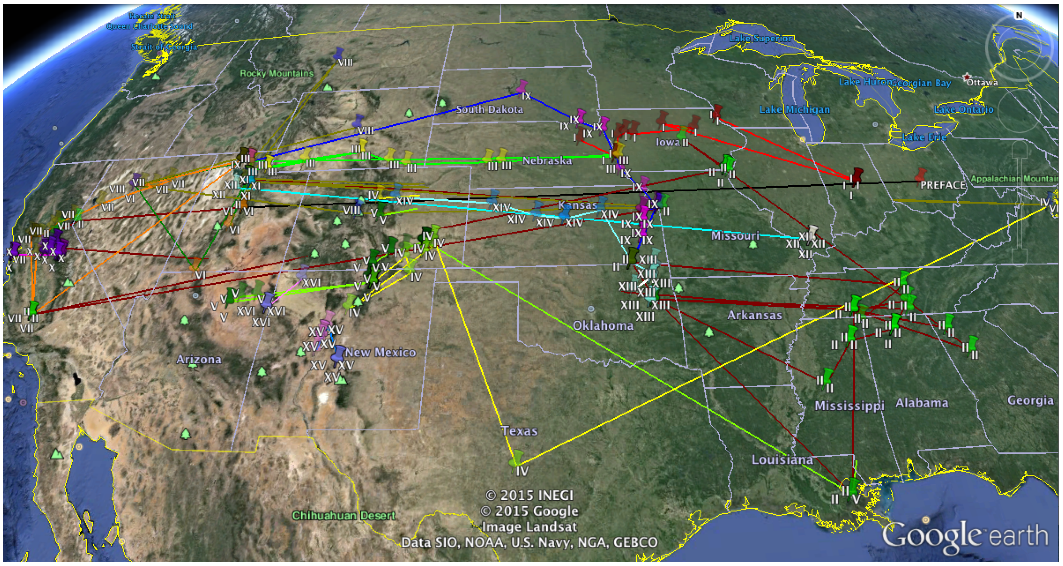


Figure 1.1: Plot of locations in Beadle's *Western Wilds* labeled by chapter, with lines connecting locations for adjacent paragraphs within a given chapter.

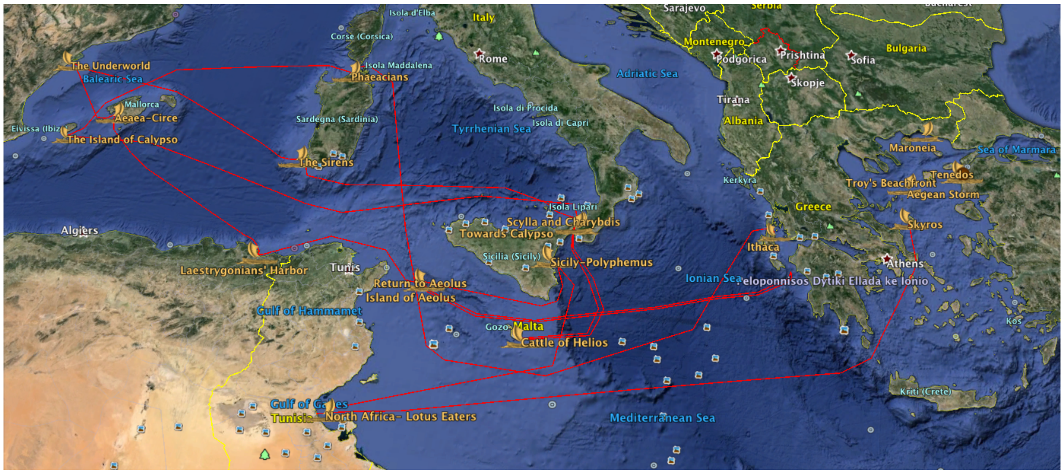


Figure 1.2: Plot of one interpretation of the paths and locations followed in Homer's *Odyssey*, from Google Lit Trips.

graphs within a given chapter, with different colors for each chapter. This map is similar to the maps produced by the Google Lit Trips project[1], as shown in Figure 1.2, which plots an interpretation of the locations found in Homer's *Odyssey*. Both are great teaching tools and serve as points of reference for further discussion of entire historical worlds. The difference between the two is that the locations in the Odyssey map are subject to a great deal of interpretation and thus the map can only be drawn by hand, but the map of *Western Wilds* could potentially be drawn automatically using document geolocation, particularly in the presence of a carefully tailed sequence model (§7.5).

This dissertation focuses on document geolocation, in particular using a document's text rather than its metadata.[2] The motivation for this, including the core theses of this dissertation, is covered in §1.1. The structure of a text-based geolocation system is described in §1.2, and §1.4 discusses previous work in document geolocation. Further discussion of the applications of geolocation in general is found in §1.5, and applications to the digital humanities, a core component of the second half of this dissertation, are discussed in §1.6. Finally, §1.7 presents an outline of the rest of this dissertation.

## 1.1 Core theses

The primary thesis of this dissertation is that **geolocation can be performed with accuracy sufficient for useful real-world applications by using only the text of a document, even without any available metadata associated with the document**. The *metadata* of a document is any information associated with the document other than its raw text, e.g. hyperlinks to other documents, the social media profile of the author of the document in a social network, a user's self-declared location, the set of other users connected to such an author through a friend or follower relationship, etc.

Metadata-based approaches can achieve great accuracy,[3] but are very specific to the partic-

[1] `http://www.googlelittrips.com/GoogleLit/Home.html`

[2] This dissertation is partly based upon Wing (2011); Wing and Baldridge (2011); Roller, Speriosu, Rallapalli, Wing and Baldridge (2012); and Wing and Baldridge (2014).

[3] For example, Schulz et al. (2013) obtain 79% accuracy within 100 miles for a US-based Twitter corpus, compared with 49% using my methods on a comparable corpus.

ular corpus and the types of metadata it makes available. Twitter,[4] for example, includes a great deal of metadata with its tweets, most of which is unavailable e.g. for Wikipedia documents.[5] In some domains, such as the digital humanities, documents are typically pure text, lacking any metadata. Text-based approaches can be applied to all types of corpora; metadata can be additionally incorporated when available.

In the first part of this dissertation I present numerous methods for supervised text-based geolocation, including novel ones I have developed. These methods divide the Earth into a grid of cells, each covering a particular region, using either a uniform or adaptive (*k*-d tree) grid (Roller, Speriosu, Rallapalli, Wing and Baldridge, 2012). I then treat geolocation as a classification problem. In this case, the number of cells is often very large. This leads to another thesis of this dissertation, that **geolocation as a classification problem can be successfully solved using discriminative methods despite having thousands or tens of thousands of classes**. I do this using a *hierarchical classification* method that I introduce, which achieves state-of-the-art results in text-based classification.

In the second part of this dissertation I investigate a particular use case of my methods, text-only historical corpora in the digital humanities (§1.6). By hand-annotating part of some digital humanities texts, I show that, **if even a fraction of the paragraphs of a document can be annotated with geographic coordinates, sufficient accuracy can be achieved to facilitate interesting real-world applications**. I hand-annotated part of a 19th-century travel log and supervised the annotation of a portion of a major primary-source archive of Civil War documents (§2.3). I train models on this annotated data and use these models to label the remainder of the data, which allows for applications relevant to the digital humanities, such as geographic topic models (§5.5.1).

One way to increase the accuracy of predictions given a fairly small set of labeled data is through *domain adaptation* (Daumé III, 2007; Chen et al., 2011; Daumé III et al., 2010), incorporating a model trained on a different (*out-of-domain*) corpus for which abundant labeled data is available (in my case, the English Wikipedia) in addition to or instead of the smaller amount of *in-domain* labeled data available. My experiments show that purely using Wikipedia with no in-domain labeled data performs surprisingly well. **Judiciously combining the two sources of labeled data**

[4] http://www.twitter.com/

[5] http://www.wikipedia.org/

**can produce *learning curves* (curves showing performance over differing amounts of training data) that significantly outperform only in-domain data** when only a small amount of in-domain data is available, and at least do no worse when a larger amount is available. This produces a curve that is flatter and with greater averaged accuracy over differing amounts of training data. Such curves can be computed dynamically while annotation is taking place to determine how much data needs to be annotated.

Another avenue I explore is **using document-geolocation techniques to inform and improve upon toponym resolution**, expanding upon the work of Speriosu (2013). Among other things, I demonstrate a means of using co-training to simultaneously train a document geolocator and a toponym resolver on a combination of document-geolocated Wikipedia text and toponym-resolved Civil War text. The potential advantages of this method are great; as in other methods for joint inference, the two knowledge sources can inform each other, each with their own separate training data and each one potentially providing separate sets of constraints. (For example, some of Speriosu's toponym resolution methods, such as SPIDER (§6.2), do joint inference over toponyms.) I produce improvements on a metric that generates a document-geolocation-inspired error distance from a set of toponym resolutions. In addition, I develop some new toponym-resolution techniques, variants on existing techniques which incorporate feedback from document-level annotations when it is available. It is hoped that this work will serve as a springboard for further research in the combination of document geolocation and toponym resolution.

## 1.2 Structure of a text-based geolocation system

Geocoordinates are real-valued latitude/longitude pairs, and thus theoretically geolocation should be treated as a regression problem. However, the function that needs to be predicted is highly complex and irregular, and for this reason it is common to discretize the Earth's surface into grid cells, which allows the task to be treated as a classification problem. This is the approach followed in this dissertation, and it allows for geolocation techniques that fall under the rubric of *language modeling approaches in information retrieval* (Ponte and Croft, 1998; Manning et al., 2008). The general strategy is to associate each training set document with the discrete grid cell that contains

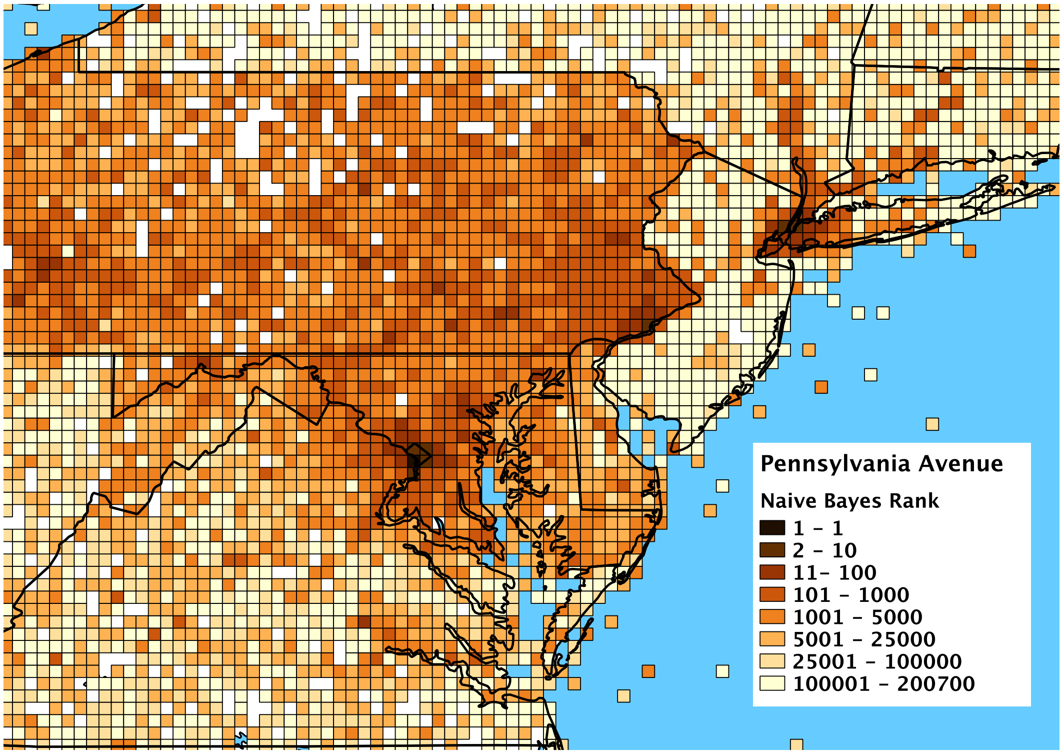


Figure 1.3: Ranking of a test document in a uniform 0.1° grid. Relative Naive Bayes rank is shown for cells for test document *Pennsylvania Avenue (Washington, DC)* in ENWIKI13 (§2.2.2), surrounding the true location. (Constructed with assistance from Grant DeLozier.)

it. The set of training documents associated with a cell is concatenated into a *pseudo-document*, and the *language model* of this pseudo-document (i.e. the statistical description of the distribution of words in the document) computed. The language model of the test document is likewise computed, and some method (e.g. Naive Bayes) is then used to compare the two, generating a score. The centroid of the highest-scoring cell is then chosen as the predicted location of the test document. Figure 1.3 shows an example of this process for a document in the English Wikipedia that describes Pennsylvania Avenue in Washington, DC.

The location that a document geolocation system associates with a document can be specified at various levels of resolution, for example an exact point in space, an address, a city, or a higher-level *administrative region* (e.g. a state, province or country), and can be identified by name (i.e. in the form of a *toponym*), by latitude/longitude coordinate, or by a polygon of such coordinates. Such a polygon can be either in the form of a *bounding box* (a rectangle in latitude/longitude space, identified by the coordinates of two opposite corners) or a more complex shape. It should be understood that the identification of a location by coordinates does not necessarily indicate that the level of resolution of this location should be taken as an exact point. For example, a document whose focus is on a state or other higher-level administrative region may nonetheless be *geotagged* in a corpus with a point coordinate, typically chosen as either the geographic center, population centroid, or location of the seat of government. In this dissertation, I typically identify locations by point coordinates, but it should be understood that this can be problematic for large administrative regions (e.g. treating the United States as a point in Kansas or Washington, D.C.). An alternative, suggested by Speriosu (2013), is to treat administrative regions as sets of representative points derived from a *gazetteer* (a list of named locations and associated coordinates), spanning the entire region.

Document geolocation makes the assumption that a document can be adequately associated with a *single* location—akin to the *one sense per discourse* assumption commonly used in word sense disambiguation (Yarowsky, 1995). This is only a well-posed problem for certain documents, generally of fairly small size. Nonetheless, there are many natural situations in which such collections arise. For example, a great number of articles in Wikipedia have been manually geotagged with a coordinate; this allows those articles to appear in their geographic locations while geobrowsing in an application like Google Earth. Images in social networks such as Flickr[6] may be geotagged by a camera and their textual tags can be treated as documents. Likewise, tweets in Twitter are often geotagged, particularly when sent from a mobile phone in which the geotagging feature has been enabled. In this case, it is possible to view either an individual tweet or the collection of tweets for a given user as a document, respectively identifying the document's location as the place from which the tweet was sent or the home location of the user (assumed to be nearby the tweets sent by that user). In the case of a long document, e.g. a book, the treatment normally used in this dissertation is to break up the document into chunks, either at the paragraph level or a fixed number of sentences in length (e.g. 10 or 20).

The document geolocation methods used in this dissertation break the Earth into a *geodesic grid*, i.e. a grid of cells that tile the Earth into regions. This allows geolocation to be treated as a type of classification problem, as discussed in §1.1. An algorithm is used to provide a *ranking* of the cells based on the document being geolocated, and the top-ranked cell is used to predict the document's location. In particular, the predicted location is the *centroid* of the training documents in the cell, where the centroid is computed by separately averaging the latitudes and longitudes of the training documents. (This was shown by Roller et al. (2012) to be superior to using the geographic center, as was done in Wing and Baldridge (2011), because it better handles the situation where the training documents—and hence, by assumption the test document—are bunched near one of the edges of the cell in question.)

The ranking of a cell is determined by scoring each cell separately, comparing the language model of the test document with the language model of the concatenation of the training documents in a given cell. Simple unigram (bag-of-words) language models are generally used. (Experiments using bigram models performed by me and by Han et al. (2014), yielded little gain, but van Laere et al. (2014) was able to demonstrate improvements.) The technique used to compute the ranking may be Naive Bayes (Lewis, 1998), Kullback-Leibler (KL) divergence (Zhai and Lafferty, 2001), logistic regression (Hosmer Jr. and Lemeshow, 2004), a hierarchical classifier (Silla Jr. and Freitas, 2011), or other strategies.

In some situations, the actual scores of cells other than the top-ranked are used. For exam-

[6] http://www.flickr.com/

ple, an alternative to simply using the centroid of the top-ranked cell as the predicted cell is to use the *mean shift* algorithm. The motivation is that the cells near the top in rank may tend to cluster in a particular region whereas the top-ranked cell may happen to be located elsewhere; in this case the weight of evidence may be taken to be in favor of the cluster. (However, experiments I performed using mean shift did not produce improvements.) More generally, the set of scores of cells can be viewed as a probability distribution over the cells—an approximation to a continuous probability distribution over the Earth. Some applications, for example some of the toponym-resolution techniques discussed in §6.2, make use of this entire distribution.

The layout of the grid is an important component of the document-geolocation process. Perhaps the simplest layout is a *uniform grid* of cells, each of which forms a square in latitude-longitude space (for example, having 1° per side). Such squares are not all of equal area, because longitude lines move closer together as one moves away from the Equator towards either pole. Equal-area alternatives, such as the *quaternary triangular mesh*, have been considered (Dias et al., 2012). A different issue with uniform grids is that they over-represent rural areas at the expense of urban areas, which I handle through the use of an adaptive *k-d grid* (§3.2.2). Another alternative is to directly use a grid constructed from a gazetteer of cities (Han et al., 2014).

Measurement of the accuracy of a particular geolocation prediction can be done in various ways. A simple possibility, used by some researchers, is *cell accuracy*, i.e. the fraction of correctly predicted cells. However, this has the disadvantage that the metric cannot easily be used to compare different grid layouts, particularly with different-sized cells. I use alternative metrics based on *error distance* (the distance between the predicted and correct location), namely *mean* and *median*. I also use *acc@161*, the fraction of documents where the error distance is within 161 km (100 miles), from Cheng et al. (2010) and approximating the concept of "within the same metro area".

It is possible to do away entirely with a grid and directly model the continuous nature of the Earth's surface. For example, Eisenstein et al. (2010) use Gaussian distributions and variational Bayes methods to model the locations of Twitter users in the United States of America. However, there are two problems with this. One is that Gaussian distributions are unsuited to modeling spherical surfaces, which have no natural endpoints; instead, spherical distributions like the von

Mises-Fisher distribution (Dhillon and Sra, 2003) would need to be employed. More fundamentally, however, this and related works (Ahmed et al., 2013; Hong et al., 2012; Eisenstein et al., 2011b) have been tested only on quite small corpora, and there are serious questions as to whether the methods can be scaled to large corpora of the sort I consider in this dissertation. Grid-based models, on the other hand, are simpler to construct and are known to have good performance over large datasets.

## 1.3 Toponym identification, toponym resolution and document geolocation

Closely related to document geolocation are *toponym identification* and *toponym resolution*. All three fall under the general rubric of georeferencing. It is important to understand how document geolocation differs from the other two, and why it is not sufficient to simply identify and resolve the toponyms in a text.

Toponym identification is the extraction of place names in text, such as identifying the place name “Springfield” and determining that it is, in fact, a place name rather than e.g. a personal name. This is normally viewed as a subtask of *named entity recognition* (NER). Toponym resolution (Leidner, 2008; Yuan, 2010; Speriosu, 2013; Rupp et al., 2013) is the attachment of coordinates to place names, including the resolution of ambiguous place names, such as disambiguating the textual mention “Springfield” to the city of Springfield, Massachusetts.

Document geolocation is clearly related to toponym resolution in that both seek to resolve text to a location, but the scope is quite different. Toponym resolution often occurs with the aid of a *gazetteer*, which lists, for each ambiguous toponym, the possible locations that it can be mapped to. The resolution step then boils down to a choice among fixed alternatives, making use of various sorts of context information such as the surrounding text or the possible or actual identities of other, nearby toponyms. (Muddying the waters somewhat is the fact that it is possible to do toponym resolution without a gazetteer, making use of document-level geolocation annotations, such as in DeLozier et al. (2015).)

Document geolocation is much less constrained, in that it seeks to identify the location that forms the primary focus of an entire document. That document may well contain toponyms, which may be a strong hint as to the location of the document. Thus, toponym resolution could serve as an ancillary component of a document geolocation system. However, toponym resolution by itself is insufficient for document geolocation, both because there are often non-toponym words that are highly geographically indicative and because some or all of the toponyms in a text may not be near the actual location of the document.

The relation between them can be seen in the following text from John Beadle, *Western Wilds* (§2.3.1):

> From this region goes most of the lumber used along the road, as far as **Salt Lake City**; but over all that interior there is an ever increasing scarcity of good timber. Woods are found only upon the mountains; the inner plains of the **Great Basin** are as bare of trees as if blasted by the breath of a volcano. At **Verdi Station**, 5,000 feet above sea-level, we pass the State line and enter **California**. Crossing the **Truckee**, we take an additional locomotive and enter upon the steepest ascent of the **Sierras**. The first large curve brings us above **Donner Lake**, so named in memory of those unfortunate emigrants from **Quincy, Illinois**, who here starved and froze and suffered away the long cold winter of 1846. Next we look down upon **Lake Bigler**, and another hour brings us to **Summit Station**, highest point on the Central Pacific, 7,042 feet above sea-level, 1,669 miles from **Omaha**, and 105 from **Sacramento**.

This text is discussing an area in the Sierra Nevada mountains of eastern California, along the Central Pacific railroad, but includes toponyms corresponding to several locations in the United States, some of which are directly relevant to the subject at hand and some of which are not, and some of which are ambiguous (e.g. “Truckee” and “Sacramento” are both towns and rivers, and there are in fact several places in the United States named “Omaha” and “Sacramento”).

Toponym resolution rather than whole-document geolocation is more common in real-world georeferencing tools, perhaps because document-level geolocation isn’t always a well-defined task for a given document, even when divided up into small chunks such as paragraphs. However, to the extent that it is applicable, **I assert that document geolocation is more useful than toponym**

**resolution because of its ability to summarize the whole topic of the document (or document chunk) in one location.** Examples of these summaries are found elsewhere in this dissertation; see, for example, Figure 1.1 and Figure 5.11.

## 1.4 Previous work

Early work on document geolocation used heuristic algorithms, predicting locations based on toponyms in the text (Ding et al., 2000; Smith and Crane, 2001). More recently, various researchers have used topic models for document geolocation (Ahmed et al., 2013; Hong et al., 2012; Eisenstein et al., 2011b; Eisenstein et al., 2010) or other types of geographic document summarization (Mehrotra et al., 2013; Adams and Janowicz, 2012; Hao et al., 2010). More recent work in document and/or user geolocation tends to make use of either the text of the document in the form of a language model—as this dissertation does—or metadata of various sorts, such as document links and social network connections. This research has sometimes been applied to Wikipedia (Overell, 2009; van Laere et al., 2013), Facebook (Backstrom et al., 2010) or Flickr (Serdyukov et al., 2009; O'Hare and Murdock, 2013), but more commonly to Twitter (see below). Some work involving domain adaptation has been done, such as applying data from Twitter to Flickr estimation (C. Hauff, 2012) and data from both Twitter and Flickr to Wikipedia (van Laere et al., 2014).

Much work on Twitter makes use of the extensive metadata provided with tweets and users, focusing on features such as time zone (Mahmud et al., 2012), declared location (Hecht et al., 2011), language identification (Graham et al., 2014), or a combination of these (Schulz et al., 2013). A recent, fruitful area of research has been the creation of *network-based* models that make use of either the friends and followers (Compton et al., 2014; McGee et al., 2013; Sadilek et al., 2012) or the unidirectional or reciprocal @-mentions in tweets, i.e. cases where one user refers to another user in a tweet (Jurgens, 2013; Rahimi et al., 2015b). This makes the assumption that the *ego network* of users directly connected to a given user tend to be located nearby that user, an assumption that has been demonstrated in the case of mutual friend/follower relationships (Takhteyev et al., 2011; McGee et al., 2011). Using this assumption, a global distribution of locations can be computed using techniques such as label propagation (Talukdar and Crammer, 2009) and total variation minimization

(Rudin et al., 1992).

The primary alternative line of research, followed by this dissertation, focuses on text-based geolocation using language models. The overall structure of such a system is described in detail in §1.2. Earlier models (Wing and Baldridge, 2011; Serdyukov et al., 2009; O'Hare and Murdock, 2013) used Naive Bayes models over a uniform grid, which was then extended to an adaptive *k*-d grid (Roller, Speriosu, Rallapalli, Wing and Baldridge, 2012) and to the use of logistic regression (Wing and Baldridge, 2014; Han et al., 2014), and further to hierarchical logistic regression (Wing and Baldridge, 2014). An additional area of research has been the use of smoothing of neighboring areas to increase geolocation accuracy, such as through kernel density methods (Hulden et al., 2015; Lichman and Smyth, 2014; Thom et al., 2012). van Laere et al. (2013) proposed a two-step process for geolocating Flickr images in which a language-model-based approach is followed by a similarity search within a given grid cell, greatly improving the accuracy. Current research involves the combination of network-based and language-model-based methods (Rahimi et al., 2015a).

## 1.5 Applications of document geolocation

Document-level geolocation has numerous applications. It is a critical component of *location-based services*, which are concerned more generally with locating users of cell phones, social networks, etc. and directing location-specific content to them, for example navigational directions; recommendations for local social events or restaurants (Quercia et al., 2010); directions to nearest businesses of various types; alerts of traffic, adverse weather, or local sales; person-to-person location services; and targeted advertisements. Other applications are possible as well given the ability to locate the source of a document in geographic space, e.g. trend detection of epidemic dispersion (Lampos et al., 2010; Paul and Dredze, 2011), earthquake prediction (Sakaki et al., 2010), or election forecasting (Tumasjan et al., 2010).

Numerous applications of document-level geolocation focus specifically on the text of the document in question. One major issue is the grounding of word meaning and language usage in geography. For example, distributions of words in geographic space can be generated using a method similar to the Average Cell Probability (ACP) inference algorithm (§3.4.4), as in Figure 1.4

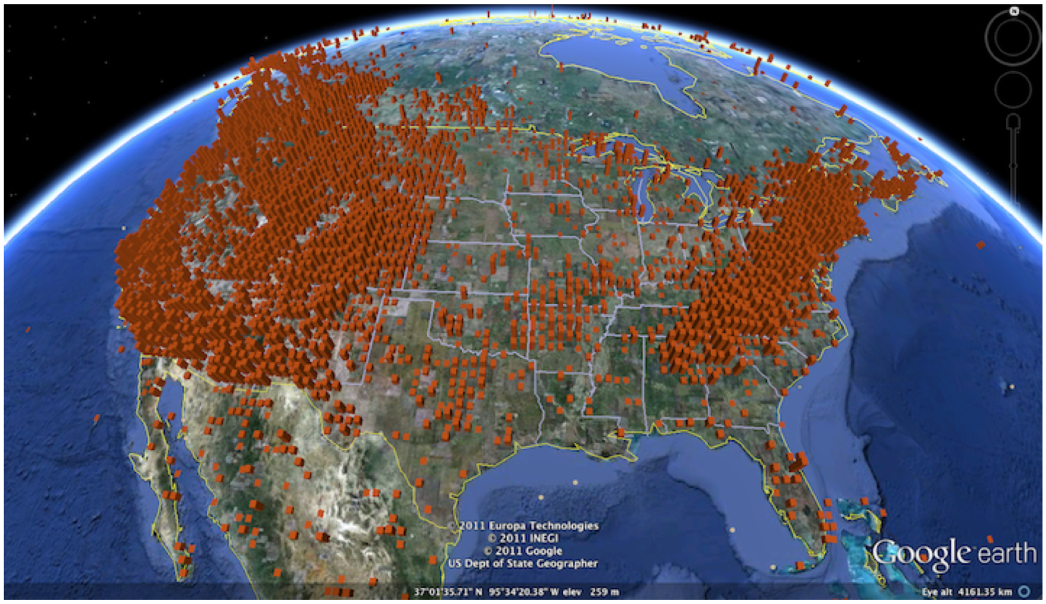
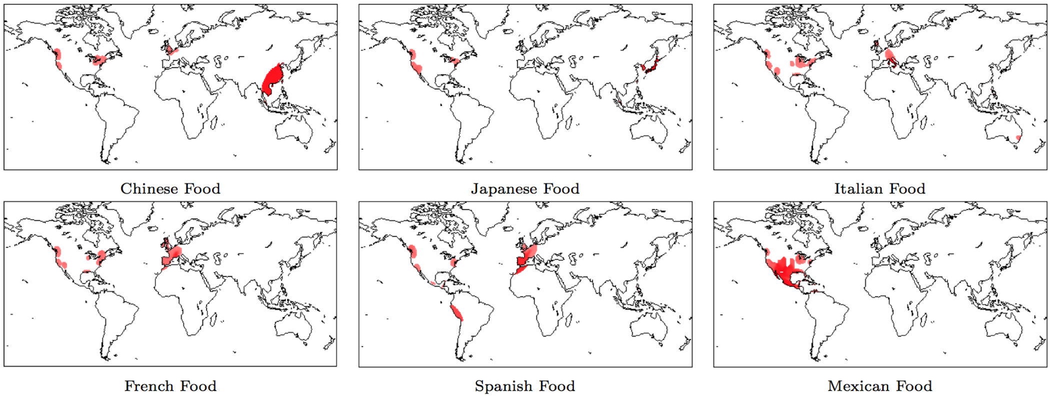


Figure 1.4: Wikipedia distribution of *mountain* in geotagged articles, plotted using Google Earth.

| Chinese Food | Japanese Food | Italian Food | French Food | Spanish Food | Mexican Food |
|---|---|---|---|---|---|
| chinese 0.552 | japanese 0.519 | italian 0.848 | french 0.564 | spanish 0.488 | mexican 0.484 |
| noodles 0.067 | ramen 0.104 | cappuccino 0.067 | bistro 0.070 | tapas 0.269 | tacos 0.069 |
| dimsum 0.064 | soba 0.066 | latte 0.048 | patisserie 0.056 | paella 0.076 | taco 0.059 |
| hotpot 0.039 | noodle 0.065 | gelato 0.030 | bakery 0.049 | pescado 0.059 | salsa 0.036 |
| rice 0.038 | sashimi 0.039 | pizza 0.002 | resto 0.044 | olives 0.032 | cajun 0.031 |
| noodle 0.035 | yakitori 0.030 | pizzeria 0.002 | pastry 0.033 | stickyrice 0.017 | burrito 0.027 |
| tofu 0.020 | okonomiyaki 0.026 | mozzarella 0.001 | tarte 0.026 | tortilla 0.013 | crawfish 0.023 |
| dumpling 0.018 | udon 0.026 | pasta 0.001 | croissant 0.021 | mediterranean 0.010 | guacamole 0.022 |
| duck 0.018 | tempura 0.020 | ravioli 0.000 | baguette 0.019 | mussels 0.008 | margarita 0.020 |
| prawn 0.017 | curry 0.016 | pesto 0.000 | mediterranean 0.018 | octopus 0.008 | cocktails 0.020 |


Chinese Food

Japanese Food

Italian Food

French Food

Spanish Food

Mexican Food

Figure 1.5: Geographic topics found in a food dataset based on geotagged, term-tagged images, from Yin et al., 2011.

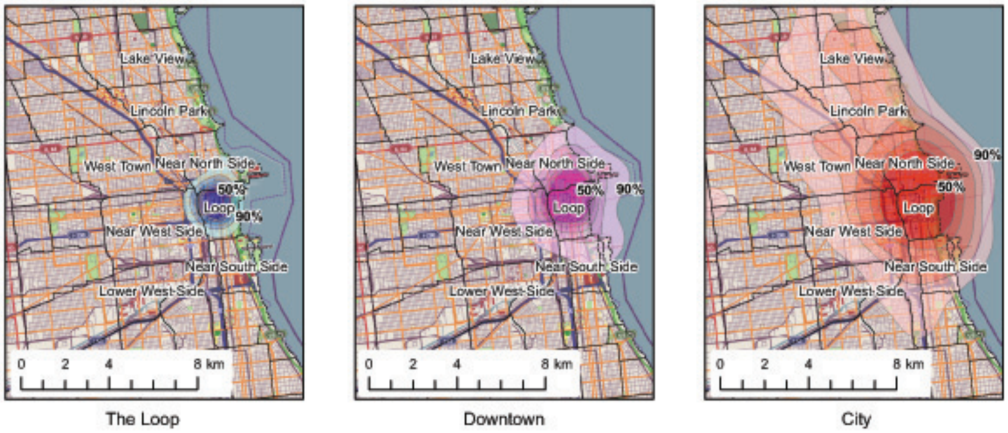


Figure 1.6: Usage of vague geographic terms to refer to areas of Chicago, based on geotagged, term-tagged images from Flickr. From Hollenstein and Purves, 2010.

(Baldridge et al., 2012). These can be viewed as representations of word meaning complementary to the context-based vector space models of word meaning in distributional semantics (Erk, 2013). Topic models can also be adapted to geographic space, as in Figure 1.5 (Yin et al., 2011).

Geolocation, especially of social media, can also serve in sociological studies of word meaning and usage (Eisenstein et al., 2011a). Examples are the prevalence of different second languages across a metropolitan area (Mocanu et al., 2013) (Figure 1.7) or the extent of use of vague geographic terms such as "downtown" to refer to particular neighborhoods in a city (Hollenstein and Purves, 2010) (Figure 1.6).

Only about 2.02% of tweets are geotagged with a location, either a city or neighborhood chosen from a list (1.8%) or exact latitude-longitude coordinates (1.6%); these numbers do not add up to the total of 2.02% because many tweets have both types of geotags (Leetaru et al., 2013). Only 8.2% of all users in the period studied by Leetaru et al. produced any geotagged tweets, with over half sending only one geotagged tweet. This suggests that there is a great deal of room for automatic geolocation techniques that make use of other information.

As an example, Leetaru et al. describe a fairly simple algorithm to deduce a location from the free-form user-declared location in users' profiles, along with related profile information. They claim this allows 34% of tweets to be geolocated, although only at the level of a user rather than

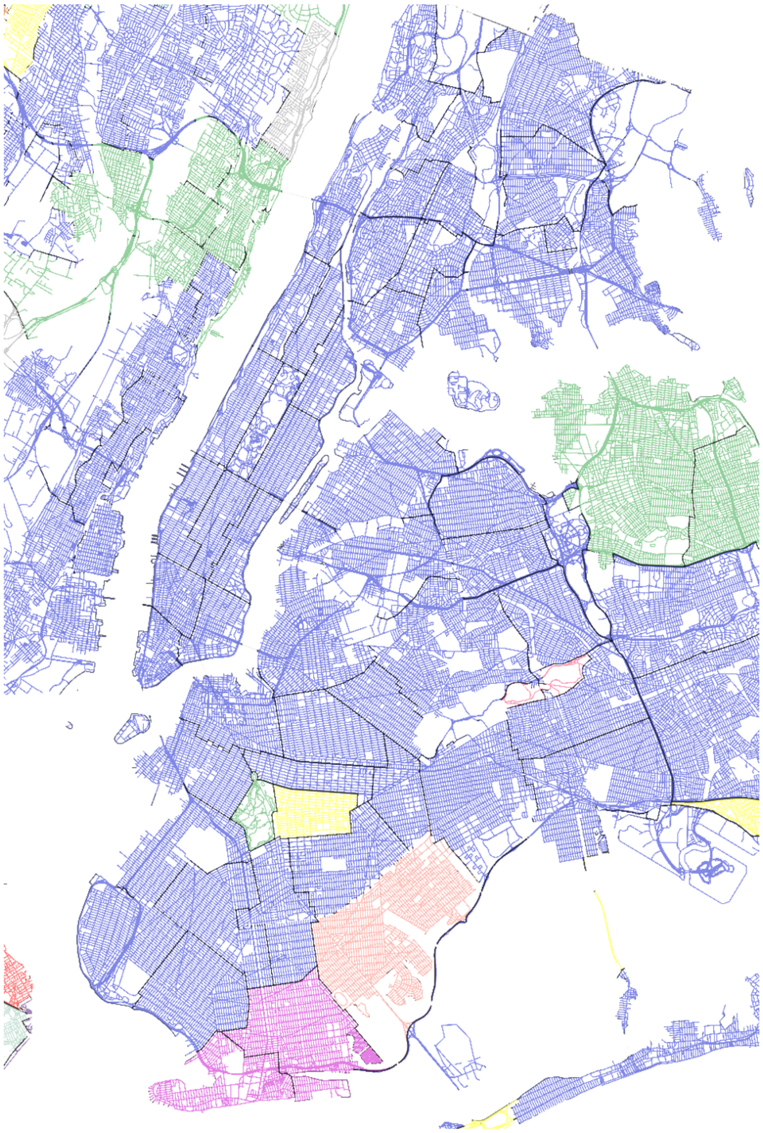


Figure 1.7: Prevalence of second languages in the New York metropolitan area, based on geotagged tweets. Blue = Spanish, Light Green = Korean, Fuchsia = Russian, Red = Portuguese, Yellow = Japanese, Pink = Dutch, Grey = Danish, Coral = Indonesian. From Mocanu et al., 2013.

an individual tweet. In practice, of course, there will be some loss of both precision and accuracy in such a technique compared with annotated tweet-level geotags, especially exact coordinates, although in this case the loss may be acceptable: a 0.72 correlation—generally considered high—is claimed between the new predicted locations and the annotated tweet-level geotags.

An additional factor arguing for automatic geolocation is that the set of geotagged tweets may not follow the same word/topic/etc. distribution as the overall set of tweets (Pavalanathan and Eisenstein, 2015). Tweets geotagged with exact coordinates come primarily from cell phones, as opposed to the various other ways of creating tweets (e.g. desktops, laptops, pads), and only when the cell phone user has explicitly enabled this feature, which is not on by default. This suggests that the average user producing geotagged tweets is more likely to use his/her cell phone as the primary means of digital communication and is less concerned about privacy than others. Together they point to young "digital natives", and anecdotal investigations of geotagged tweets bear this out, with high-school students and Internet slang heavily represented (Pavalanathan and Eisenstein, 2015).

There are significant ethical issues involved in social media geolocation. This is especially the case with methods such as text-based geolocation that are capable of recovering a latent signal representing location that a user might want to hide but has no clear means of doing so. Geolocation techniques that rely on settings over which the user has control, such as explicit latitude/longitude geotags or a voluntarily provided location field, are less problematic, but there is often no way to defeat a text-based or network-based geolocation algorithm other than not to use social media at all. Most users in fact greatly value the privacy of their location (Junglas and Spitzmuller, 2005). Smith et al. (1996) identify four areas of privacy that usually trigger concerns in users:

1. the collection of personal information;
2. the unauthorized use of that information;
3. unauthorized access to that information;
4. errors in the information.

Item #2, unauthorized use, is the most worrisome per the authors, leading to negative outcomes ranging from mildly annoying (e.g. receiving spam) to potentially life-threatening. For this reason, it is considered critically important to develop standards to protect the privacy of location information (McMullan, 2014). Various researchers have sought to develop such standards (Michael et al., 2008; Anuar and Gretzel, 2011; ISACA, 2011), but laws to enforce these standards are still in the process of being developed.[7]

## 1.6 Application to digital humanities

Millions of historical documents exist, and humanities researchers traditionally faced a prohibitive task doing large-scale analyses of such primary sources. As a result, they had to be content with close reading and analysis of a small set of carefully selected sources. However, the development of accurate optical character recognition (OCR) software, combined with computational data analysis techniques, has recently facilitated the development of the field of *digital humanities* (Burdick et al., 2012), allowing such large-scale analyses to be done. Geolocation can be of great assistance in quickly extracting and summarizing the geographic data available in such datasets, and in fact an entire field, known as the *spatial humanities* (Bodenhamer et al., 2010), has developed around the marriage of geographic information systems (GIS) and the digital humanities. The quantitative methods in this field have allowed for a revolution in the detailed and large-scale understanding of historical and literary phenomena that heretofore had resisted analysis. The techniques pursued in this dissertation are of particular interest to this field because the documents used in the field typically lack the metadata common to social media datasets, forcing geolocation to rely primarily or exclusively on the text itself.

Mapping has long been important to the humanities due to its ability to compactly represent large amounts of data. In the mid 19th century, Charles Joseph Minard produced a series of such maps variously representing the movement of goods and people across Europe, including what "may well be the best statistical graphic ever drawn" (Tufte, 1986) — Minard's famous 1869 map depicting Napoleon's disastrous 1812 invasion of Russia (Figure 1.8). At the end of the 19th century, Charles Booth produced a famous series of maps of poverty and crime in London (Booth, 1902).

[7] In the United States, as of September, 2015 there are laws in various states to protect the privacy of location information but no federal law, although congressional bills have been introduced to this effect, such as S. 237 and H.R. 491 in 2015, S. 2171 in 2014 and H.R. 983 in 2013.

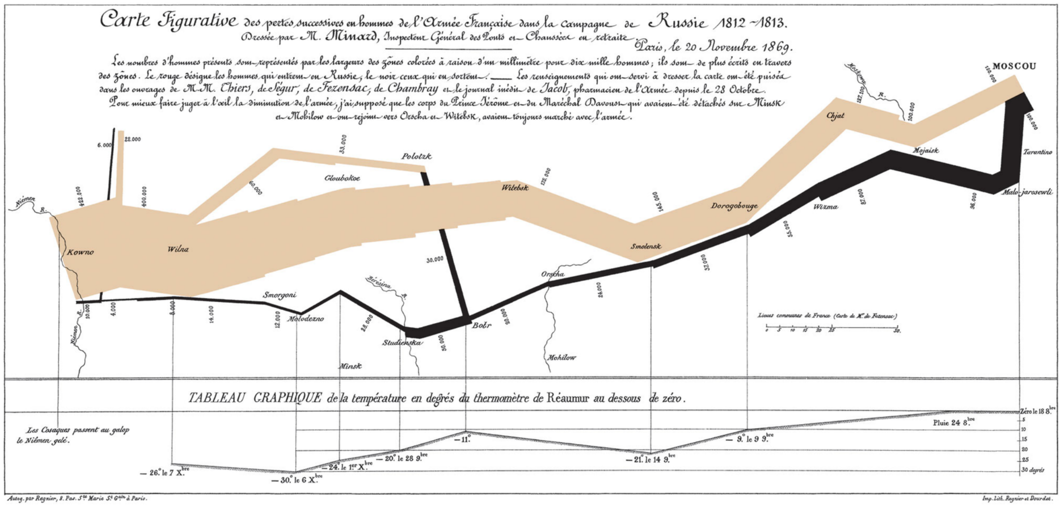

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Niemen R.
422.000
6.000
22.000
400.000
Kowno
Wilna
60.000
Gloubokoe
33.000
Polotzk
Witebsk
175.000
145.000
Smolensk
Dorogobouge
Wixma
Chjat
Mojaisk
MOSCOU
Tarantino
Malo-jarosewli
Orscha
Mohilow
Smorgoni
Molodezno
Minsk
Berezina R.
Studienska
Bobr
30.000
50.000
28.000
20.000
24.000
12.000
14.000
8.000
4.000
10.000
Niemen R.
Lieues communes de France (Carte de M. de Fezensac)
0 5 10 15 20 25 50.
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24.° le 1.er X.bre
— 20.° le 28 9.bre
— 11.°
— 21.° le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. S.te Marie S.t G.ain à Paris.
Imp. Lith. Regnier et Dourdet.

Figure 1.8: Charles Joseph Minard's famous 1869 map of Napoleon's Russia Campaign.

However, before the so-called "spatial turn" in the humanities beginning in the 1990's (Guldi, 2009) that led to the emergence of the spatial humanities, such maps were difficult and time-consuming to produce. Recent years have seen an explosion of map-related research projects in the humanities (Cohen, 2011). Through mining historical texts, researchers have mapped topics as disparate as the spread and retreat of cholera and other diseases in 19th century Britain, as shown in Figure 1.9 (Murrieta-Flores et al., in press); the maritime transmission of Buddhism from India to China along trade routes (Lancaster, 2014); Robert E. Lee's knowledge (or lack thereof) of troop movements during the Battle of Gettysburg (Knowles, 2013); the spread of accusations of witchcraft during the Salem witch trials of the late 1600's (Ray, 2002); the reasons underlying the Dust Bowl of the 1930's (Cunfer, 2008); and other issues. Current work is focused on moving beyond simply connecting GIS and the humanities to incorporate advances in computational linguistics (Gregory et al., 2013).

To aid the digital humanities, I have defined a new task—text-based historical-corpus document geolocation. For use with this task I annotated, or supervised the annotation of, two historical digital humanities corpora, a larger one (War of the Rebellion or WOTR, based on U.S. Civil War archives) and a smaller one (*Western Wilds* or BEADLE, a 19th-century travel log). These corpora come with careful annotations as well as a larger set of unannotated data from the same distribution. I demonstrate good accuracy using the same methods I developed earlier in my dissertation,

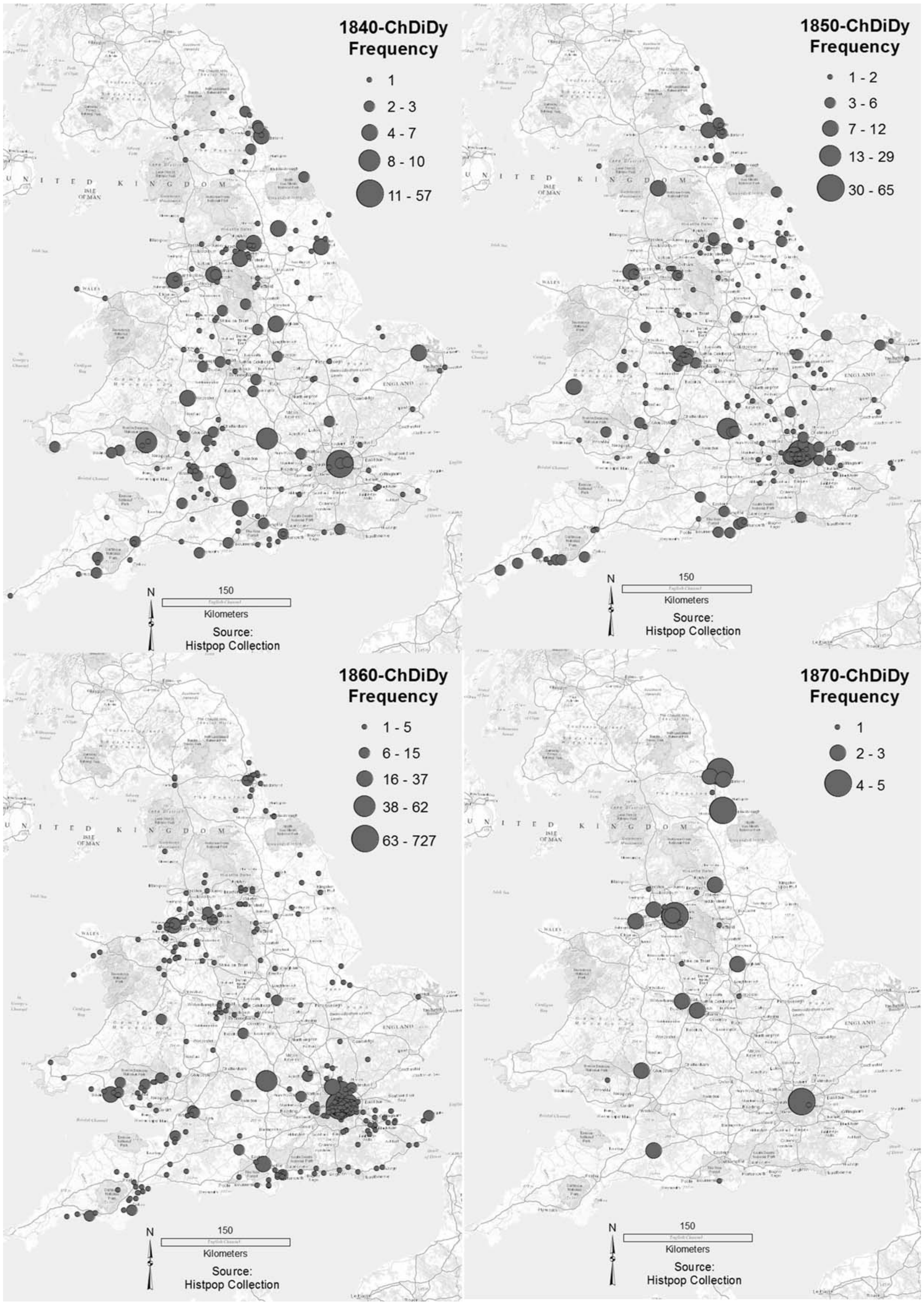


Figure 1.9: Map of the occurrence of cholera, diarrhea and dysentery in 19th century Britain, from Murrieta-Flores et al. (in press).

achieving 72% accuracy with a median of less than 50 km error on WOTR, and 59% accuracy and less than 100 km median error on the smaller BEADLE data set. I also investigate various types of domain adaptation using Wikipedia as an out-of-domain training corpus, and I show with learning curves the additional benefit that out-of-domain data yields.

Researchers in the digital humanities often make use of *topic models* (Blei and Lafferty, 2009), particularly those derived using *latent Dirichlet allocation* (Blei et al., 2003). Topic models are automatically derived collections of statistically related words, where words that tend to co-occur in the same contexts are grouped together into "topics" that frequently (although not always) can be identified with a coherent, real-world subject. Topic models, when properly analyzed — often with the aid of well-designed visualizations — can reveal a great deal about the a data set. In §5.5, I compute topic models segmented in various ways by geography, which is possible through taking my geolocation models trained on the annotated portion of the data (possibly in conjunction with out-of-domain data in a domain adaptation setting), applying them to the full set of unannotated data (§2.3), and dividing up the unannotated text according to the predicted location. This allows me to compute *dynamic topic models* (Blei and Lafferty, 2006) that show the change in topic membership of particular words over geography and/or over time, producing a detailed picture of differences in subject matter and approach over time and space. This in turn allows for careful variationist analysis to be performed (§5.5.2). The results reveal a mixture of expected and unexpected results, where the expected results can be used to calibrate the accuracy of the topic model and the unexpected results used to produce genuine new insights.

Finally, in Chapter 6 I develop a new text-based document geolocation method based on co-training between document-level annotations and toponym identification and resolution. This method works on text-only corpora, such as the digital humanities corpora described above, and allows me to introduce additional, outside domain knowledge (in the form of a gazetteer) while still remaining within a pure-text scenario. Co-training has many variants; I include careful analysis of the relative strengths and weaknesses of different approaches. I also develop and justify a metric for evaluating the success of my co-training algorithm.

## 1.7 Outline

Chapter 2 describes my sources of data, including three separate Twitter corpora (originally constructed by Eisenstein et al. (2010), Roller et al. (2012) and Han et al. (2014)); recent processed dumps of the English, German and Portuguese versions of Wikipedia; the COPHIR corpus of tagged Flickr images (Bolettieri et al., 2009); the BEADLE corpus derived from the 19th century travel log *Western Wilds* by John Beadle, which I annotated myself; and WOTR, derived from the official Civil War archives (*War of the Rebellion*), whose annotation I supervised. In addition, I describe three toponym-resolution corpora that are used for applying document geolocation techniques to toponym resolution.

Chapter 3 describes in detail the construction of a grid of cells and the supervised models used for document geolocation that are built on them. The cells can be constructed using either uniform or adaptive (*k*-d tree) grids. The various models described include those based on information retrieval techniques, as well as higher-accuracy techniques that rely on logistic regression, either by itself or as part of a hierarchical process that uses multiple logistic-regression classifiers. I also implement a feature-selection technique based on information gain ratio (IGR), for comparison with Han et al. (2014).

Chapter 4 describes the experiments I carried out on the various modern corpora (from Wikipedia, Twitter and Flickr) and the results I obtained. Hierarchical classification is the clear winner, beating the other methods on all of the large corpora I evaluate. Flat logistic regression is also fairly effective and able to beat Naive Bayes on many of the corpora, while IGR works only on the Twitter corpora. Naive Bayes and KL divergence are of comparable performance.

Chapter 5 investigates applications of document geolocation to some 19th-century digital humanities corpora that I either annotated myself or supervised the annotation of. I show how my methods can be extended to work well in this context, with improved results obtained using domain adaptation with the English Wikipedia as an out-of-domain source of labeled data. I then take the predicted locations of the documents in the full set of Civil War documents (§2.3), group them according to membership in a set of hand-drawn "theater of war" regions, and apply dynamic topic models to these regions, which allows for careful, variationist analysis across the geographic and temporal scope of the Civil War.

Chapter 6 describes experiments in informing toponym resolution with document-level geolocation, expanding upon the work of Speriosu (2013). I develop a means of using co-training to simultaneously train a document geolocator and a toponym resolver on a combination of document-geolocated Wikipedia text and toponym-resolved Civil War text. This allows me to jointly exploit the complementary knowledge contained in both sources of geographic information, including the outside knowledge contained in a gazetteer of toponyms and their possible resolutions.

Chapter 7 summarizes the work performed for this dissertation and further directions to take the research.

## 1.8 Contributions

This dissertation includes the following contributions to the field of natural language processing:

- An investigation of various effective methods for supervised geolocation of a test document, i.e. associating the document with a particular set of latitude/longitude coordinates on the Earth. I consider methods that divide the Earth's surface into rectangular grid cells—either of constant degree size or using an adaptive *k*-d tree (Bentley, 1975)—and find the single best grid cell, relying exclusively on the text of a document. Many of these methods are simple to implement and fast to run, but give comparable accuracy to more complicated and slower Bayesian methods. These methods are fast enough to be scaled up to a large amount of training material, even with a fine-scale grid mesh, and easy to parallelize. Among these methods are Naive Bayes, KL divergence, Average Cell Probability (which involves inverting unigram distributions to determine a distribution of cells for a given word) and a few different baselines.

- An application and careful analysis of the performance of these methods, along with various smoothing techniques as well as the information gain ratio (IGR) feature selection technique of Han et al. (2014), to several different corpora: Three corpora of Twitter user feeds of distinct natures; dumps of the English, German and Portuguese versions of Wikipedia, processed by custom-written software; and a large set of image tags corresponding to geotagged Flickr images. I consider a number of different evaluation metrics and show that none of the geolocation techniques consistently outperforms Naive Bayes on all the corpora, including IGR, which performs well on the Twitter corpora for which is was designed, but not on the other corpora.

- The application of logistic regression to geolocation. Contrary to the claims of Han et al. (2014), I show that logistic regression can be more accurate than Naive Bayes and KL divergence (including variants incorporating feature selection) and fast enough to run on large corpora. Logistic regression itself very effectively picks out words with high geographic significance. In addition, because logistic regression does not assume feature independence, complex and overlapping features of various sorts can be employed.

- A new method for supervised geotagging, which involves a hierarchical discriminative classifier that creates multiple individual classifiers at different grid resolutions and combines them to achieve better results than could be done using a classifier at a single level. This method scales well to large training sets and greatly improves results across a wide variety of corpora. In fact, I am able to achieve state-of-the-art results on all of the large corpora I evaluate on. Importantly, this is the first method that improves upon straight uniform-grid Naive Bayes on all of these corpora, in contrast with *k*-d trees (Roller et al., 2012) and the current state-of-the-art technique for Twitter users of geographically-salient feature selection Han et al. (2014).

- The development of a new NLP task, *text-based document geolocation of historical corpora*, to assist the application of document geolocation techniques to the digital humanities. I apply my techniques to two new corpora (see below), establishing a baseline for further research and showing how good accuracy can be achieved even with text-only documents and with relatively little training material. I further show how domain adaptation techniques that make use of the large amount of geographic annotation available in the English Wikipedia can dramatically reduce the amount of annotated training data required to achieve equivalent levels of performance. I create learning curves to investigate the minimal amount of annotation required to achieve a given level of performance, to assist further researchers in deciding how much money and effort to spend on annotation.

- Two new annotated historical corpora for use with the new NLP task I developed (see above). These two corpora are of significantly different size and subject matter—a 19th-century American travel log (*Western Wilds* by John Beadle) and a large set of primary-source documents from the American Civil War archives (*War of the Rebellion*). The annotations on the Civil War archives are in the form of polygons or multipoints when appropriate, allowing for more sophisticated analyses than can be achieved with typical single-point annotations. I put a siginificant amount of work into cleaning up the entire set of Civil War archives (not just the annotated portion) and dividing it into individual documents—some 255,000 in all, spread over 126 volumes. This alone should be of great benefit to digital humanities researchers interested in further work on *War of the Rebellion*. All the data will be released publicly, along with the source code required to process the data and detailed instructions on how to operate it.

- A new technique for creating *geographic topic models*, based on David Blei's dynamic topic models (Blei and Lafferty, 2006) and applied to the above Civil War archives. This involves identifying, using a domain expert, a set of regions corresponding to coherent *theaters of war*. These are then linearized and treated similarly to the timeslices of a standard dynamic topic model. This has the effect of creating topics that vary over geography, allowing for broad-ranging variationist analyses to be performed.

- A new geolocation technique for text-only corpora involving co-training between document geolocation and toponym resolution, building on the toponym resolution methods previously investigated by Speriosu (2013). This has the effect of introducing external information into the process in the form of a gazetteer of locations, yielding the potential to significantly increase the geolocation accuracy beyond what can be extracted from the text alone. This demonstrates superior results on some metrics, and can serve as a branching-off point for further research in this area.

- A program that implements the methods described above.

- Processed versions of all the corpora I use for evaluation (to the extent this is legally possible), and programs for recreating them. (This includes the various modern and toponym-resolution corpora I make use of, not just the two new historical corpora I annotated.) Similarly processed versions from other sources can also be created (e.g. different dumps of Wikipedia, different sets of tweets). These programs are written in a modular fashion, so that the components that do various types of processing can be reused in other programs needing such processing, and new components can be created to do similar operations, e.g. as might be required to analyze the dump of a foreign-language Wikipedia.

# Chapter 2

# Data

## 2.1 Introduction

I work with a number of datasets annotated with document-level geotags (locations in the form of latitude/longitude coordinates).[1] I have nine such datasets available for evaluation — seven modern (three of tweets, three of Wikipedia articles, and one of Flickr photos) and two historical (a single-book travel log and a multi-volume collection of Civil War archives). Of the modern corpora, one of the Twitter datasets is fairly small, but all of the others are much larger, consisting of at least several hundred thousand training instances. The two historical corpora were annotated manually and as a result consist of fewer annotations than the large modern corpora, but still differ significantly among each other in size. Two of the three Twitter datasets and the two historical datasets are primarily localized to the United States, while the remaining datasets cover the whole world. See Table 2.1 for a summary of the datasets.

For the toponym-resolution work described in Chapter 6, I also make use of various other datasets, which are generally not annotated with document-level geotags but usually do have individual toponyms annotated. This is documented more below.

[1] This chapter is partly based on Wing (2011), Wing and Baldridge (2011) and Wing and Baldridge (2014). Jason Baldridge was my advisor for these works and helped edit the papers.

| Dataset | Corpus<br>Source | Document<br>Source | Scope | #Docs<br>(Training) | #Tweets<br>(Total) | #Types<br>(Training) | #Tokens<br>(Training) | #Tokens<br>/Doc |
|---|---|---|---|---|---|---|---|---|
| GeoText | Twitter | User feed | US | 5.69K | 378K | 114K | 1.58M | 277.3 |
| TwUS | Twitter | User feed | US | 430K | 38M | 4.75M | 244M | 568.8 |
| TwWorld | Twitter | User feed | World | 1.37M | 12M | 95.4K | 41.8M | 30.6 |
| EnWiki13 | English Wikipedia | Article | World | 691K | — | 4.32M | 174M | 251.7 |
| DeWiki14 | German Wikipedia | Article | World | 259K | — | 4.09M | 129M | 497.5 |
| PtWiki14 | Portuguese Wikipedia | Article | World | 105K | — | 608K | 18.4M | 175.3 |
| CoPhIR | Flickr | Single image tags | World | 2.27M | — | 629K | 20.5M | 9.04 |
| Beadle | *Western Wilds* | Paragraph | US | 244 | — | 6.16K | 16.6K | 67.8 |
| WOTR | *War of the Rebellion* | Article | US | 4008 | — | 25.9K | 526K | 131.3 |

Table 2.1: Summary of datasets with document-level geotags used in the dissertation. Note that type and token counts exclude stopwords, and TwWorld was pre-filtered to exclude non-alphabetic words, words shorter than 3 characters in length and words occurring less than 10 times in the entire corpus.

## 2.2 Modern geolocation datasets

### 2.2.1 Twitter datasets

This dissertation uses a number of datasets collected from Twitter. These datasets are taken from tweets collected using one of the public streaming API's. Some of these API's yield a sample of all publicly-available tweets created (e.g. the Spritzer and Gardenhose API's). Others allow for tweets to be searched using particular characteristics; for this dissertation, tweets were requested that were geotagged with a specific latitude/longitude coordinate, generally derived from a GPS device embedded in the cell phone used to send the tweet.

Because tweets are so short (at most 140 characters), documents serving as training instances are constructed by amalgamating the tweets of a given user. The location of a user is deemed to be the earliest tweet with specific, GPS-assigned latitude/longitude coordinates. This choice follows Eisenstein et al. (2010). It is possible that other methods (e.g. choosing the centroid of all available coordinates) may prove to be more appropriate.

An additional issue to be considered is that the distribution of tweets created by users who post to Twitter through cell phones and allow tweet geotagging may not be the same as the overall distribution. For example, casual inspection reveals that these are generally young users, often in high school, who use a great deal of Internet slang (see example below).

The following is an example of some tweets from a particular user (with references to other users anonymized):

| Date/time | Coordinates | Text |
|---|---|---|
| 2010-03-03T02:02:04 | 40.2015,-74.806535 | Watching LOST |
| 2010-03-03T12:01:41 | 40.221968,-74.734795 | @USER_89a3500b i did |
| 2010-03-03T20:06:19 | 40.221968,-74.734795 | Maneuver so that I can put my team on, hopefully sooner so that we can live our dreams on |
| 2010-03-03T23:30:45 | 40.221333,-74.732688 | Darko was eating hamburgers in the locker room before they played the knicks. Lol |
| 2010-03-04T02:58:43 | 40.220681,-74.758761 | Girl pack ya bags i'm bout to take you on a ride! |
| 2010-03-04T15:26:50 | 40.194523,-74.756427 | @USER_a9cf8f82 lol, yeah check it out bro |
| 2010-03-04T20:03:07 | 40.289891,-74.678256 | RT @USER_5eae722d: #inhighschool me & Mr. Stavisky dnt lk each other, his breath smelled lk straight ass! - lmao he use 2 chase us dn the hall |
| 2010-03-04T23:57:58 | 40.221968,-74.734795 | #inhighschool trenton high girls basketball team always had the best record out of all the highschool sports teams. Nothing has changed |
| 2010-03-05T00:17:13 | 40.221968,-74.734795 | The cheerleading team need Mrs. Grady back |
| 2010-03-05T00:39:56 | 40.221968,-74.734795 | This girl Ashley Hines from is a beast. They can't stop her |
| 2010-03-05T01:02:32 | 40.221968,-74.734795 | Da High always had the best fans. Going way back before I was #inhighschool |
| ... | | |

The same user, with tweets amalgamated and converted to a unigram distribution, appears as follows:

```
USER_6197f95d    40.2015,-74.806535    had:4 everybody:1 we:1 bub:2 funk:1 lights:1 u:1 said:1 he:3 who:1 to:6 of:1 lol:3 she:1 knicks:1
                                       pack:1 room:1 played:1 nothing:1 #inhighschool:3 bags:1 ride:1 &:1 mr:1 be:2 changed:1 any:1
                                       or:1 is:3 i'm:2 if:1 reopen:1 up:2 can't:1 that:5 eating:1 darko:1 dreams:1 our:1 live:1 sooner:1
                                       hopefully:1 team:3 did:1 locker:1 hamburgers:1 needs:1 union:1 repaired:1 hurry:1 ewing:1 da:1
                                       looking:2 pause:1 down:2 dj:3 life:1 does:2 all:2 see:1 from:1 has:1 2:1 em:1 bro:1 they:2 one:1
                                       so:3 and:1 oh:1 new:2 freezer:1 watching:1 lost:1 ass:1 yo:2 there:1 maneuver:1 put:1 looks:1 cheer:1
                                       sound:1 approach:1 caution:1 burger:1 hit:1 jack:1 daniels:1 stop:1 liquor:1 credit:1 hard:1 best:2
                                       ...
```

**GEOTEXT** is a small dataset consisting of 377,616 tweets from 9,475 users tweeting inside of a bounding box consisting of the 48 American states (and some parts of Canada and Mexico), compiled by Eisenstein et al. (2010). It was compiled from tweets collected using the Gardenhose API during the first week of March 2010. Tweets without geotags (GPS-assigned latitude/longitude coordinates) were discarded, as were users with fewer than 20 geotagged tweets. Also discarded were users following or followed by 1,000 or more other users, in order to eliminate marketers, celebrities, news media sources, etc. (Kwak et al., 2010).

**TWUS** is a dataset of tweets compiled by Roller et al. (2012), designed to address the sparsity problems resulting from the small size of GEOTEXT. Tweets were collected using both the Spritzer and location-search API's over the period from September 4 to November 29, 2011.

Filtering and amalgamation were done similar to GEOTEXT. However, tweets without geotags were not discarded; instead all users with at least one geotagged tweet were considered. (This may have the effect of lessening somewhat the potential distribution mismatch between geolocated and non-geolocated tweets.) The resulting dataset contains 38M tweets from 450K users, of which 10,000 each are reserved for the development and test sets.

**TWWORLD** is a dataset of tweets compiled by Han et al. (2012). It was collected using the Spritzer API over the period from September 21, 2011 to February 29, 2012 and differs from TWUS in that it covers the entire Earth instead of primarily the United States, and consists only of geotagged tweets. Non-English tweets and those not near a city were removed, and non-alphabetic, overly short and overly infrequent words were filtered. The resulting dataset consists of 12M tweets from 1.4M users, with 10,000 each reserved for the development and test sets. Note that, even though this dataset contains more users than TWUS, it consists of fewer tweets, meaning that the average document size is significantly smaller (8.6 tweets/user, vs. 84.4 tweets/user for TWUS).

### 2.2.2 Wikipedia datasets

As of November 2014, Wikipedia has some 34.0 million content-bearing articles in 241 language-specific encyclopedias.[2] Among these, 52 have over 100,000 articles and 12 have over 1 million articles, including 4.8 million articles in the English-language edition alone. Wikipedia articles generally cover a single subject; in addition, most articles that refer to geographically fixed subjects are *geotagged* with their coordinates. Such articles are well-suited as a source of supervised content for document geolocation purposes. Furthermore, the existence of versions in multiple languages means that the techniques in this paper can easily be extended to cover documents written in many of the world's most common languages.

Wikipedia's geotagged articles encompass more than just cities, geographic formations and landmarks. For example, articles for events (like the shooting of JFK) and vehicles (such as the frigate USS *Constitution*) are geotagged. The latter type of article is actually quite challenging to geolocate based on the text content: for example, though the USS *Constitution* is moored in Boston,

[2] `http://stats.wikimedia.org/EN/Sitemap.htm`

most of the page discusses its role in various battles along the eastern seaboard of the USA. However, such articles make up only a small fraction of the geotagged articles.

For the experiments in this paper, we used full dumps of versions of Wikipedia in three different languages:[3]

1. English (**ENWIKI13**) from November 4, 2013, with 864K geotagged articles out of 4.44M total

2. German (**DEWIKI14**) from July 5, 2014, with 324K geotagged articles out of 1.71M total

3. Portuguese (**PTWIKI14**) from June 24, 2014 131K geotagged articles out of 817K total

These dumps include not only the content-bearing articles but various types of special-purpose articles used primarily for maintaining the site (specifically, redirect articles and articles outside the main namespace), which were filtered out. For example, although the English Wikipedia version mentioned above has 4.44M content-bearing articles, the dump actually has 14.0M articles in it—i.e. almost 10M of the articles in the dump are special-purpose, non-content-bearing articles.

It is necessary to process the raw dump to obtain the plain text, as well as metadata such as geotagged coordinates. Extracting the coordinates, for example, is not a trivial task, as coordinates can be specified using multiple templates and in multiple formats. Automatically-processed versions of the English-language Wikipedia site are provided by Metaweb,[4] which at first glance promised to significantly simplify the preprocessing. Unfortunately, these versions still need significant processing and they incorrectly eliminate some of the important metadata. In the end, we wrote our own code to process the raw dump, involving about 4,600 lines of Python code and 1,200 lines of shell script. It should be possible to extend this code to handle other languages with little difficulty. (An alternative strategy, perhaps better in hindsight, would have been to download and run the MediaWiki software used to process the Wikipedia article source code into HTML, and parse the resulting HTML.) See Lieberman and Lin (2009) for more discussion of a related effort to extract and use the geotagged articles in Wikipedia.

[3] `http://download.wikimedia.org/`

[4] `http://download.freebase.com/wex/`

The entire set of articles was split 80/10/10 in round-robin fashion into training, development, and testing sets after randomizing the order of the articles, which preserved the proportion of geotagged articles.

### 2.2.3 Flickr datasets

**COPHIR** (Bolettieri et al., 2009) is a large dataset of images from the photo-sharing social network Flickr. It consists of 106M images, of which 8.7M are geotagged. Most images contain user-provided tags describing them. I follow algorithms described in O'Hare and Murdock (2013) in order to make direct comparison possible. This involves removing photos with empty tag sets and performing *bulk upload filtering*, retaining only one of a set of photos from a given user with identical tag sets. The resulting reduced set of 2.8M images is then divided 80/10/10 into training, development and test sets. The tag set of each photo is concatenated into a single piece of text (in the process losing user-supplied tag boundary information in the case of multi-word tags). The resulting documents tend to be extremely short (often less than 10 words) but consist of words that tend to have high geographic salience.

## 2.3 Historical texts

No document-level annotations exist for the historical texts I am interested in studying. For this reason, I annotated part of a book-length historical travel log, John Beadle's *Western Wilds, and the Men Who Redeem Them*, published in 1878. In addition, as part of a project funded by the New York Community Trust (NYCT), I supervised the annotation of parts of *The War of the Rebellion: a Compilation of the Official Records of the Union and Confederate Armies*,[5] a set of over 100 volumes of archives of the American Civil War.[6]

I use these texts to demonstrate the feasibility of text-based geolocation with smaller amounts of annotated material, especially in conjunction with domain adaptation (Chapter 5) involving additional training on the English Wikipedia (ENWIKI13, §2.2.2).

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

[6] This work was done in conjunction with Professors Scott Nesbit and Jason Baldridge, as well as a colleague, Grant DeLozier.

### 2.3.1 *Western Wilds*

*Western Wilds* by John Beadle is one of the books from the PCL Travel collection of 19th-century travel texts.[7] This book is an account of Beadle's travels over a seven year period throughout the Western part of the United States. It includes both direct descriptions of Beadle's travels and interpolated travel stories of people that Beadle encountered. Because it is in the form of a travelogue, it is generally possible to identify a location with each stretch of text—for example, the location that the narrator is assumed to have been at when a story is being narrated, or the geographic topic of interest when a description of a location is interpolated into the narrative. The resulting corpus is termed BEADLE.

I annotated both at the paragraph and sub-paragraph level. For the latter type of annotation, I subdivided the paragraph into chunks (as large or as small as necessary) covering a unified geographic topic, whether a political feature (e.g. city, county, state or Indian reservation) or a natural feature (e.g. a lake, river, mountain or mountain range).

All annotations were done in the form of single points, even when the topic of the paragraph was more naturally described by a polygon (e.g. regions, states or rivers). This was done for consistency with the automatic annotations of the various modern datasets previously described in this chapter, and for computational convenience. This is admittedly not an ideal situation; however, both the algorithms and their evaluation get significantly more complex when polygonal or even rectangular annotations are introduced. Note that in the other text annotated as part of this dissertation, *War of the Rebellion* (§2.3.2), textual spans were in fact annotated with polygons when appropriate, but I still derived a point location from the polygons for use in training and testing.

My actual mechanism for annotating a span of text with a location was in most cases to specify the name of a geolocated feature in Wikipedia, so that I could then use a script to automatically tag each such feature with the appropriate coordinates as found in Wikipedia. However, when doing this I avoided specifying a linear feature such as a river or mountain range in favor of using the point feature in Wikipedia (e.g. a city or landmark) that is as close as possible to the portion of the river or mountain range that the text in question is about. This was done to maximize the accuracy of

[7] This corpus consists of 94 books and approximately 7.7M words, and was collected by the Perry-Castañeda Library at the University of Texas at Austin.

the geolocations given the large geographic scope of many rivers and mountain ranges. (Rivers are especially problematic since they are tagged in Wikipedia by the coordinates of their mouth, rather than something more desirable such as a point along their middle stretch.)

Similarly, I avoided choosing a state as an annotation if the geographic topic could be identified as a sub-region of the state, instead using the point feature as close as possible to the middle of the sub-region in question. When no such sub-region could be identified, I went ahead and annotated using the name of the state. Note, however, that this is problematic because of the need to map the state to a single point, which introduces a large, unavoidable source of error when using a distance-based, point-based evaluation metrics derived from the *error distance* between the predicted and true locations (§4.1.2, §7.7). Regions larger than a state (e.g. the Western United States or the United States as a whole) are even more problematic in this respect, and for this reason I refused to annotate any paragraphs in such a fashion, instead annotating them as "various", which leaves them without coordinates.

In some cases, it was not possible to find a point feature in Wikipedia that was close enough to the desired location, either because no such feature exists at all, because there is no Wikipedia article corresponding to the feature, or because such an article exists but lacks a geolocation. In these cases I manually entered a latitude/longitude coordinate. In some cases, this was obtained elsewhere on the Internet (e.g. the location of the ghost town of Benton, Wyoming, a place not in Wikipedia). In other cases, I estimated approximate coordinates using a map. (An example of this is a paragraph whose location was identified through context as being between the ends (sinks) of the Humboldt and Truckee Rivers in Western Nevada. For this paragraph, I chose a point halfway between the two river sinks.) Some stretches of text without clear geographic focus were annotated as "unclear" and left without coordinates.

When annotating a paragraph with multiple sub-paragraph annotations, one of these annotations was chosen as the paragraph annotation if it appeared to apply to the majority of the paragraph; else, the union of the annotations was determined and the above considerations applied to chose the actual coordinates.

All in all, *Western Wilds* consists of 37 chapters plus a preface, for a total of 1,437 paragraphs. The preface and the first 10 chapters were annotated in their entirety, up through paragraph 408, as well as the first part of several more chapters XI through XVI: paragraphs 409-414 (chapter XI), 439-458 (chapter XII), 486-496 (chapter XIII), 539-547 (chapter XIV), 576-594 (chapter XV), 620-625 (chapter XVI). This is a total of 479 paragraphs annotated, but this produced only 408 data instances because some paragraphs could not be assigned coordinates (e.g. those annotated as "various", "unclear" or "Western United States", as described above).

An example of a paragraph from *Western Wilds* is shown in §1.3; this paragraph was annotated with latitude 39.3422, longitude -120.2036 (the coordinates of Truckee, California). A summary plot of the locations in the book is shown in Figure 1.1. Each location is labeled by the chapter it occurs in (using a Roman numeral), and lines are drawn connecting adjacent paragraphs within a given chapter, with different colors for each chapter. As can be seen, the narrative jumps around a good deal both within a given chapter and across chapters. This partly reflects the multiple times that Beadle traveled across the country and back, and partly reflects stories of other adventurers that Beadle interpolated into his own narrative.

### 2.3.2 *Official Records of the War of the Rebellion*

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) is a large set of American Civil War archives.[8] It was published in 128 books (grouped into 70 volumes, which are further grouped into four series) by the United States Government between 1881 and 1901. The archives consist of military orders and reports, governmental correspondence, proclamations, court reports, maps, and other primary sources generated during the war. Each volume is about 1,000 pages, for a total of 138,579 pages.[9]

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 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 ac-

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

[9] See http://en.wikipedia.org/wiki/Official_Records_of_the_American_Civil_War.

|  | Annotated subset | Full data |
|---|---:|---:|
| Total tokens | 1,743,331 | 57,557,037 |
| Total types | 40,416 | 315,564 |
| Number of volumes | 118 | 126 |
| Number of documents | 7,533 | 254,744 |
| Average tokens per volume | 14,773.99 | 453,205.02 |
| Average tokens per document | 231.43 | 225.94 |
| Average documents per volume | 63.84 | 2,005.86 |
| Average types per volume | 2,402.57 | 16,943.35 |
| Average types per document | 125.50 | 118.53 |
| Number of geometries | 5,010 | 0 |
| Average geometries per volume | 42.46 | 0 |
| Fraction of documents with geometries | 0.665 | 0 |

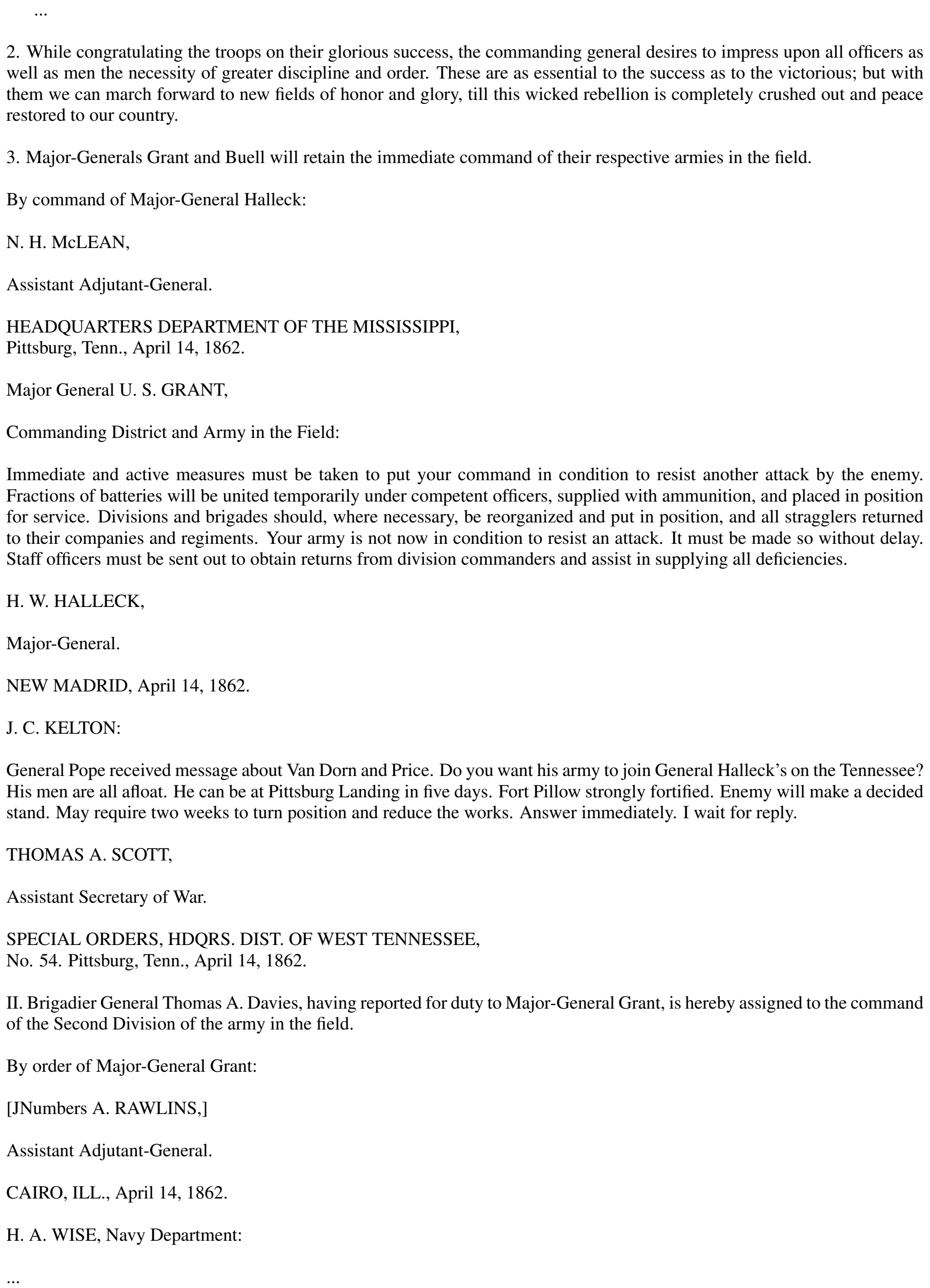
Table 2.2: Statistics on WOTR, annotated subset and full data (using documents predicted based on a sequence model derived from the annotated data).

tion, 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, although not always, the same as the geographic theme; it depends 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.

The total number of documents resulting from this process is 254,744. Statistics on the full WOTR corpus and the annotated subset are shown in Table 2.2.

### Data preparation

Preparation of the data required multiple steps. The source data was taken from an OCR (optical character recognition) scan of the pages of the original printed books, hand-corrected and then dumped directly into a Drupal-based web site, with one HTML document per physical printed page. The web site was then crawled and provided to me as-is. No attempt was made to eliminate the page breaks resulting from this process, which often fall in the middle of a hyphenated word, with lengthy footnotes frequently intervening between the two halves of the separated word.



...

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

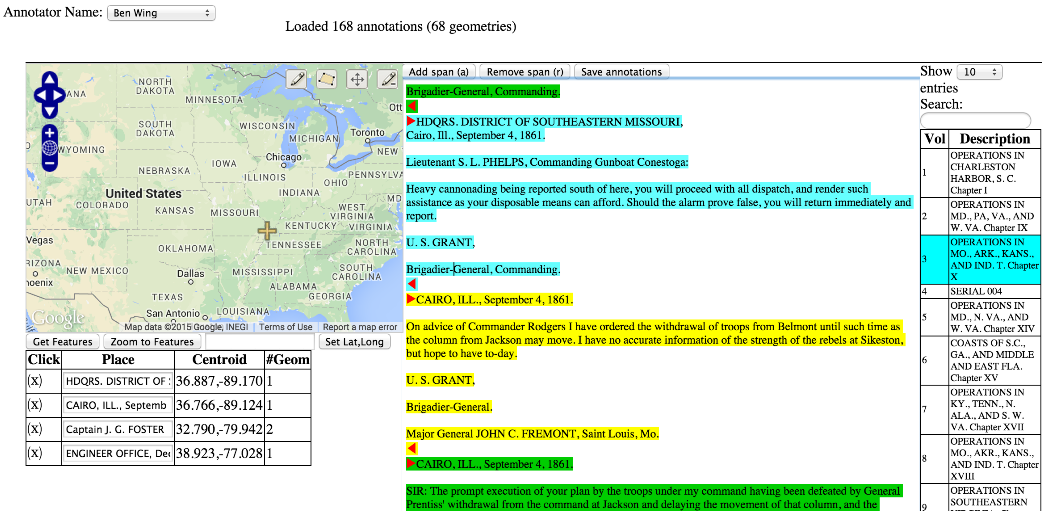
The original volumes are highly structured, with multiple maps, diagrams and tables, and heavy use of typographical conventions (italics, indentation, horizontal rules, etc.) to provide structure, such as to notate the beginning and end of source documents (letters, reports, proclamations, etc.), to offset salutations, closings and headers, and to indicate quoted text embedded in a document. Almost all of this information was lost in the OCR scan, and only sporadic and highly inconsistent attempts were made to recover some of this structure during the hand correction process. The result is that the text of the various source documents runs together, and it is often difficult to determine where one document starts and ends.

The following steps were necessary to produce the final annotated 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. Hire annotators to create the geographic annotations.
4. Create a sequence model using a CRF (conditional random field) to automatically split up the continuous text into documents, training it on the documents manually marked up by the annotators.

Figure 2.1 is an example of part of the source text of a volume, after preprocessing to stitch up page breaks and remove footnotes, headers, footers, etc., but before splitting into individual documents.

**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 often interrupting a paragraph in the middle of a word (broken with a hyphen), interposed with further text such as footnotes, stray headers and footers and the like, often in an inconsistent fashion. In order to derive a set of uninterrupted documents, it was first necessary to rejoin the text across these page breaks. Given that there are over 100,000 pages of text, doing this by hand was out of the question. A program was written that



Annotator Name: Ben Wing

Loaded 168 annotations (68 geometries)

| Click | Place | Centroid | #Geom |
|---|---|---|---|
| (x) | HDQRS. DISTRICT OF : | 36.887,-89.170 | 1 |
| (x) | CAIRO, ILL., Septemb | 36.766,-89.124 | 1 |
| (x) | Captain J. G. FOSTER | 32.790,-79.942 | 2 |
| (x) | ENGINEER OFFICE, Dec | 38.923,-77.028 | 1 |

Brigadier-General, Commanding.
◀
▶HDQRS. DISTRICT OF SOUTHEASTERN MISSOURI,
Cairo, Ill., September 4, 1861.

Lieutenant S. L. PHELPS, Commanding Gunboat Conestoga:

Heavy cannonading being reported south of here, you will proceed with all dispatch, and render such assistance as your disposable means can afford. Should the alarm prove false, you will return immediately and report.

U. S. GRANT,

Brigadier-General, Commanding.
◀
▶CAIRO, ILL., September 4, 1861.

On advice of Commander Rodgers I have ordered the withdrawal of troops from Belmont until such time as the column from Jackson may move. I have no accurate information of the strength of the rebels at Sikeston, but hope to have to-day.

U. S. GRANT,

Brigadier-General.

Major General JOHN C. FREMONT, Saint Louis, Mo.
◀
▶CAIRO, ILL., September 4, 1861.

SIR: The prompt execution of your plan by the troops under my command having been defeated by General Prentiss' withdrawal from the command at Jackson and delaying the movement of that column, and the

Show 10 entries
Search:

| Vol | Description |
|---|---|
| 1 | OPERATIONS IN CHARLESTON HARBOR, S. C. Chapter I |
| 2 | OPERATIONS IN MD., PA, VA., AND W. VA. Chapter IX |
| 3 | OPERATIONS IN MO., ARK., KANS., AND IND. T. Chapter X |
| 4 | SERIAL 004 |
| 5 | OPERATIONS IN MD., N. VA., AND W. VA. Chapter XIV |
| 6 | COASTS OF S.C., GA., AND MIDDLE AND EAST FLA. Chapter XV |
| 7 | OPERATIONS IN KY., TENN., N. ALA., AND S. W. VA. Chapter XVII |
| 8 | OPERATIONS IN MO., AKR., KANS., AND IND. T. Chapter XVIII |
| 9 | OPERATIONS IN SOUTHEASTERN |

Figure 2.2: Screenshot of annotation tool used for adding geometries to document spans.

used various heuristics to do the majority of work, although several more steps and a good deal of manual editing was required to achieve satisfactory results.

**GUI annotation tool** In conjunction with my colleague Grant DeLozier, I wrote a GUI annotation tool that allows document spans to be selected in a text box and points or polygons added on a map. Figure 2.2 shows a screenshot 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). As shown, the blue span has a point geometry, indicated as a large cross on the map at a point in extreme southeast Missouri (slightly to the south of Cairo, Illinois). Points can be added directly on the map, by entering a latitude/longitude coordinate into the text box and clicking **Set Lat/Long**, or by using the list of recent locations below the map.

The annotation tool is written in JavaScript, with data stored using Parse, a *backend-as-a-service* (BAAS) which allows for free data storage within certain storage and bandwidth limits.

**Hiring annotators** 5 annotators were hired, with the intention of having each work for about 50 hours. Detailed instructions were given as to how to correctly divide spans and how to decide what

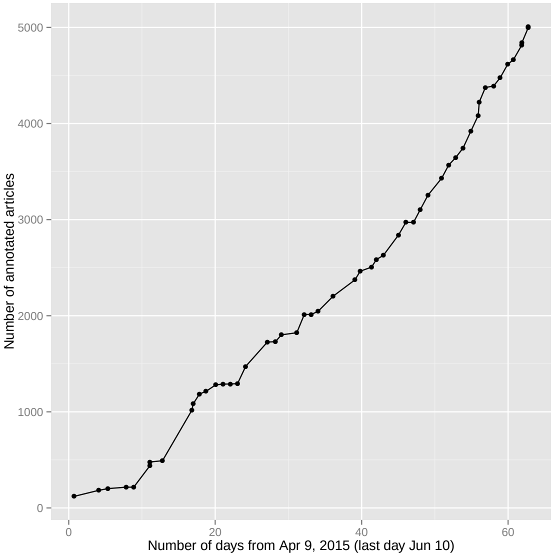


Figure 2.3: Graph of number of annotated articles as a function of time.

counts as the geographic theme. 100 pages (pages 100-199) were selected from each volume, and each annotator was originally assigned 10 volumes, based on an assumed annotation time of 5 hours per volume. This was eventually changed so that 25 pages of each volume were annotated, in an attempt to get at least some annotations on every volume, to increase the geographic and temporal diversity of the annotations.

In the end, 118 out of 126 volumes had annotations provided for them. As it turned out, nearly all of the annotations were done by one annotator, who was responsible for about 200 of the 250 total hours. This was because this annotator was the only one willing to work consistently; the others worked for a few hours and then become unresponsive.

Figure 2.3 shows a graph of the number of annotated articles as a function of time. During the first 25 days or so, a number of annotators were working, but fitfilly, leading to the bumps in the graph. After this, all work was done by the single annotator. His output accelerated slightly over time as we gradually increased the number of hours he worked.

See below for more discussion on how the annotation process was guided.

**Automatically locating document spans** As mentioned above, 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 screenshot in Figure 2.2, 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 is manually annotated.

To do this, a sequence model was created using a CRF (conditional random field) in 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) in 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

- the presence of a date at the end of a line, possibly followed by a time;
- the presence of certain place-related terms typically indicating a header line, such as *HEADQUARTERS*, *HDQRS* or *FORT*;
- the presence of a rank-indicating word (e.g. *Brigadier*, *General* or *Commanding*) at the beginning of or within a line;
- the presence of a line beginning with a string of capital letters, typically indicating a header line;
- the presence of certain words (e.g. *obedient servant*) typically indicating a salutation;
- the combination of the above features with certain punctuation at the end of the line (comma, period, or colon);
- the length of a line;

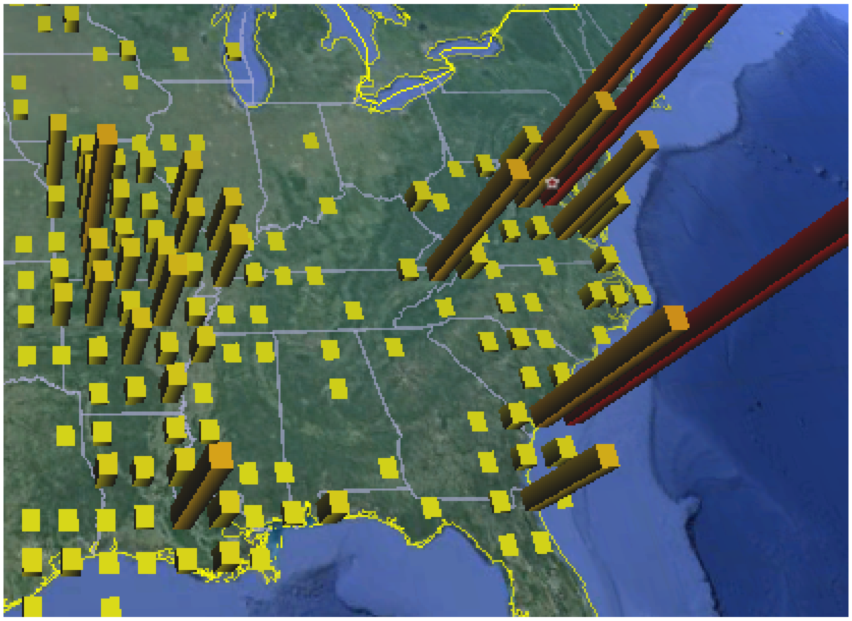


Figure 2.4: KML distribution of the annotated corpus on May 3rd.

- all of the above features for the actual paragraph in question as well as the previous, second-previous, next, second-next, and combinations thereof;
- the first and last words of the paragraph, after stripping out punctuation.

The resulting model performed rather well, but did not consistently handle correctly 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.

### Guiding the annotation process

To guide the annotation process, I selected individual volumes from among the 126 total volumes in approximately ten batches, informed by the distribution of articles produced so far. As new annotations came in, I generated KML graphs of the article distributions and used them to choose both the next set of volumes and which section of each volume to annotate, to maximize the spread of annotations. Three stages are shown in Figure 2.4, Figure 2.5 and Figure 2.6. The most dramatic change is apparent from Figure 2.4 to Figure 2.5, where the entire middle of the graph (particularly

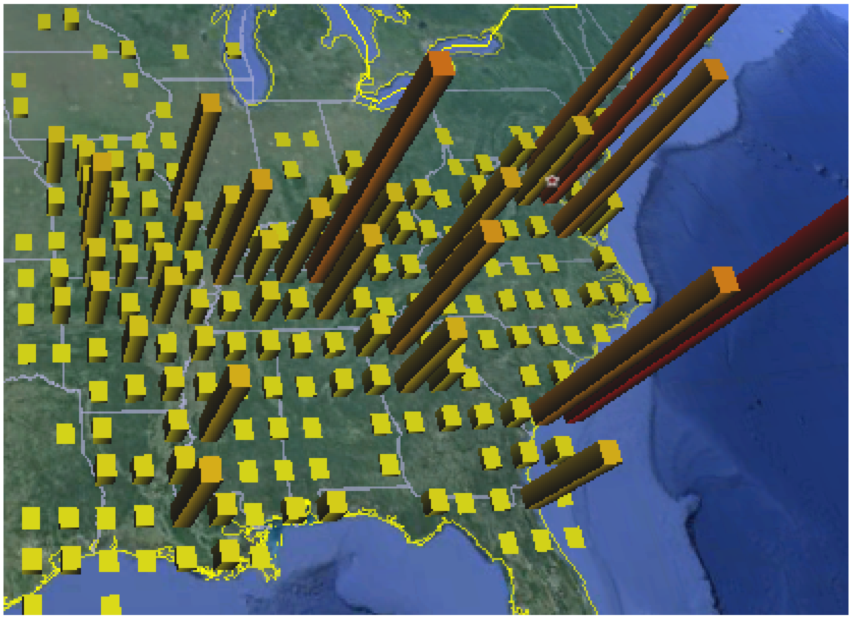


Figure 2.5: KML distribution of the annotated corpus on May 18th.

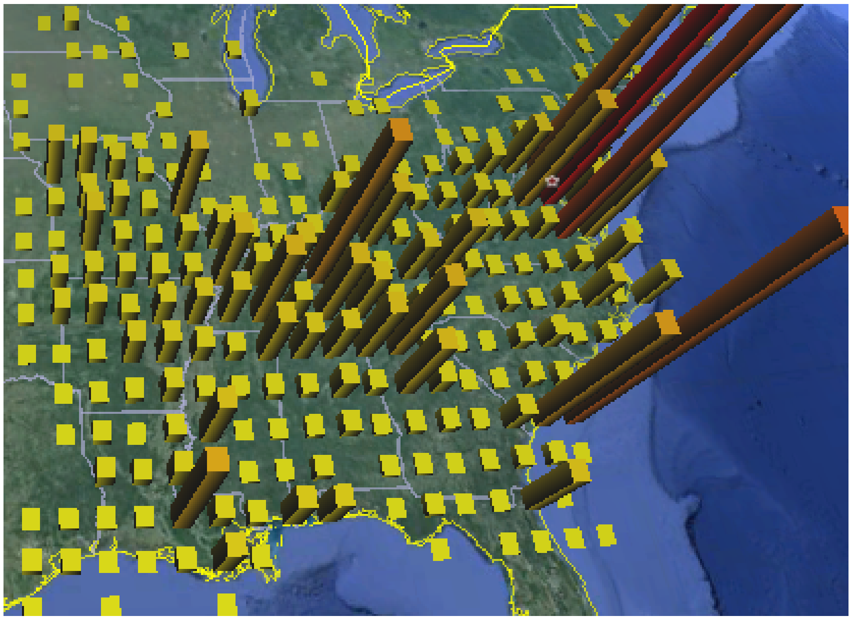


Figure 2.6: KML distribution of the annotated corpus on Jun 10th.

Tennessee and Georgia) starts to fill in. The differences are less apparent in Figure 2.6, but by comparing the two it can be seen that Alabama (especially Northern Alabama), South Carolina, North Carolina, and West Virginia have been significantly filled in. (Keep in mind also that the June 10th graph has been rescaled to keep the highest bar at the same place, causing the heights of all bars to drop by a factor of about 1.5. The total number of annotations increased from approximately 2,500 to 5,000 during this time period.)

## 2.4 Toponym resolution datasets

The following is a description of the datasets used in Chapter 6. These corpora have individual toponym annotations, but most do not have document-level geotags.

**CWAR** is the Perseus Civil War and 19th Century American Collection, a corpus of 341 books (2.5M lines and 58M words of text) from the Perseus Digital Library project (Crane, 2012), primarily concerning the American Civil War. It contains toponym-level annotations, which were generated by a named entity recognizer and then hand-corrected. Approximately 1.1M toponym instances (comparable to word tokens) are annotated with TGN codes (from the Getty Thesaurus of Geographic Names[10]), corresponding to about 56K distinct toponym types. Figure 2.7 shows an example of resolved toponyms in one of the Perseus texts.

Prior to mid-2014, obtaining the full list of latitude/longitude coordinates for TGN codes was difficult, requiring scraping a large number of web pages. As a result, Speriosu (2013) used a partial list of approximately 2,000 common toponyms with both TGN codes and coordinates, corresponding to locations where Union Army units were organized or posted at; with these, he was able to attach coordinates to around 240K toponym instances.

I proceeded to download the full set of TGN codes and produce a new version with all 1.1M toponym instances tagged with coordinates. This was used to produce updated experimental values for the various methods described in Speriosu (2013).

**CWARPORTAL** consists of the subset of articles from the November 4, 2013 English Wikipedia that belong to the Civil War Portal and are in the main namespace. It consists of 4,149

[10] http://www.getty.edu/research/tools/vocabularies/tgn/

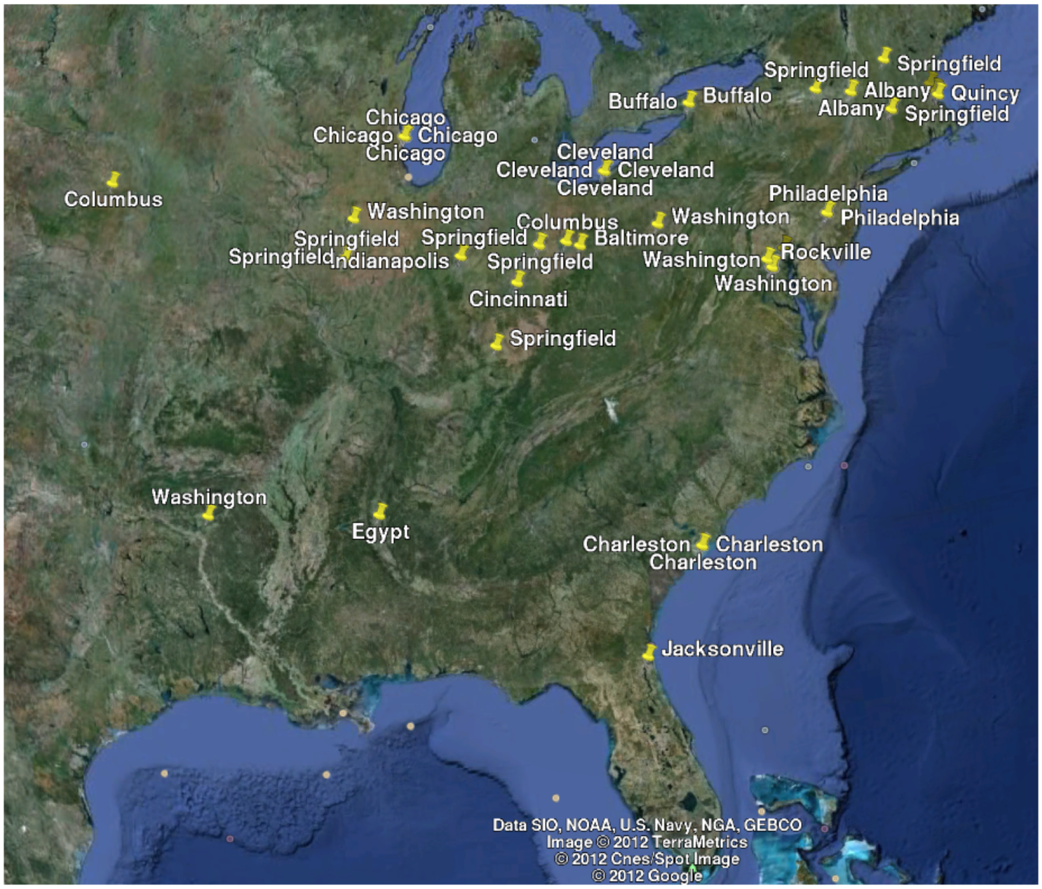


Figure 2.7: KML visualization of predicted locations, situated primarily in the North, for toponyms found towards the end of *Abraham Lincoln: The True Story of a Great Life* (1892), from Speriosu (2013).

articles comprising approximately 4,475,000 words. Of these, only 218 are geotagged. A toponym resolution corpus was created by treating links from a given article to another geolocated article as a toponym, with its candidate set determined by matching the link's anchor text to a gazetteer and the resolved candidate for the toponym determined by the toponym candidate closest to the location of the linked article, as long as it is within 100km (or 500km if the candidate is a state or other higher-level administrative entity, since all such entities are identified by points that may differ between Wikipedia and the gazetteer). Other stretches of text in the same article that match the anchor text of the link are taken as further instances of the same toponym with the same resolution. In addition, links whose anchor text is in the form *CITY, STATE* for states within the United States are converted into two toponyms, one for the city and one for the state, where the correct candidate for the city must be identified in the gazetteer as belong to the state in question, and the correct candidate for the state must be identified as a state within the United States.

**TOPOWIKI13** is a combined document geolocation/toponym resolution dataset that uses the same methods used to create CWARPORTAL, but applied to the entire November 4, 2013 English Wikipedia.

# Chapter 3

# Document geolocation models

## 3.1 Introduction

I implement a number of different ranking models of various levels of complexity and requiring varying degrees of training.[1]

During the evaluation stage, I consider each document in the evaluation set in turn, and produce a ranking of all the grid cells. Normally, I then choose the top-ranked cell and identify its centroid (§3.3) as the "correct" location of the document. In general, given a ranking over all grid cells, it is possible to make use of cells other than the top-ranked to choose the location, and in fact I have implemented the *mean shift* algorithm, which selects the top $K$-ranked cells for some value $K$, and then attempts to cluster them. The idea is that it is possible the top-ranked cell is simply incorrect but the majority of cells near the top are clustered around the correct cell. Preliminary experiments, however, produced results worse than simply selecting the top-ranked cell.

My methods use only the text in the documents; predictions are made based on the distributions $\theta$, $\kappa$, and $\gamma$ introduced in the previous chapter. No use is made of metadata, such as links, followers/friends, or user-declared location (§1.1).

Table 3.1 lists the distributions and other symbols used in the formulas presented in this

[1] This chapter is partly based on Wing (2011), Wing and Baldridge (2011) and Wing and Baldridge (2014). Jason Baldridge was my advisor for these works and helped edit the papers.

| | |
|---|---|
| $\theta_{c_ij}$ | $P(w_j\|c_i)$ = probability of word $w_j$ occurring in cell $c_i$ |
| $\theta_{d_kj}$ | $P(w_j\|d_k)$ = probability of word $w_j$ occurring in document $d_k$ |
| $\theta_{Dj}$ | $P(w_j)$ = overall probability of word $w_j$ occurring across all documents |
| $\theta_{Dj}^{(-d_k)}$ | similar to $\theta_{Dj}$ but the words in document $d_k$ have been assigned zero probability and the remaining probabilities renormalized |
| $\tilde{\theta}_{\cdot j}$ | unsmoothed (maximum likelihood) estimate of word $w_j$ occurring in some context |
| $\kappa_{ji}$ | $P(c_i\|w_j)$ = for a given token of word type $w_j$, probability that $c_i$ is the cell where it occurs |
| $\gamma_i$ | $P(c_i)$ = prior probability of cell $c_i$ occurring, based on the number of documents in the cell |
| $G$ | a geodesic grid, i.e. a division of the Earth's surface into non-overlapping cells |
| $c_i$ | cell number $i$ in grid $G$ |
| $D$ | the set of all documents |
| $d_k$ | document number $k$ in the set of all documents $D$ |
| $V$ | the set of observed vocabulary items |
| $w_j$ | word type (i.e. vocabulary item) number $j$ in vocabulary $V$ |
| $\hat{c}$ | the cell predicted for a given test document |
| $\alpha_{d_k}$ | Good-Turing-style smoothing factor: amount of mass reserved for words unseen in $d_k$ |
| $V_{d_k}$ | the set of observed vocabulary items for document $d_k$ |

Table 3.1: Symbols used in the formulas describing the various geolocation strategies in this chapter.

dissertation.

## 3.2 Grid types

In the context of the general grid-based approach to geolocation followed by this dissertation and described in §1.2, there are several options for constructing the grid and for modeling.

### 3.2.1 Uniform grid

The simplest grid is a *uniform grid* with rectangular cells of equal-sized degrees, such as 1° by 1° or 100 km by 100 km, a strategy followed by Serdyukov et al. (2009) and O'Hare and Murdock (2013) for Flickr, Cheng et al. (2010) and Wing and Baldridge (2011) for Twitter, and Wing and Baldridge (2011) for Wikipedia. Compared to a grid that takes document density into account, it over-represents rural areas at the expense of urban areas. Furthermore, the rectangles are not equal-area, but shrink in width away from the Equator. (However, the shrinkage is mild until near the poles. For example, at 45° latitude, the ratio of width to height is better than 0.7 to 1.)

Figure 1.3 in Chapter 1 shows a choropleth map demonstrating the uniform grid construction. The rank of cells for the test document *Pennsylvania Avenue (Washington, DC)* in ENWIKI13 is plotted, for a uniform 0.1° grid. The top-ranked cell is the correct one. The highest-ranked cells are near Washington, DC, but other culturally similar areas — nearby large cities (Baltimore, Philadelphia, New York City, Pittsburgh) and suburban northern Virginia — are also highly ranked. The importance of certain words in the article is visible in the delineation of the states of Pennsylvania (due to "Pennsylvania" occurring in the article's topic) and Maryland (three-fourths of Pennsylvania Avenue is in Maryland).

A truly equal-area grid can be constructed by means of a quaternary triangular mesh (Dutton, 1996). Dias et al. (2012) used such a construction for Wikipedia, but it did not yield consistently better results. For this reason, as well as ease-of-implementation reasons and the fact that most of the populated regions of interest for this dissertation are far from the poles (where the worst distortion occurs), I construct rectangular grids.

### 3.2.2 Adaptive *k*-d tree grid

Roller, Speriosu, Rallapalli, Wing and Baldridge (2012) introduced an adaptive grid based on *k*-d trees (Bentley, 1975), which I make use of in this dissertation. The idea is to use variable-sized cells so that the number of documents per cell is approximately the same. A *k*-d tree in 2 dimensions starts out with a single grid cell and adds documents to this cell one by one. When the number of documents reaches a threshold termed the *bucket size*, the cell is split in two along the dimension with the greatest range of points seen, following Friedman et al. (1977). Roller et al. (2012) considered splitting at either the midpoint of the range of points or at the median of the dimension in question for all points in the cell, and found that neither method was clearly superior. In my preliminary experiments I found midpoint splitting to work at least as well, and I use that in my subsequent experiments.[2]

[2] But see Figure 5.5 for learning-curve experiments performed using median splitting.

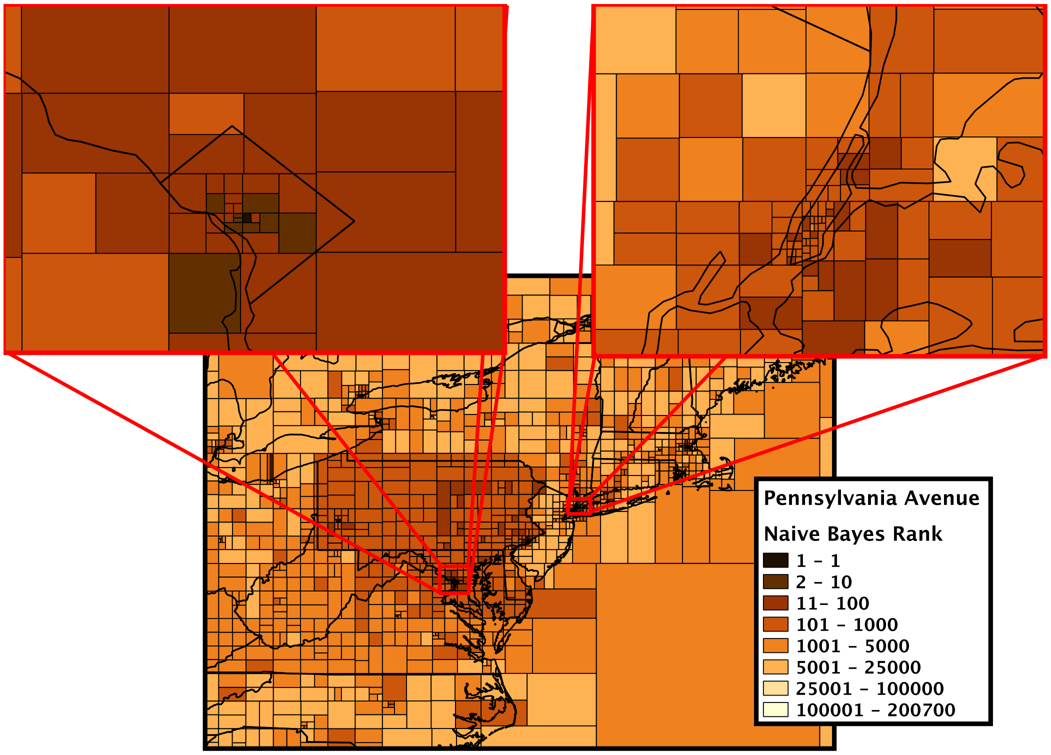


Figure 3.1: *k*-d tree grid construction. Relative Naive Bayes rank is shown for cells for EnWiki13 test document *Pennsylvania Avenue (Washington, DC)*, surrounding the true location. (Constructed with assistance from Grant DeLozier.)

Figure 3.1 shows a sample *k*-d tree grid in the form of a choropleth map. Increased cell density with correspondingly smaller cells occurs on land compared with over the sea, especially in coastal regions of the Northeast of the United States. Map callouts zoom in on Washington, DC and New York City, showing the particularly increased concentration of cells in city centers.

### 3.2.3 City-based grid

Some researchers have used a city-based representation, either with a full set of cities covering the Earth and taken from a comprehensive gazetteer (Han et al., 2014) or a limited, pre-specified set of cities (Kinsella et al., 2011; Sadilek et al., 2012). This is somewhat comparable to *k*-d trees in that it adapts to areas of greater population. Han et al. (2014)'s construction, for example, determines a set of *city attractors* by reducing the total set of cities in a gazetteer through amalgamating cities into nearby larger cities in the same second-level administrative district (in the same state, in the case of the United States). Training documents are then assigned to a pseudo-document corresponding to the nearest city. An even more direct method would use census-tract boundaries when available.

An advantage of city-based grids compared especially with coarser-scale rectangular grids is that in the latter, the boundary between cells may run through the middle of a city. This has the effect of splitting a presumably unitary linguistic area, and grouping the different parts of the city with the heterogeneous linguistic areas of other cities. For example, a coarse grid that passes through the middle of Austin, Texas might group one half with San Antonio and the other half with Houston, making it more difficult to correctly geolocate a document whose location is in Austin. The resulting statistical bias is known as the *modifiable areal unit problem* (Gehlke and Biehl, 1934; Openshaw, 1983). With finer grids, however, this is less likely to be an issue. It is also possible to mitigate this issue in *k*-d trees by dividing a cell in such a way as to produce the maximum margin between the dividing line and the nearest document on each side. (This was implemented in Roller et al. (2012)'s code but not investigated in their paper.)

A disadvantage of city-based grids is that they are unable to resolve locations at a finer scale than an entire city, whereas rectangular grids can be made as fine-scale as desired. This is a particular advantage of *k*-d trees, which will naturally increase their resolution in the vicinity of populated regions, leading to grids that may be able to distinguish cities from suburbs or even identify individual neighborhoods in a city, as shown in Figure 3.1.

Other disadvantages of these methods are the dependency on time-specific population data, making them unsuitable for some corpora (e.g. 19th-century documents); the difficulty in adjusting grid resolution in a principled fashion; and the fact that not all documents are near a city. Han et al. (2014) in fact find that 8% of tweets are "rural" and cannot predicted by their model. This may be worse for Wikipedia, which includes coverage of many small towns and villages.

For these reasons, I do not consider city-based grids in my experiments.

## 3.3 Grid construction

With such a discrete representation of the earth's surface, there are four distributions that form the core of all my geolocation methods. The first is a standard multinomial distribution over the vocabulary for every cell in the grid. Given a grid $G$ with cells $c_i$ and a vocabulary $V$ with words $w_j$, we have $\theta_{c_i j} = P(w_j|c_i)$. The second distribution is the equivalent distribution for a single test document $d_k$ with vocabulary $V_{d_k}$, i.e. $\theta_{d_k j} = P(w_j|d_k)$. The third distribution is the reverse of the first: for a given word, its distribution over the earth's cells, $\kappa_{ji} = P(c_i|w_j)$. The final distribution is over the cells, $\gamma_i = P(c_i)$.

The first and second distributions are in fact particular types of *language models*, i.e. methods of assigning a probability to a sequence of words. Specifically, the first distribution is a language model of a document, and the second one is a model of the concatenation of all training documents within a given cell. For the purposes of this dissertation, I use a simple unigram model that ignores word ordering. As a result, it will have difficulty when presented with a multiword toponym such as the Texas city of *College Station*. In Chapter 7 this issue is addressed further.

The grid representation I use ignores all higher level regions, such as states, countries, rivers, and mountain ranges, but is consistent with the geocoding in both the Wikipedia and Twitter datasets. Note that the $\kappa_{ji}$ for words referring to such regions is likely to be quite flat (spread out) but with most of the mass concentrated in a set of connected cells. Those for highly focused point-locations will jam up in a few disconnected cells—in the extreme case, toponyms like *Springfield* which are connected to many specific point locations around the earth.

I use grids with cell sizes of varying granularity $d \times d$. For example, with $d{=}0.5°$, a cell at the equator is roughly 56x55 km and at 45° latitude it is 39x55 km. At this resolution, there are a total of 259,200 cells, of which 35,750 are non-empty when using the ENWIKI13 training set. For comparison, at the equator a cell at $d{=}5°$ is about 557x553 km (2,592 cells; 1,747 non-empty) and at $d{=}0.1°$ a cell is about 11.3x10.6 km (6,480,000 cells; 170,005 non-empty).

The geolocation methods predict a cell $\hat{c}$ for a document, and the latitude and longitude of the centroid of the cell (the mean of all observed points in the cell in the training data) is used as the predicted location. (Wing and Baldridge (2011) used the midpoint of the cell, but better results stem from using the centroid, which often reflects the location of the major city or area of concentration within the grid cell, as shown by Roller et al. (2012).) Prediction error is the great-circle distance from these predicted locations to the locations given by the gold standard. This differs from the evaluation metrics used by Serdyukov et al. (2009), which are all computed relative to a given grid size. With their metrics, results for different granularities cannot be directly compared because using larger cells means less ambiguity when choosing $\hat{c}$. With distance-based evaluation, large cells are penalized by the distance from the centroid to the actual location even when that location is in the same cell. Smaller cells reduce this penalty and permit the word distributions $\theta_{c_i j}$ to be much more specific for each cell, but they are harder to predict exactly and suffer more from sparse word counts compared to courser granularity. For large datasets like the English Wikipedia, fine-grained grids work very well, but the trade-off between resolution and sufficient training material shows up more clearly for GEOTEXT (the small Twitter dataset). See §4.1.2 for a fuller discussion of evaluation metrics.

## 3.4 Information retrieval models

A geodesic grid of sufficient granularity creates a large decision space, when each cell is viewed as a label to be predicted by some classifier. This situation naturally lends itself to simple, scalable language-modeling approaches, motivated by the techniques used in information retrieval. For this general strategy, each cell is characterized by a *pseudo-document* constructed from the concatenation of the training documents that it contains. A test document's location is then chosen based on the cell with the most similar language model according to standard measures such as Kullback-Leibler (KL) divergence (Zhai and Lafferty, 2001), which seeks the cell whose language model is closest to the test document's, or Naive Bayes (Lewis, 1998), which chooses the cell that assigns the highest probability to the test document, according to Bayes' Law.

These models are quick to train, which allows them to expand to encompass fine-scale grid resolutions with potentially thousands or even hundreds of thousands of non-empty grid cells to choose among.

Other scalable models I implemented are what I term Average Cell Probability (ACP), which inverts the set of grid-cell language models to produce distributions over grid cells for a given word and averages the distributions of the test document's words; cosine similarity; TF/IDF; and some very basic baselines, such as selecting a random cell or always choosing the cell containing the greatest number of training documents. In preliminary experiments, I did not get good results from cosine similarity or TF/IDF, and do not consider them further.

### 3.4.1 Training

The training material specifies the location of each document. Using that, I aggregate documents into grid cells, from which I acquire $\theta$ and $\kappa$ straightforwardly.

#### Word distributions

The unsmoothed estimate of word $w_j$'s probability in a test document $d_k$ is:[3]

$$\tilde{\theta}_{d_k j} = \frac{\#(w_j, d_k)}{\sum\limits_{w_l \in V} \#(w_l, d_k)} \tag{3.1}$$

Similarly for a cell $c_i$, I compute the unsmoothed word distribution by aggregating all of the documents located within $c_i$:

[3] I use $\#()$ to indicate the count of an event.

$$\tilde{\theta}_{c_i j} = \frac{\sum\limits_{d_k \in c_i} \#(w_j, d_k)}{\sum\limits_{d_k \in c_i} \sum\limits_{w_l \in V} \#(w_l, d_k)} \tag{3.2}$$

I compute the global distribution $\theta_{Dj}$ over the set of all documents $D$ in the same fashion.

To compute the smoothed word distribution of a document $d_k$, I can either interpolate the global distribution $\theta_{Dj}$, or back off to it when a word is not seen in the document's distribution. A general interpolation model looks like

$$\theta_{d_k j} = (1 - \lambda_{d_k}) \tilde{\theta}_{d_k j} + \lambda_{d_k} \theta_{Dj} \tag{3.3}$$

where the *discount factor* $\lambda_{d_k}$ indicates how much probability mass to reserve for unseen words. I consider two possibilities. *Jelinek smoothing* simply sets $\lambda_{d_k}$ to a constant value, while *Dirichlet smoothing* assigns it as follows:

$$\lambda_{d_k} = 1 - \frac{|d_k|}{|d_k| + m} \tag{3.4}$$

where $|d_k|$ is the size of the document and $m$ is a tunable parameter. This has the effect of relying more on $d_k$'s distribution and less on the global distribution for larger documents that provide more evidence than shorter ones.

A general back-off model looks like

$$\theta_{Dj}^{(-d_k)} = \frac{\theta_{Dj}}{1 - \sum\limits_{w_l \in d_k} \theta_{Dl}} \tag{3.5}$$

$$\theta_{d_k j} = \begin{cases} \alpha_{d_k} \theta_{Dj}^{(-d_k)}, & \text{if } \tilde{\theta}_{d_k j} = 0 \\ (1 - \alpha_{d_k}) \tilde{\theta}_{d_k j}, & \text{o.w.} \end{cases} \tag{3.6}$$

where $\theta_{Dj}^{(-d_k)}$ is an adjusted version of $\theta_{Dj}$ that is normalized over the subset of words not found in document $d_k$. This adjustment ensures that the entire distribution is properly normalized.

$\alpha_{d_k}$ is the probability mass reserved for unseen words. I set it using a non-parametric method I devised called *pseudo-Good-Turing*. Motivated by Good-Turing smoothing, I determine $\alpha_{d_k}$ by the empirical probability of having seen a word once in the document:

$$\alpha_{d_k} = \frac{|w_j \in V\, s.t.\, \#(w_j, d_k){=}1|}{\sum\limits_{w_j \in V} \#(w_j, d_k)} \tag{3.7}$$

$$\tag{3.8}$$

For whichever smoothing method I use, the cell distributions are treated analogously.

### Cell distributions

The distributions over cells for each word simply renormalize the $\theta_{c_ij}$ values to achieve a proper distribution:

$$\kappa_{ji} = \frac{\theta_{c_ij}}{\sum\limits_{c_i \in G} \theta_{c_ij}} \tag{3.9}$$

A useful aspect of the $\kappa$ distributions is that they can be plotted in a geobrowser using thematic mapping techniques (Sandvik, 2008) to inspect the spread of a word over the earth. I used this as a simple way to verify the basic hypothesis that words that do not name locations are still useful for geolocation. Figure 3.2 is an example of the Wikipedia distribution for *mountain*, plotted using Google Earth[4]; Figure 3.3 is a similar plot for *beach*. Not surprisingly, it shows high density over the Rocky Mountains, Smokey Mountains, the Alps, and other ranges.[5] Similarly, *beach* has high density in coastal areas. Words without inherent locational properties also have intuitively correct

[4] `http://earth.google.com`

[5] The red line in the upper right corresponds to Little Hall Island off the coast of Baffin Island, in northern Canada. The color stems from Google Earth's lighting algorithm, but it is unclear why this particular location is so significant for the word *mountain*.

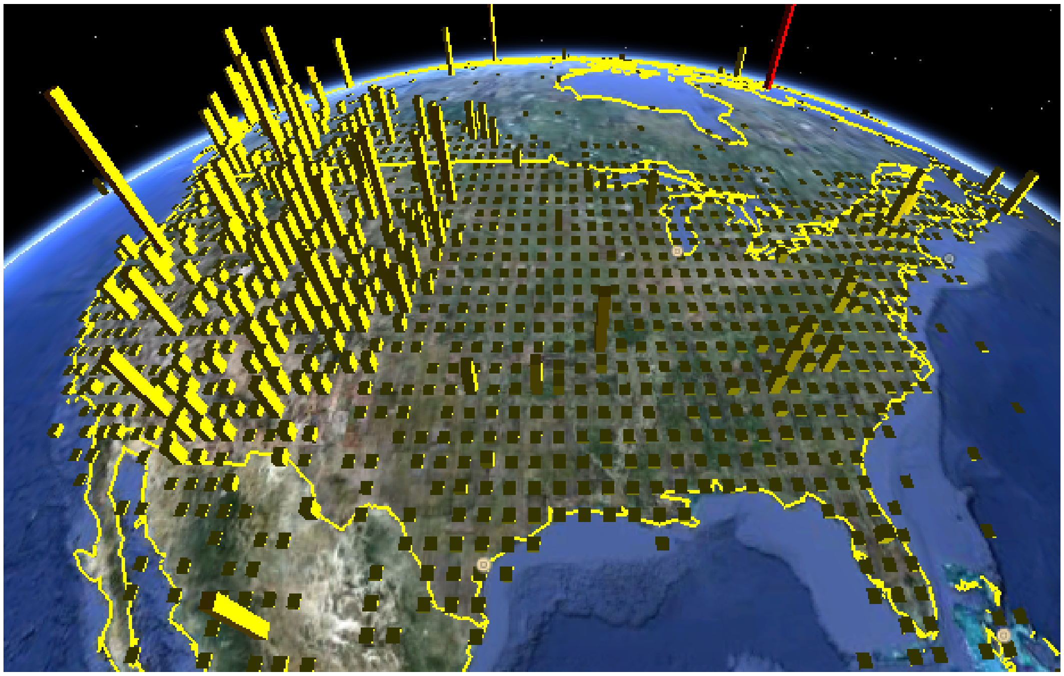


Figure 3.2: Wikipedia distribution of *mountain*, plotted using Google Earth.

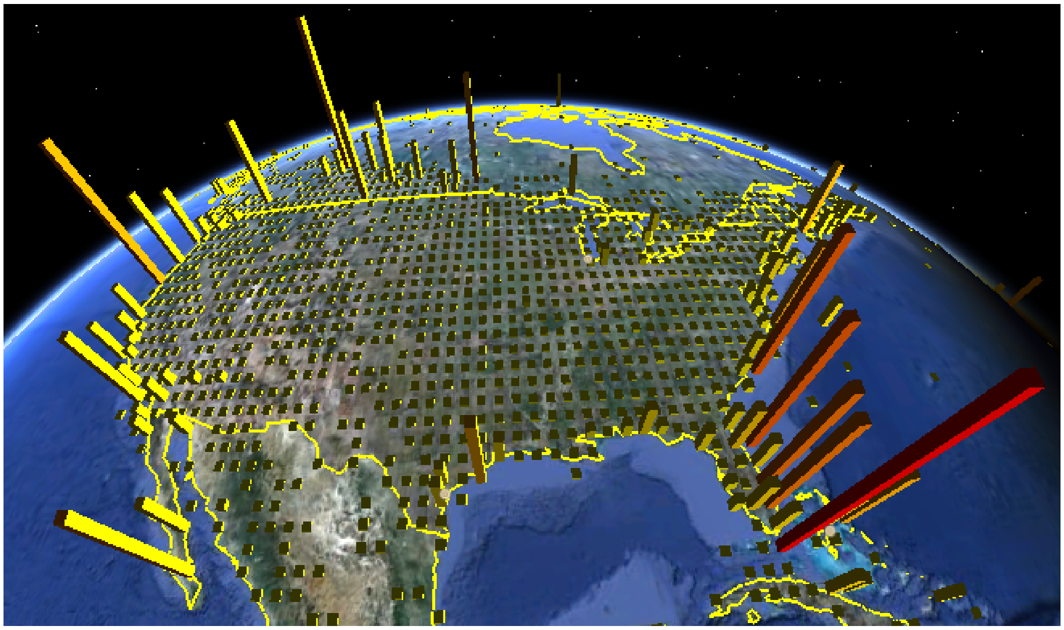


Figure 3.3: Wikipedia distribution of *beach*, plotted using Google Earth.

distributions: e.g., *barbecue* has high density over the south-eastern United States, Texas, Jamaica, and Australia, while *wine* is concentrated in France, Spain, Italy, Chile, Argentina, California, South Africa, and Australia.[6]

Finally, the cell distributions are simply the relative frequency of the number of documents in each cell: $\gamma_i = \frac{|c_i|}{|D|}$.

A standard set of stopwords are ignored. Also, all words are lowercased except in the case of the most-common-toponym baselines, where uppercase words serve as a fallback in case a toponym cannot be located in the article.

### 3.4.2 Kullback-Leibler divergence

Given the distributions for each cell, $\theta_{c_i}$, in the grid, I use an information retrieval approach to choose a location for a test document $d_k$: compute the similarity between its word distribution $\theta_{d_k}$ and that of each cell, and then choose the closest one. Kullback-Leibler (KL) divergence is a natural choice for this (Zhai and Lafferty, 2001). For distribution $P$ and $Q$, KL divergence is defined as:

$$KL(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{3.10}$$

This quantity measures how good $Q$ is as an encoding for $P$—the smaller it is the better. The best cell $\hat{c}_{KL}$ is the one which provides the best encoding for the test document:

$$\hat{c}_{KL} = \operatorname*{arg\,min}_{c_i \in G} KL(\theta_{d_k}||\theta_{c_i}) \tag{3.11}$$

The fact that KL is not symmetric is desired here: the other direction, $KL(\theta_{c_i}||\theta_{d_k})$, asks which cell the test document is a good encoding for. With $KL(\theta_{d_k}||\theta_{c_i})$, the log ratio of probabilities for each word is weighted by the probability of the word in the test document, $\boldsymbol{\theta_{d_k j}} \log \frac{\theta_{d_k j}}{\theta_{c_i j}}$, which means that the divergence is more sensitive to the document rather than the overall cell.

As an example for why non-symmetric KL in this order is appropriate, consider geolocating a page in a densely geotagged cell, such as the page for the Washington Monument. The distribution

[6] This also acts as an exploratory tool. For example, due to a big spike on Cebu Province in the Philippines I learned that Cebuanos take barbecue very, very seriously.

of the cell containing the monument will represent the words from many other pages having to do with museums, US government, corporate buildings, and other nearby memorials and will have relatively small values for many of the words that are highly indicative of the monument's location. Many of those words appear only once in the monument's page, but this will still be a higher value than for the cell and will weight the contribution accordingly.

Rather than computing $KL(\theta_{d_k}||\theta_{c_i})$ over the entire vocabulary, I restrict it to only the words in the document to compute KL more efficiently:

$$KL(\theta_{d_k}||\theta_{c_i}) = \sum_{w_j \in V_{d_k}} \theta_{d_k j} \log \frac{\theta_{d_k j}}{\theta_{c_i j}} \tag{3.12}$$

Early experiments showed that it makes no difference in the outcome to include the rest of the vocabulary. Note that because $\theta_{c_i}$ is smoothed, there are no zeros, so this value is always defined.

### 3.4.3 Naive Bayes

Naive Bayes is a natural generative model for the task of choosing a cell $c_i$, given the distributions $\theta_{c_i}$ and $\gamma$. To generate a document $d_k$, choose a cell $c_i$ according to $\gamma$ and then choose the words in the document according to $\theta_{c_i}$:

$$\begin{aligned}
\hat{c}_{NB} &= \arg\max_{c_i \in G} P_{NB}(c_i|d_k) \\
&= \arg\max_{c_i \in G} \frac{P(c_i)P(d_k|c_i)}{P(d_k)} \\
&= \arg\max_{c_i \in G} \gamma_i \prod_{w_j \in V_{d_k}} \theta_{c_i j}^{\#(w_j, d_k)} \qquad (3.13)
\end{aligned}$$

This method maximizes the combination of the *likelihood* of the document $P(d_k|c_i)$ and the *cell prior probability* $\gamma_i$.

### 3.4.4 Average cell probability

For each word, $\kappa_{ji}$ gives the probability of each cell in the grid. A simple way to compute a distribution for a document $d_k$ is to take a weighted average of the distributions for all words to compute the average cell probability (ACP):

$$\begin{aligned}
\hat{c}_{ACP} &= \arg\max_{c_i \in G} P_{ACP}(c_i|d_k) \\
&= \arg\max_{c_i \in G} \frac{\sum\limits_{w_j \in V_{d_k}} \#(w_j, d_k)\kappa_{ji}}{\sum\limits_{c_l \in G} \sum\limits_{w_j \in V_{d_k}} \#(w_j, d_k)\kappa_{jl}} \\
&= \arg\max_{c_i \in G} \sum_{w_j \in V_{d_k}} \#(w_j, d_k)\kappa_{ji} \qquad (3.14)
\end{aligned}$$

This method, despite its conceptual simplicity, works well in practice. It could also be easily modified to use different weights for words, such as TF/IDF or relative frequency ratios between geolocated documents and non-geolocated documents.

## 3.5 Logistic regression

The use of discrete cells over the Earth's surface allows any classification strategy to be employed, including discriminative classifiers such as logistic regression (also known as *maximum entropy modeling*). Logistic regression does not assume that the set of features are independent, as does Naive Bayes, but instead learns how to properly weight the features, automatically down-weighting those that largely duplicate the evidence supplied by other features. As a result, logistic regression often produces better results than generative classifiers at the cost of more time-consuming training, which limits the size of the problems it may be applied to. Training is generally unable to scale to encompass several thousand or more distinct labels, as is the case with fine-scale grids of the sort I employ in various models. Nonetheless I find flat logistic regression to be effective on most of my large-scale corpora, and the hierarchical classification strategy discussed in §3.7 allows me to take advantage of logistic regression without incurring such a high training cost.

Logistic regression, when operating as a binary classifier, models the log-odds of the probability of a positive response as a linear combination of the input features and a set of learned weights. When doing multi-way classification, the normal procedure is to identify one of the possible labels as a pivot and model the log of the probability ratio of seeing each of the other classes relative to the pivot as a linear combination of features and separate sets of weights.

A $K$-way logistic regression classifier is normally written as

$$\begin{aligned}
\ln p(y_i = 1) &= \beta_1 \cdot x_i - \ln Z \\
\ln p(y_i = 2) &= \beta_2 \cdot x_i - \ln Z \\
&\ldots \\
\ln p(y_i = K) &= \beta_k \cdot x_i - \ln Z
\end{aligned} \tag{3.15}$$

for a multinomial label $y_i$, with a set of weights $\beta_k$ for $k = 1 \ldots K$, where each choice has its own weight vector, and a normalizing constant $Z$ is introduced so that the probabilities all sum to 1:

$$\sum_{k=1}^{K} p(y_i = k) = 1 \tag{3.16}$$

## 3.6 Feature selection

Naive Bayes assumes that features are independent, which penalizes models that must accommodate many features that are poor indicators and which can gang up on the good features. Large improvements have been obtained by reducing the set of words used as features to those that are geographically salient. Cheng et al. (2010; 2013) model word locality using a unimodal distribution taken from Backstrom et al. (2008) and train a classifier to identify geographically local words based on this distribution. This unfortunately requires a large hand-annotated corpus for training. Han et al. (2014) systematically investigate various feature selection methods for finding geo-indicative words, such as information gain ratio (IGR) (Quinlan, 1993), Ripley's K statistic (O'Sullivan and Unwin, 2010) and geographic density (Chang et al., 2012), showing significant improvements on

TWUS and TWWORLD (Chapter 2).

Both papers use information retrieval methods for doing the actual geolocation. Han et al. (2014) compare Naive Bayes with KL divergence, while Cheng et al. use a method similar to my ACP method (§3.4.4).

For comparison with Han et al. (2014), I test against an additional baseline: Naive Bayes combined with feature selection done using IGR. Following Han et al., I first eliminate words which occur less than 10 times, have non-alphabetic characters in them or are shorter than 3 characters. I then compute the IGR for the remaining words across all cells at a given cell size or bucket size, select the top $N\%$ for some *cutoff percentage* $N$ (which I vary in increments of 2%), and then run Naive Bayes at the same cell size or bucket size.

## 3.7 Hierarchical classification

To overcome the limitations of discriminative classifiers in terms of the maximum number of cells they can handle, I introduce hierarchical classification (Silla Jr. and Freitas, 2011) for geolocation. Dias et al. (2012) use a simple two-level generative hierarchical approach using Naive Bayes, but to my knowledge no previous work implements a multi-level discriminative hierarchical model with beam search for geolocation.

To construct the hierarchy, I start with a root cell $c_{root}$ that spans the entire Earth and from there build a tree of cells at different scales, from coarse to fine. A cell at a given level is subdivided to create smaller cells at the next level of resolution that altogether cover the same area as their parent.

I use the *local classifier per parent* approach to hierarchical classification (Silla Jr. and Freitas, 2011) in which an independent classifier is learned for every node of the hierarchy above the leaf nodes. The probability of any node in the hierarchy is the product of the probabilities of that node and all of its ancestors, up to the root. This is defined recursively as:

$$\begin{aligned} P(c_{root}) &= 1.0 \\ P(c_j) &= P(c_j|\uparrow c_j)P(\uparrow c_j) \end{aligned} \tag{3.17}$$

where $\uparrow c_j$ indicates $c_j$'s parent in the hierarchy.

In addition to allowing one to use many classifiers that each have a manageable number of outcomes, the hierarchical approach naturally lends itself to beam search. Rather than computing the probability of every leaf cell using equation 3.17, I use a stratified beam search: starting at the root cell, keep the $b$ highest-probability cells at each level until reaching the leaf node level. With a tight beam—which I show to be very effective—this dramatically reduces the number of model evaluations that must be performed at test time.

For example, if I have 40 cells at level 1, and each level-$L$ cell subdivides into 4 cells at level $L + 1$, and I have a beam size of 8, then I proceed as follows:

1. Run the top-level classifier over the 40 level-1 cells.

2. Select the 8 highest-probability cells at level 1.

3. For each such cell, run the classifier associated with this cell, which yields probabilities over the 4 subdivided cells at level 2; combine them with the level-1 cell's probability to get a total probability for 32 level-2 cells.

4. Select the 8 highest-probability cells at level 2.

5. Repeat if there are any more levels.

**Grid size parameters** Two factors determine the size of the grids at each level. The first-level grid is constructed the same as for Naive Bayes or flat logistic regression and is controlled by its own parameter. In addition, the *subdivision factor* $N$ determines how I subdivide each cell to get from one level to the next. Both factors must be optimized appropriately.

For the uniform grid, I subdivide each cell into $NxN$ subcells. In practice, there may actually be fewer subcells, because some of the potential subcells may be empty (contain no documents).

For the *k*-d grid, if level 1 is created using a bucket size $B$ (i.e. I recursively divide cells as long as their size exceeds $B$), then level 2 is created by continuing to recursively divide cells that exceed a smaller bucket size $B/N$. At this point, the subcells of a given level-1 cell are the leaf

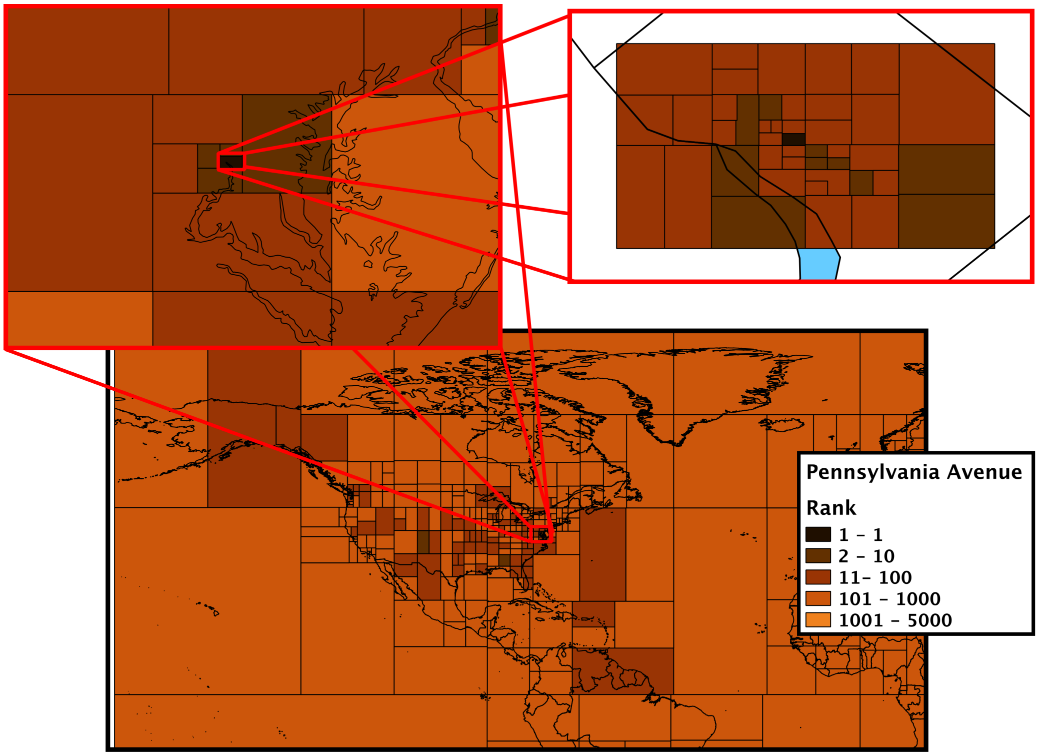


Figure 3.4: Relative hierarchical LR rank of cells for EnWiki13 test document *Pennsylvania Avenue (Washington, DC)*, surrounding the true location. The first callout simply expands a portion of level 1, while the second callout shows a level 1 cell subdivided down to level 2.

cells contained with the cell's geographic area. The construction of level 3 proceeds similarly using bucket size $B/N^2$, etc.

Note that the subdivision factor has a different meaning for uniform and *k*-d tree grids. Furthermore, because creating the subdividing cells for a given cell involves dividing by $N^2$ for the uniform grid but $N$ for the *k*-d tree grid, greater subdivision factors are generally required for the *k*-d tree grid to achieve similar-scale resolution.

Figure 3.4 shows the behavior of hierarchical LR using *k*-d trees for the test document *Pennsylvania Avenue (Washington, DC)* in EnWiki13. After ranking the first level, the beam zooms in on the top-ranked cells and constructs a finer *k*-d tree under each one (one such subtree is shown in the top-right map callout).

## 3.8 Simple baselines

There are several natural baselines to use for comparison against the methods described above.

**Random** Choose $\hat{c}_{rand}$ randomly from a uniform distribution over the entire grid $G$.

**Cell prior maximum** Choose the cell with the highest prior probability according to $\gamma$: $\hat{c}_{cpm} = \arg\max_{c_i \in G} \gamma_i$.

**Most frequent toponym** Identify the most frequent toponym in the article and the geotagged Wikipedia articles that match it. Then identify which of those articles has the most incoming links (a measure of its prominence), and then choose $\hat{c}_{mft}$ to be the cell that contains the geotagged location for that article. This is a strong baseline method, but can only be used with Wikipedia.

Note that a toponym matches an article (or equivalently, the article is a candidate for the toponym) either if the toponym is the same as the article's title, or the same as the title after a parenthetical tag or comma-separated higher-level division is removed. For example, the toponym *Tucson* would match articles named *Tucson*, *Tucson (city)* or *Tucson, Arizona*. In this fashion, the set of toponyms, and the list of candidates for each toponym, is generated from the set of all geotagged Wikipedia articles.

I implemented a fourth baseline, but ended up not using it in my final experiments. It chooses the most frequent toponym in the article and then chooses the grid cell with the maximum probability for this word, according to $\kappa_{ji}$ (see §3.4.4 above). If no toponym is found in an article, it falls back to the most frequent capitalized word. In my early experiments on Wikipedia, it consistently performed worse than the *most frequent toponym* strategy described above that uses incoming-link prominence. However, this strategy might be useful as a baseline for a corpus where such a prominence measure is unavailable, as in the Twitter corpus I use for evaluation.

# Chapter 4

# Experiments on modern corpora

The eventual goal of this dissertation is to apply the methods of the previous chapter to historical corpora in the digital humanities.[1] These corpora, however, tend to be small and to lack document-level or paragraph-level annotations, which makes it difficult to evaluate their performance and requires us to develop domain-adaptation techniques for the annotated material that we do have. We first need to understand the performance of these methods standing by themselves, and we do so by training and evaluating them on modern corpora where we have plenty of in-domain material. As described in Chapter 2, we evaluate on three types of corpora: Twitter user feeds, Wikipedia articles, and Flickr image tags.

## 4.1 Experimental setup

### 4.1.1 Configurations

The most important parameters in my experiments are those related to *grid construction* and *grid scoring*. Additional parameters cover *choice of representative point*, *smoothing*, *filtering* and *logistic regression*.

[1] This chapter is partly based on Wing (2011), Wing and Baldridge (2011) and Wing and Baldridge (2014). Jason Baldridge was my advisor for these works and helped edit the papers.

**Grid construction** For grid construction, the possibilities are either a **uniform** or ***k*-d** tree grid. For uniform grids, the main tunable parameter is *grid size* (in **degrees**), while for *k*-d trees it is *bucket size* (**BK**), i.e. the number of documents above which a node is divided in two.

**Grid scoring** For grid scoring, the options are:

- **RAND**: Random baseline
- **PRIOR**: Cell prior maximum
- **NB**: Naive Bayes
- **KL**: KL divergence
- **ACP**: Average cell probability
- **IGR**: Naive Bayes using features selected by information gain ratio
- **FLATLR**: Logistic regression model over all leaf nodes
- **HIERLR**: Product of logistic regression models at each node in a hierarchical grid (eq. 3.17)

Some of these methods are associated with additional parameters, which must be tuned on the dev set:

- For IGR, there is one additional parameter, the *cutoff* (CU), a percentile. For a given value $c$, we eliminate the bottom $(100 - c)\%$ of words, as measured by information gain ratio.
- For HIERLR, there are three additional parameters: subdivision factor (**SF**), beam size (**BM**), and hierarchy depth (**D**). See §3.7 and §4.2.5 for more discussion. All of our test-set results use a depth of three levels.

**Choice of representative point** Once grid cells have been scored, a single point representing the top-ranked cell needs to be chosen. This can be done using the geographic center of a cell or the centroid of the training documents in the cell. The latter produces consistently better results and is used in further experiments (§3.3), but has a significant dependence on the particular set of training documents, which especially matters when this set is small (§5.2.2). Another possibility is to take into account cells further down in the ranking, using an algorithm such as *mean shift* (§1.2, §7.2.2), although preliminary experiments with this algorithm were not promising.

**Smoothing** As discussed in §3.4.1, I consider three types of smoothing of language models: *Dirichlet*, *Jelinek*, and my own method *pseudo-Good-Turing*. Based on preliminary experiments, I choose Dirichlet smoothing in conjunction with Naive Bayes, with the Dirichlet parameter set to $m = 1,000,000$. For KL divergence, I did not have good luck with Dirichlet smoothing, and instead use pseudo-Good-Turing, which has no tunable parameter.

**Filtering** For the most part, I do not pre-filter words out of a language model, except for applying standard language-dependent sets of stopwords. Some methods that I compare against, however (e.g. GEOTEXT, §4.2.1), do pre-filter words, and I investigate whether this is needed.

**Logistic regression** Due to its speed and flexibility, I use Vowpal Wabbit (Agarwal et al., 2014) for logistic regression.[2] I estimate parameters with limited-memory BFGS (Nocedal, 1980; Byrd et al., 1995), as I found that stochastic gradient descent (SGD) (Bottou, 2010) yielded significantly worse results.[3] Unless otherwise mentioned, I use 26-bit feature hashing (Weinberger et al., 2009) and 40 passes over the data (optimized based on early experiments on development data). For the subcell classifiers in hierarchical classification, which have fewer classes and much less data, I use 24-bit features and 12 passes.

Vowpal Wabbit has a hold-out mechanism, which holds out a portion of the training data and uses it to determine when to stop training, to avoid potential overfitting problems. I turn this mechanism off due to poor performance with it enabled. This means I have to carefully optimize the number of passes using the dev set, to avoid both underfitting (not enough passes) and overfitting (too many passes), both of which cause significant decreases in accuracy. This is in contrast to the

[2] I also investigated some other tools, including the `mlogit` package of R (Croissant, 2013) and Rob Malouf's TADM (Tools for Advanced Data Modeling) package (Malouf, 2002).

[3] SGD holds out the promise of being faster than BFGS. However, I found that attempting to tune SGD to achieve similar results to BFGS produced even slower running times than BFGS. One possibility I did not consider, which may produce comparable accuracy and faster running time, was to use SGD to produce a preliminary solution and optimize further with BFGS.

|  | Feature bits |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Passes | 22 | 23 | 24 | 25 | 26 | 27 |
| 16 | 394 | 355 | 363 | 380 | 390 | 391 |
| 24 | 346 | 309 | 287 | 302 | 299 | 287 |
| 32 | 277 | 266 | 250 | 259 | 254 | 257 |
| 40 | 267 | 259 | 256 | **247** | 249 | 255 |
| 48 | 275 | 266 | 267 | 254 | 254 | 253 |
| 64 | 301 | 281 | 286 | 286 | 276 | 277 |

Table 4.1: Median prediction error (km) on the TWUS dev set for various combinations of feature-hashing bit size and number of BFGS passes.

number of bits used for feature hashing, where it is merely necessary to use a large enough feature space to avoid clashes, and using more bits than necessary does not materially hurt performance.

The effect of different numbers of feature bits and passes can be seen in Table 4.1, which shows median prediction error on TWUS-LARGE with a uniform 5° grid under FLATLR. In this case 25 bits is slightly better than 26, but in other experiments (e.g. in HIERLR, and for ENWIKI13, which has more features) I found better performance from 26 bits, which is what I ultimately selected.

### 4.1.2 Evaluation metrics

A number of different evaluation metrics have been used by various authors to gauge the performance of geolocation. Serdyukov et al. (2009) used various cell-based accuracy metrics, measuring the fraction of documents successfully geolocated to the correct cell, or to a square of $K$ cells surrounding the correct cell. This is a simple and accessible metric but has the disadvantage that it is sensitive to the size of the cell grid, making comparisons across different-sized grids difficult.

Serdyukov et al. (2009) also use mean reciprocal rank, commonly used in the learning-to-rank community (Liu, 2011), which measures the accuracy of an entire ranking, including those cells ranked below the best cell. This measure has the same flaws as cell accuracy. In addition, its emphasis on the entire ranking makes it fundamentally different from the other metrics considered here. This would be useful in e.g. a context where the user is presented with a number of possible locations and asked to select one; typically, however, the goal is to find the correct location. This minimizes the need for user-based assistance and allows for a much wider range of applications, e.g. local content (where limited space typically does not allow display of content from multiple locations) or indications of the current distance or time to the user's home. This suggests that metrics that only consider the top-ranked cell are more appropriate.

Eisenstein et al. (2010) and Cheng et al. (2010) use metrics based on *error distance*, i.e. the distance between the correct location and the chosen location. (As mentioned above, for cell-based methods the chosen location methods is normally the centroid of the training documents in the chosen cell.) Either the **mean** or **median** error distance can be calculated. This has the advantage over cell accuracy in that it allows comparisons across distinct grid sizes. Furthermore, performance at greater precision than the size of a cell can be measured. This metric is also not tied to a cell-based representation, and can be employed e.g. when the mean shift algorithm is used to select a geolocated location other than a cell centroid.

A potential issue with error distance, however, is that at a fine grain the metric becomes dependent on the exact location chosen for a given evaluation document (which may not be accurate to more than city-level granularity). In addition, in some cases geolocation to a metro area is sufficient, and in these cases a measure like "accuracy within the metro area" might be desired. A proxy for this is *accuracy at 161 km* (**acc@161**), introduced by Cheng et al. (2010), which measures the fraction of documents whose error distance is at most 161 km (originally chosen as 100 miles).

In the rest of my experiments, I use mean and median error distance and accuracy at 161 kilometers (acc@161). As noted above, all of these metrics allow for direct comparison across different cell sizes. Following Han et al. (2014), I use acc@161 on development sets when choosing algorithmic parameter values such as cell and bucket sizes.

## 4.2 Results

### 4.2.1 Small Twitter corpus

In Eisenstein et al. (2010)'s experiments, all vocabulary items that appear in fewer than 40 users were ignored. This *thresholding* takes away a lot of very useful material, including many relatively rare but highly indicative toponyms. This suggests that a lower threshold would be better, and this

| Method | Parameters | A@161 | Mean | Med. |
|---|---:|---:|---:|---:|
| KL Uniform | 1° | 35.4 | 954 | 546 |
| KL *k*-d | BK250 | 32.9 | 910 | 539 |
| NB Uniform | 1° | 36.1 | 1009 | 552 |
| NB *k*-d | BK100 | 33.9 | 1007 | 598 |
| IGR Uniform | 2.5°, CU88% | 36.7 | 972 | 496 |
| IGR *k*-d | BK250, CU100% | 33.1 | 968 | 570 |
| FlatLR Uniform | 2.5° | **42.0** | 837 | **312** |
| FlatLR *k*-d | BK250 | 38.4 | 860 | 425 |
| HierLR Uniform | 5°, SF3, BM1 | 41.6 | 808 | 317 |
| HierLR *k*-d | BK250, SF2, BM2 | 38.7 | 877 | 460 |
| Eisenstein et al. (2010) | Geographic topic model | — | 900 | 494 |
| Hong et al. (2012) | Full model | — | — | 373 |
| Hulden et al. (2015) | $\mathrm{NB}_{kde2d}$ | — | **765** | 357 |
| Hulden et al. (2015) | KL | — | 802 | 333 |

Table 4.2: Performance on the test set of GeoText for different methods and metrics.

is borne out by my experiments, where a threshold of 5 is best.

Test set results are shown in Table 4.2 and are compared with a number of other papers that evaluate on the same dataset. Best acc@161 and median come from FlatLR. This is contrary to all the other corpora I consider, for which HierLR is consistently better — although even for this corpus, HierLR's value for mean beats all my other methods.[4] Mean is the only metric that considers the performance of points for which the predicted location is significantly inaccurate, and the take-home significance of this is that HierLR is doing a better job than all my other methods at reducing the likelihood of more extreme errors.

IGR outperforms NB with a uniform grid, similarly to the other Twitter datasets (and unlike the remaining datasets in this chapter), but in this case the gain is fairly slight, and IGR is actually worse than NB with a *k*-d grid.

Both FlatLR and HierLR manage to beat all values for median error distance reported in other papers. However, Hulden et al. (2015) report a mean value that is better than all my methods, using a kernel density estimation (KDE) technique. What is surprising about their results is the values they report for plain KL divergence, which differ drastically from the KL divergence figures that I obtain, with a median that even beats their own KDE method. I have no explanation for this

[4] For this dataset, FlatLR and HierLR were run with 15 BFGS passes for uniform and 12 passes for *k*-d, and HierLR was run with 9 passes and 22-bit features in sublevels.

discrepancy. Their figures are reported for a threshold of 5, just like my figures. The only other possible difference is in smoothing methods, yet they report using simple add-$n$ smoothing, and it is hard for me to believe that this can account for the difference in values.

### 4.2.2 Large Twitter corpora

I show the effect of varying cell size in Table 4.3 and *k*-d tree bucket size in Figure 4.1. The number of non-empty cells is shown for each cell size and bucket size. For NB, this is the number of cells against which a comparison must be made for each test document; for FlatLR, this is the number of classes that must be distinguished. For HierLR, no figure is given because it varies from level to level and from classifier to classifier. For example, with a uniform grid and subdivision factor of 3, each level-2 subclassifier will have between 1 and 9 labels to choose among, depending on which cells are empty.

| Method | Cell Size | | #Class | Acc. | Mean | Med. |
|---|---|---|---|---|---|---|
| | (Deg) | (km) | | @161 | (km) | (km) |
| NB | 0.17° | 18.9 | 11,671 | 36.6 | 929.5 | 496.4 |
| | 0.50° | 55.6 | 2,838 | 35.4 | 889.3 | 466.6 |
| IGR, cu90% | 1.5° | 167 | 501 | 45.9 | 787.5 | 255.6 |
| FlatLR | 5° | 556 | 59 | 35.4 | 727.8 | 248.7 |
| | 4° | 445 | 99 | 44.4 | 718.8 | 227.9 |
| | 3° | 334 | 159 | 47.3 | 721.3 | 186.2 |
| | 2.5° | 278 | 208 | 47.5 | 743.9 | 198.9 |
| | 2° | 223 | 316 | 46.9 | 737.7 | 209.9 |
| | 1.5° | 167 | 501 | 46.6 | 762.6 | 226.9 |
| | 1° | 111 | 975 | 43.0 | 810.0 | 303.7 |
| HierLR, d2, sf2, bm5 | 4° | – | – | 48.6 | **695.2** | 182.2 |
| HierLR, d2, sf2, bm2 | 3° | – | – | **49.0** | 725.1 | 174.6 |
| HierLR, d3, sf2, bm2 | 3° | – | – | **49.0** | 718.9 | **173.8** |
| HierLR, d2, sf2, bm5 | 2.5° | – | – | 48.2 | 740.9 | 187.7 |

Table 4.3: Dev set performance for TwUS, with uniform grids. HierLR and IGR parameters optimized using acc@161. Best metric numbers for a given method are underlined, except that overall best numbers are in bold.

FlatLR does much better than NB and IGR, and HierLR is still better. This is despite logistic regression needing to operate at a much lower resolution.[5] Interestingly, uniform-grid 2-level HierLR does better at 4° with a subdivision factor of 2 than the equivalent FlatLR run at

[5] The limiting factor for resolution was the 24-hour per job limit on my computing cluster.

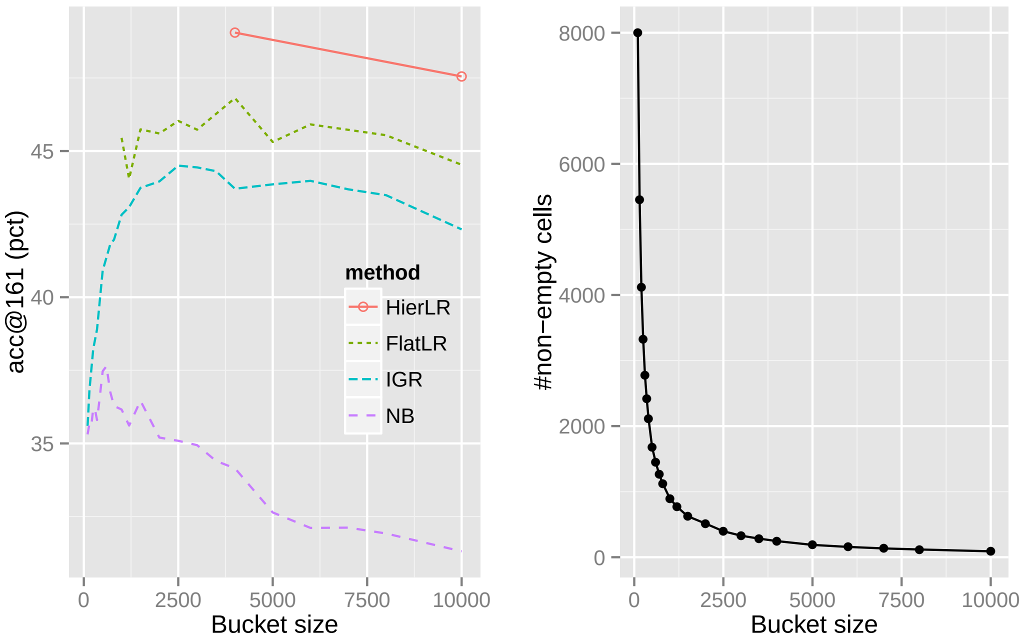


Figure 4.1: Dev set performance for TwUS, with *k*-d tree grids.

2°.

Table 4.4 shows the test set results for the various methods and metrics described in §4.1, on both TwUS and TwWorld.[6] HierLR is the best across all metrics; the best acc@161km and median error is obtained by HierLR with a uniform grid, while HierLR with *k*-d trees obtains the best mean error.

Compared with vanilla NB, my implementation of NB using IGR feature selection obtains large gains for TwUS and moderate gains for TwWorld, showing that IGR can be an effective geolocation method for Twitter. This agrees in general with Han et al. (2014)'s findings. I can only compare my figures directly with Han et al. (2014) for *k*-d trees—in this case they use a version of the same software I use and report figures within 1% of mine for TwUS. Their remaining results are computed using a city-based grid and an NB implementation with add-one smoothing, and are significantly worse than my uniform-grid NB and IGR figures using Dirichlet smoothing, which is

[6] Note that for TwWorld, it was necessary to modify the parameters normally passed to Vowpal Wabbit, moving up to 27-bit features and 96 passes, and 24-bit features with 24 passes in sublevels of HierLR.

| Corpus | TwUS | | | | TwWorld | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Parameters | A@161 | Mean | Med. | Parameters | A@161 | Mean | Med. |
| NB Uniform | 0.17° | 36.2 | 913.8 | 476.3 | 1° | 30.2 | 1690.0 | 537.2 |
| NB *k*-d | BK1500 | 36.2 | 861.4 | 444.2 | BK500 | 28.7 | 1735.0 | 566.2 |
| IGR Uniform | 1.5°, CU90% | 46.1 | 770.3 | 233.9 | 1°, CU90% | 31.0 | 2204.8 | 574.7 |
| IGR *k*-d | BK2500, CU90% | 44.6 | 792.0 | 268.6 | BK250, CU92% | 29.4 | 2369.6 | 655.0 |
| FlatLR Uniform | 2.5° | 47.2 | 727.3 | 195.4 | 3.7° | 32.1 | 1736.3 | 500.0 |
| FlatLR *k*-d | BK4000 | 47.4 | 692.2 | 197.0 | BK12000 | 27.8 | 1939.5 | 651.6 |
| HierLR Uniform | 3°, SF2, BM2 | **49.2** | 703.6 | **170.5** | 5°, SF2, BM1 | **32.7** | 1714.6 | **490.0** |
| HierLR *k*-d | BK4000, SF3, BM1 | 48.0 | **686.6** | 191.4 | BK60000, SF5, BM1 | 31.3 | **1669.6** | 509.1 |

Table 4.4: Performance on the test sets of TwUS and TwWorld for different methods and metrics.

| Corpus | EnWiki13 | | | | CoPhIR | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Parameters | A@161 | Mean | Med. | Parameters | A@161 | Mean | Med. |
| NB Uniform | 1.5° | 84.0 | 326.8 | 56.3 | 1.5° | 65.0 | 1553.5 | 47.9 |
| NB *k*-d | BK100 | 84.5 | 362.3 | 21.1 | BK3500 | 58.5 | 1726.9 | 70.0 |
| IGR Uniform | 1.5°, CU96% | 81.4 | 401.9 | 58.2 | 1.5°, CU92% | 60.8 | 1683.4 | 56.7 |
| IGR *k*-d | BK250, CU98% | 80.6 | 423.9 | 34.3 | BK1500, CU62% | 54.7 | 2908.8 | 83.5 |
| FlatLR Uniform | 7.5° | 25.5 | 1347.8 | 259.4 | 2.0° | 60.6 | 1942.3 | 73.7 |
| FlatLR *k*-d | BK1500 | 74.8 | 253.2 | 70.0 | BK3000 | 57.7 | 1961.4 | 72.5 |
| HierLR Uniform | 7.5°, SF3, BM5 | 86.2 | 228.3 | 34.0 | 7°, SF4, BM5 | 65.3 | 1590.2 | **16.7** |
| HierLR *k*-d | BK1500, SF12, BM2 | **88.9** | **168.7** | **15.3** | BK100000, SF15, BM5 | **66.0** | **1453.3** | 17.9 |

Table 4.5: Performance on the test sets of EnWiki13 and CoPhIR for different methods and metrics.

known to significantly outperform add-one smoothing (Smucker and Allan, 2006). For example, for NB they report 30.8% acc@161 for TwUS and 20.0% for TwWorld, compared with my 36.2% and 30.2% respectively. I suspect an additional reason for the discrepancy is due to the limitations of their city-based grid, which has no tunable parameter to optimize the grid size and requires that test instances not near a city be reported as incorrect.

My NB figures also beat the KL divergence figures reported in Roller et al. (2012) for TwUS (which they term UTGeo2011), perhaps again due to the difference in smoothing methods.

### 4.2.3 Wikipedia

Table 4.5 shows results on the test set of EnWiki13 for various methods. Table 4.7 shows the corresponding results for DeWiki14 and PtWiki14. In all cases, the best parameters for each method were determined using acc@161 on the development set, as above.

HierLR is clearly the stand-out winner among all methods and metrics, and particularly so for the *k*-d tree grid. This is achieved through a high subdivision factor, especially in a 2-level

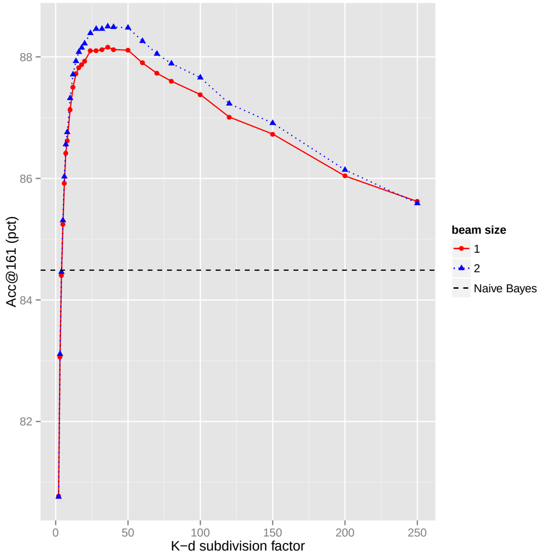


Figure 4.2: Plot of subdivision factor vs. acc@161 for the EnWiki13 dev set with 2-level *k*-d tree HierLR, bucket size 1500. Beam sizes above 2 yield little improvement.

hierarchy, where a factor of 36 is best, as shown in Figure 4.2 for EnWiki13. (For a 3-level hierarchy, the best subdivision factor is 12.)

Unlike for TwUS, FlatLR simply cannot compete with NB in the larger Wikipedias (EnWiki13 and DeWiki14). EnWiki13 especially has dense coverage across the entire world, whereas TwUS only covers the United States and parts of Canada and Mexico. Thus, there are a much larger number of non-empty cells at a given resolution and much coarser resolution required, especially with the uniform grid. For example, at 7.5° there are 933 non-empty cells, comparable to 1° for TwUS. Table 4.6 shows the number of classes and runtime for FlatLR and HierLR at different parameter values. The hierarchical classification approach is clearly essential for allowing me to scale the discriminative approach to handle a large, dense dataset covering the whole world.

Moving from larger to smaller Wikipedias, FlatLR becomes more competitive. In particular, FlatLR outperforms NB and is close to HierLR for PtWiki14, the smallest of the three (and significantly smaller than TwUS). In this case, the relatively small size of the dataset and its greater geographic specificity (many articles are located in Brazil or Portugal) allows for a fine enough resolution to make FlatLR perform well—comparable to or even finer than NB.

| Method | Param | #Class | A@161 | Med. | Runtime |
|---|---:|---:|---:|---:|---:|
| FlatLR Uniform | 10° | 648 | 19.2 | 314.1 | 11h |
| | 8.5° | 784 | 26.5 | 248.5 | 16h |
| | 7.5° | 933 | 30.1 | 232.0 | 19h |
| FlatLR *k*-d | BK5000 | 257 | 57.1 | 133.5 | 5h |
| | BK2500 | 501 | 67.5 | 94.9 | 9h |
| | BK1500 | 825 | 74.7 | 69.9 | 16h |
| HierLR Uniform | 7.5°,SF2,BM1 | — | 85.2 | 67.8 | 23h |
| | 7.5°,SF3,BM5 | — | 86.1 | 34.2 | 27h |
| HierLR *k*-d | BK1500,SF5,BM1 | — | 88.2 | 19.6 | 23h |
| | BK5000,SF10,BM5 | — | 88.4 | 18.3 | 14h |
| | BK1500,SF12,BM2 | — | 88.8 | 15.3 | 33h |

Table 4.6: Performance/runtime for FlatLR and 3-level HierLR on the EnWiki13 dev set, with varying parameters.

In all of the Wikipedias, NB *k*-d outperforms NB uniform, and HierLR outperforms both, but by greatly varying amounts, with only a 1% difference for DeWiki14 but 12% for PtWiki14. It's unclear what causes these variations, although it's worth noting that Roller et al. (2012)'s NB *k*-d figures on an older English Wikipedia corpus were noticeably higher than my figures: They report 90.3% acc@161, compared with our 84.5%. I verified that this is due to corpus differences: I obtain their performance when I run on their Wikipedia corpus. This suggests that some of the differences between methods may be due to vagaries of the individual corpora, e.g. the presence of differing numbers of geotagged stub articles, which are very short and thus hard to geolocate.

As for IGR, though it is competitive for Twitter, it performs badly here—in fact, it is even worse than plain Naive Bayes for all three Wikipedias (likewise for CoPhIR, in the next section).

### 4.2.4 CoPhIR

Table 4.5 shows results on the test set of CoPhIR for various methods, similarly to the EnWiki13 results. HierLR is again the clear winner. Unlike for EnWiki13, FlatLR is able to do fairly well. IGR performs poorly, especially when combined with *k*-d.

In general, as can be seen, for CoPhIR the median figures are very low but the mean figures very high, meaning there are many images that can be very accurately placed while the remainder are very difficult to place. (The former images likely have the location mentioned in the tags, while

| Corpus | DEWIKI14 | | | | PTWIKI14 | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Parameters | A@161 | Mean | Med. | Parameters | A@161 | Mean | Med. |
| NB Uniform | 1° | 88.4 | 257.9 | 35.0 | 1° | 76.6 | 470.0 | 48.3 |
| NB *k*-d | BK25 | 89.3 | 192.0 | **7.6** | BK100 | 77.1 | 325.0 | 45.9 |
| IGR Uniform | 2°, CU82% | 87.1 | 312.9 | 68.2 | 2°, CU54% | 71.3 | 594.6 | 89.4 |
| IGR *k*-d | BK50, CU100% | 86.0 | 226.8 | 10.9 | BK100, CU100% | 71.3 | 491.9 | 57.7 |
| FLATLR Uniform | 5° | 55.1 | 340.4 | 150.1 | 2° | 88.9 | 320.0 | 70.8 |
| FLATLR *k*-d | BK350 | 82.0 | 193.2 | 24.5 | BK25 | 86.8 | 320.8 | 30.0 |
| HIERLR Uniform | 7°, SF3, BM5 | 88.5 | 184.8 | 30.0 | 7°, SF2, BM5 | 88.6 | 223.5 | 64.7 |
| HIERLR *k*-d | BK3500, SF25, BM5 | **90.2** | **122.5** | 8.6 | BK250, SF12, BM2 | **89.5** | **186.6** | **27.2** |

Table 4.7: Performance on the test sets of DEWIKI14 and PTWIKI14 for different methods and metrics.

the latter do not.)

My NB results are not directly comparable to O'Hare and Murdock (2013)'s results on COPHIR because they use various cell-based accuracy metrics while I use cell-size-independent metrics. The closest to my acc@161 metric is their Ac1 metric, which at a cell size of 100 km corresponds to a 300km-per-side square at the equator, roughly comparable to my 161-km-radius circle. They report Ac1 figures of 57.7% for term frequency and 65.3% for user frequency, which counts the number of distinct users in a cell using a given term and is intended to offset bias resulting from users who upload a large batch of similar photos at a given location. My term frequency figure of 65.0% significantly beats theirs, but I found that user frequency actually degraded my dev set results by 5%. The reason for this discrepancy is unclear.

### 4.2.5 Summary and discussion

**Summary of results.** HIERLR is the clear winner, outperforming Naive Bayes, KL Divergence, IGR (information gain ratio) and all other methods on all the large corpora, typically by a significant margin. This is a strong result, especially given the highly disparate nature of the various corpora. Among these corpora, FLATLR is able to beat NB on some of these corpora (PTWIKI14, TWUS, and TWWORLD), but fails badly on others—in particular, those that are large and with worldwide scope—due to its inability to scale down to fine enough grid sizes. Investigation reveals clearly that this is because Vowpal Wabbit cannot easily handle more than about 1,000 distinct classes (i.e. non-empty grid cells), whereas NB has no problem with even 100,000+ classes due to its linear dependence on the number of classes. (A potential trade-off here is that HIERLR requires more careful optimization than NB, with 3–4 parameters needing tuning.)

Feature selection through IGR, introduced in geolocation by Han et al. (2014) for Twitter, appears to perform well only for Twitter, but not for Wikipedia or COPHIR. Han et al.'s method involves both filtering low-frequency words and cutting out words with low IGR. In their method, only the latter step is associated with a tunable parameter. However, it is possible to separate the two steps. Indeed, if no filtering of uncommon words is done, IGR should never perform worse than Naive Bayes, since one possible setting for the cutoff value is not to cut any words, making the method equivalent to Naive Bayes.

For COPHIR, and also TWWORLD (and partly for GEOTEXT), HIERLR performs best when the root level is significantly coarser than the cell or bucket size that is best for FLATLR. The best setting for the root level appears to be correlated with cell accuracy, which in general increases with larger cell sizes. The intuition here is that HIERLR works by drilling down from a single top-level child of the root cell. Thus, the higher the cell accuracy, the greater the fraction of test instances that can be improved in this fashion, and in general the better the ultimate values of the main metrics. (The above discussion isn't strictly true for beam sizes above 1, but these tend to produce marginal improvements, with little if any gain from going above a beam size of 5.) The large size of a coarse root-child cell, and correspondingly poor results for acc@161, can be offset by a high subdivision factor, which does not materially slow down the training process.

**Optimizing for median.** Note that better values for the other metrics, especially median, can be achieved by specifically optimizing for these metrics. In general, the best parameters for median are finer-scale than those for acc@161: smaller grid sizes and bucket sizes, and greater subdivision factors. This is especially revealing in ENWIKI13 and COPHIR. For example, on the ENWIKI13 dev set, the "best" uniform NB parameter of 1.5°, as optimized on acc@161, yields a median error of 56.1 km, but an error of just 16.7 km can be achieved with the parameter setting 0.25° (which, however, drops acc@161 from 83.8% to 78.3% in the process). Similarly, for the COPHIR dev set, the optimized uniform 2-level HIERLR median error of 46.6 km can be reduced to just 8.1 km by dropping from 7° to 3.5° and bumping up the subdivision factor from 4 to 35—again, causing a drop in acc@161 from 68.6% to 65.5%.

| Salt Lake | San Francisco | New Orleans | Phoenix | Denver | Houston | Montreal | Seattle | Tulsa | Los Angeles |
|---|---|---|---|---|---|---|---|---|---|
| utah | sacramento | orleans | tucson | denver | houston | montreal | seattle | tulsa | knotts |
| slc | hella | jtfo | az | colorado | antonio | mtl | portland | okc | sd |
| salt | sac | prelaw | phoenix | broncos | texans | quebec | tacoma | oklahoma | pasadena |
| byu | niners | saints | arizona | aurora | sa | magrib | wa | wichita | diego |
| provo | berkeley | louisiana | asu | amarillo | corpus | rue | vancouver | ou | ucla |
| ut | safeway | bourbon | tempe | soopers | whataburger | habs | bellevue | kansas | disneyland |
| utes | oakland | kmsl | scottsdale | colfax | heb | canadian | oregon | ku | irvine |
| idaho | earthquake | uptown | phx | springs | otc | ouest | seahawks | lawrence | socal |
| orem | sf | joked | chandler | centennial | utsa | mcgill | pdx | shaki | tijuana |
| sandy | modesto | wya | fry | pueblo | mcallen | coin | uw | ks | riverside |
| rio | exploit | canal | glendale | larimer | westheimer | gmusic | puyallup | edmond | pomona |
| ogden | stockton | metairie | desert | meadows | pearland | laval | safeway | osu | turnt |
| lds | hayward | westbank | harkins | parker | jammin | poutine | huskies | stillwater | angeles |
| temple | cal | bayou | camelback | blake | mayne | boul | everett | topeka | usc |
| murray | jose | houma | mesa | cherry | katy | est | seatac | sooners | chargers |
| menudito | swaaaaggg | lawd | gilbert | siiiiim | jamming | je | ducks | straighht | oc |
| mormon | folsom | gtf | pima | coors | tsu | sherbrooke | victoria | kc | compton |
| gateway | roseville | magazine | dbacks | englewood | marcos | pas | beaverton | manhattan | meadowview |
| megaplex | juiced | gumbo | mcdowell | pikes | laredo | fkn | hella | boomer | rancho |
| lake | vallejo | buku | devils | rockies | texas | centre | sounders | sooner | ventura |

Table 4.8: Top 20 features selected for various regions using logistic regression on TWUS with a uniform 5° grid.

**Hierarchy depth.** We use a 3-level hierarchy throughout for the test set results. Evaluation on development data showed that 2-level hierarchies perform comparably for several datasets, but are less effective overall. We did not find improvements from using more than three levels. When using a simple local classifier per parent approach as we do, which chains together spines of related but independently trained classifiers when assigning a probability to a leaf cell, most of the benefit presumably comes from simply enabling logistic regression to be used with fine-grained leaf cells, overcoming the limitations of FLATLR. Further benefits of the hierarchical approach might be achieved with the data-biasing and bottom-up error propagation techniques of Bennett and Nguyen (2009) or the hierarchical Bayesian approach of Gopal et al. (2012), which is able to handle large-scale corpora and thousands of classes.

## 4.3 Feature selection

The main focus of Han et al. (2014) is identifying geographically salient words through feature selection. Logistic regression performs feature selection naturally by assigning higher weights to features that better discriminate among the target classes.

Table 4.8 shows the top 20 features ranked by feature weight for a number of different cells, labeled by the largest city in the cell. The features were produced using a uniform 5° grid, trained using 27-bit features and 40 passes over TwUS. The high number of bits per feature were chosen to ensure as few collisions as possible of different features (as it would be impossible to distinguish two words that were hashed together).

Most words are clearly region specific, consisting of cities, states and abbreviations, sports teams (*broncos*, *texans*, *niners*, *saints*), well-known streets (*bourbon*, *folsom*), characteristic features (*desert*, *bayou*, *earthquake*, *temple*), local brands (*whataburger*, *soopers*, *heb*), local foods (*gumbo*, *poutine*), and dialect terms (*hella*, *buku*).

| Top-IGR words | | | | Bottom-IGR words | | | |
|---|---|---|---|---|---|---|---|
| 1–10 | 11–20 | 21–30 | 31–40 | 1–10 | 11–20 | 21–30 | 31–40 |
| lockerby | ghibran | presswiches | curtisinn | plan | black | times | true |
| killdeer | briaroaks | haubrich | guymon | party | dream | end | found |
| fordville | joekins | yabbo | dakotamart | men | hey | twitter | drink |
| azilda | numerica | presswich | missoula | happy | face | full | pay |
| ahauah | bemidji | pozuelo | mimbres | show | finally | part | meet |
| hutmacher | amn | akeley | shingobee | top | easy | forget | lost |
| cere | roug | chewelah | gottsch | extra | time | close | find |
| miramichi | pbtisd | computacionales | uprr | late | live | dead | touch |
| alamosa | marcenado | bevilacqua | hesperus | facebook | wow | cool | birthday |
| multiservicios | banerjee | presswiche | racingmason | friday | yesterday | enjoy | ago |

Table 4.9: Top and bottom 40 features selected using IGR for TwUS with a uniform 1.5° grid.

As a comparison, Table 4.9 shows the top and bottom 40 features selected using IGR on the same corpus. Unlike for logistic regression, the top IGR features are mostly obscure words, only some of which have geographic significance, while the bottom words are quite common. To some extent this is a feature of IGR, since the denominator of the IGR formula contains the binary entropy of each word, which is directly related to its frequency. However, it shows why cutoffs around 90% of the original feature set are necessary to achieve good performance on the Twitter corpora. (IGR does not perform well on Wikipedia or CoPhIR, as shown above.)

# Chapter 5

# Document geolocation for the digital humanities

The previous chapters developed techniques for text-based document geolocation and demonstrated their feasibility on a number of modern-day corpora. In this and the subsequent chapter, I extend these techniques for use with historical documents in the digital humanities. There is little or no work applying document geolocation to historical corpora in the digital humanities, and no corpora available for evaluation. To facilitate further research in the field, I develop a new NLP task, *text-based document geolocation of historical corpora*, and provide two new annotated corpora for evaluation purposes. These two corpora are of significantly different natures: a 19th-century travel log (John Beadle's *Western Wilds*, aka BEADLE, §2.3.1) and a large collection of primary sources covering the American Civil War (the *War of the Rebellion*, aka WOTR, §2.3.2). This makes it possible to generalize the performance of different methods beyond a single corpus. I apply my existing methods to these corpora, yielding good accuracy despite their smaller size and significantly different nature compared with the modern-day corpora, and demonstrate further improvements through Wikipedia-based domain adaptation. These results should serve as a strong baseline for the development of further text-based geolocation techniques.

I then demonstrate the power and real-world applicability of geolocation models by combining their predictions with a dynamic topic model (Blei and Lafferty, 2006) to generate a *geographic topic model*—another thing that has not previously been done in the digital humanities—and show how it can be used to yield useful insights about the U.S. Civil War. The significance of this lies in the special place that topic models serve in the digital humanities. The typical use of NLP in the digital humanities is as a tool for exploratory data analysis of large-scale textual datasets consisting of primary sources in the humanities. Among the tools used, topic modeling is one of the most frequently used, if not *the* single most frequently used tool (Meeks and Weingart, 2012).

In the following chapter, I develop an entirely new geolocation method that specifically targets text-only historical corpora, using co-training between a toponym resolver and a document geolocator. The toponym resolver works in conjunction with a gazetteer of potentially ambiguous place names and possible resolutions to latitude/longitude coordinates. This has the effect of injecting outside knowledge into the training process beyond what can be learned from the text alone. This is somewhat analogous to how the *ego network* of friends, followers, and/or direct communication paths leading from a Twitter user to other Twitter users provides additional information that facilitates the creation of network-based models that greatly improve the accuracy of Twitter user geolocation.

It should now be clear why I have deliberately restricted my geolocation techniques to be text-only. In the context of social media corpora, this feels a bit like fighting with one hand tied behind one's back because of the wealth of metadata available; even a largely textual resource such as Wikipedia provides hyperlinks between pages that can enable similar network-based methods such as label propagation. Historical corpora in the humanities, however, typically come as text-only sources, rendering non-text-based methods largely inapplicable.

As mentioned above, there has been little or no previous work in applying document geolocation to historical corpora, although some authors (e.g. Dias et al. (2012), in addition to the various papers I have authored or coauthored) allude to it as one of the use cases of text-based document geolocation. Some authors have applied other sorts of geolocation or geocoding to historical documents. A number of authors have designed tools that interface a GIS with historical maps (Chias and Abad, 2009; Bollini et al., 2013; Ferrighi, 2015) or audiovisual resources (Zurcher, 2013). Perhaps the most relevant is Chasin et al. (2013), who geolocate individual toponyms using named entity recognition (NER) combined with toponym resolution. This is not the same as document geolocation and in fact is not actually applied to historical documents at all but to modern documents (from Wikipedia) that are *about* historical topics.

In this chapter, §5.1 describes how domain adaptation with Wikipedia as an out-of-domain training source can be effective given the relatively small amount of annotated data (in particular as compared with the large modern Twitter, Wikipedia and Flickr corpora described in Chapter 2). In §5.2 and §5.3 I do experiments on BEADLE and WOTR, respectively, of the same sort as was previously done in Chapter 4 on the modern corpora, to show the feasibility of my methods even on somewhat smaller corpora. Here I also apply domain adaptation techniques and discuss various issues involved in guiding an annotation process, including learning curves that show the tradeoff between annotation time and geolocation accuracy.

In the rest of the chapter I investigate the full set of data in WOTR, which is much larger than the annotated subset of data. (There are 5,010 annotated documents and approximately 255,000 total documents.) §5.4 discusses in detail the distribution of the full set of WOTR documents and various ways of evaluating that distribution, comparing the distribution yielded by a model trained of the annotated documents with a technique of geolocating through toponym resolution and a separate corpus of military actions. Building on the geolocation techniques in this dissertation, §5.5 includes various topic-model-based analyses, including a simple scheme for geographic topic models and an analysis based on dynamic topic models, which are designed for sets of auto-correlated topic models that vary over time or space.

## 5.1 Extending geolocation to historical corpora

The difficulty with applying the techniques of the previous chapters to the historical corpora of this chapter is the relatively small amount of annotated material. BEADLE has only 408 annotated paragraphs (with each paragraph corresponding to a data instance), while WOTR has 5,010 annotated articles. However, given the availability of the large ENWIKI13 corpus of Wikipedia articles, the framework of *domain adaptation* can be applied. Domain adaptation refers to a situation where only a small amount of in-domain but a large amount of out-of-domain training material is available. The in-domain training material is assumed to be drawn from the same distribution as the test instances. The out-of-domain material is not, but is assumed to have some properties in common that can be taken advantage of.

### 5.1.1 The applicability of domain adaptation

Why would domain adaptation be beneficial? In this case, for example, the assumption would be that many of the words that are geographically indicative of certain places in Wikipedia are indicative of those same places in BEADLE or WOTR. This assumption appears reasonable in many instances. For instance, many of the most geographically indicative words are toponyms or other geographically-salient proper nouns, such as names of Native American tribes or groups such as the Mormons.

It is true that some of these names have changed. For example, Beadle refers to the Hopi tribe as the “Moqui”, and collectively terms the mountains of New Mexico the “Sierra Madre”, whereas nowadays there is no collective term for those mountains and the term “Sierra Madre” itself normally refers to a different mountain range in Mexico. Beadle also terms the Purgatoire River in Colorado the “Las Animas River”, whereas the modern-day ”Animas River” is a different Colorado river. Furthermore, none of these older usages can be found in Wikipedia. A similar situation obtains in WOTR with places such as “Keatsville, Missouri” (modern-day “Keytesville”, whose Wikipedia entry does not list the older spelling).

A different issue comes from toponyms referring to places that no longer exist. Beadle, for example, mentions a number of railroad ghost towns that had already ceased to exist in his time, such as Deadfall, Last Chance, Murder Gulch and Painted Post in Utah, and Benton, Wyoming. Of these, only Benton can be found in Wikipedia (and not in its own article but in the article concerning the nearby town of Sinclair). Other towns are mentioned that existed at the time but no longer do, such as Red Dog, California (has its own Wikipedia article) and Hazard Station, Wyoming (no mention in Wikipedia).

WOTR is full of such toponyms. Many of them refer to temporary places such as “McCullan's Store" in Missouri (apparently a settlement containing a store) or various army camps. These camps may be given names such as "Camp McIntosh" (named after the commander James McIntosh) and may appear in the bylines of letters in WOTR, but have a strictly temporary existence and disappear as soon as the army occupying them moves on. Only somewhat less temporary are numerous forts such as Fort Lyon in Missouri and Fort Jackson in Arkansas, which existed only for a few years during the Civil War. (Beadle similarly mentions a Fort Lancaster in Colorado, which existed only from 1837-1844.) For the most part none of these places can be found in Wikipedia.

However, this is less of an issue than it may appear. For one thing, the large majority of places mentioned in both BEADLE and WOTR still exist with the same names they had 150 years ago. This includes places that may have changed their nature, such as the former territories of Arizona, Utah, Colorado and Dakota, which have since transitioned into states but largely kept the previous names. Similarly, most ethnonyms, such as the Mormons, Navajos, Apaches and Utes have remained the same. In many cases where names have changed (e.g. Davisville, California was renamed to Davis in 1907, and City Point, Virginia was annexed into Hopewell, Virginia in 1923), the old names are prominently mentioned in Wikipedia. Some civil war forts, and most places associated with battles, likewise are either featured in their own Wikipedia articles or mentioned prominently in other articles, often the article describing the battle taking place at that location.

A different and perhaps more significant issue comes with terms that have distributions that differ significantly in Wikipedia vs. BEADLE and WOTR. One issue is with names that may have a most prominent sense in Wikipedia that is different from the usage in BEADLE or WOTR. Some examples:

1. Many forts that existed during the 19th century bear the same names as modern forts in different locations (e.g. the modern-day Fort Lyon in Colorado, Fort Lancaster in Texas, and Fort Jackson in South Carolina, compared with the above-mentioned forts of the same names in Missouri, Colorado and Arkansas, respectively).

2. The place name "Columbus" tends to refer in Wikipedia to Columbus, Ohio, but in WOTR to Columbus, Kentucky, which saw significant fighting, whereas Columbus, Ohio did not.

3. "Grant" in WOTR is likely to refer to General Ulysses S. Grant, whereas its distribution in

Wikipedia is due not only to General Grant but to numerous other people and places with the same name. It is also affected by lowercase "grant", due to case-folding in the algorithm I use; this is done due to many inconsistencies in case usage, such as all-lowercase text in Twitter, all-capital text in WOTR, and of course capitalized words at the beginning of a sentence. Even if all occurrences of "Grant" in Wikipedia that refer to General Grant could be separated out, there still remains the issue that at least half of Grant's Civil War service, and hence mention in WOTR, was in the Western Theater (e.g. in Missouri and Tennessee), whereas the majority of the text in Wikipedia on Grant appears to covers his two terms as a U.S. President, during which he was located in Washington, D.C.

4. "Sherman" in Wikipedia (and WOTR) concerns General William T. Sherman, whereas "Sherman" in BEADLE is primarily the name of a town in Wyoming.

5. Most mentions of "Washington" in Wikipedia actually refer to Washington State (the term "Washington" is linked 17,127 times in the November 4, 2013 Wikipedia to the article on Washington State, but only 3,581 times to the article on Washington, D.C.), whereas nearly all mentions of "Washington" in WOTR refer to Washington, D.C.

Note that all of the above examples concern the U.S. Things get even worse when the possibility of terms referring to places, people, etc. across the whole world is considered. In practice, however, this isn't an issue: When using Wikipedia as a source I limit the regions considered to those within a bounding box surrounding the United States, due to the fact that nearly all locations in both BEADLE and WOTR are within the U.S.

### 5.1.2 Domain adaptation techniques

Daumé III (2007) has a discussion of a number of domain adaptation methods, including various baselines that are "surprisingly difficult to beat":

1. SRCONLY, which trains only using the "source" (out-of-domain) material.

2. TGTONLY, which trains only using the "target" (in-domain) material.

3. ALL, which trains on the union of the two domains.

4. WEIGHTED, which downweights examples from the out-of-domain material to avoid it swamping the in-domain material, with the weight optimized by cross-validation or a dev set.

5. PRED, which trains a SRCONLY model and uses it to generate predictions on the full set of annotated in-domain material (training, dev and test sets), which then serve as additional features in a model trained on the in-domain training data.

6. LININT, which linearly interpolates between the predictions of the SRCONLY and TGTONLY models, with the interpolation factor optimized by cross-validation or a dev set. (Note that we also discuss interpolation in the context of co-training in §6.3.3.)

Daumé III (2007) also proposes a new domain adaptation method, EASYADAPT, which works by expanding the feature space to include a combined space of features that fire only for in-domain training examples, features that fire only for out-of-domain training examples, and features that fire for both. This method is easy to implement and works well in experimental results on various domains, such as named-entity recognition, part-of-speech tagging and shallow parsing. A major limitation of this model is that it is fully supervised, meaning that we cannot take advantage of the remainder of the paragraphs in the original Beadle document that have not been annotated. This is rectified by Daumé III et al. (2010), a semi-supervised version of EASYADAPT termed EA++, although I did not perform experiments using this latter method.

## 5.2 Geolocation experiments on BEADLE

BEADLE was split 60/20/20 in a round-robin fashion into training, dev and tests. (The small size of the corpus makes it a good candidate for cross-validation, something that could be done in future research.)

### 5.2.1 Cross-domain geolocation

My first experiments involved simply training on ENWIKI13 and testing on BEADLE (Daumé's SRCONLY model), as a baseline for further work, on the assumption of similarity in the geographic word distribution of the Wikipedia and Beadle documents.

#### Naive Bayes

Early experiments using Naive Bayes showed that predictions were significantly harmed by the need to make world-wide predictions when nearly all locations in Beadle were in the United States, so a bounding box was applied consisting of latitudes in the range $[25, 49]$ and longitudes in the range $[-126, -60]$. Although this includes parts of Canada and Mexico as well as the entire contiguous United States, it resulted in significant improvements, especially in mean and median, although less so for accuracy@161, as can be seen in Figure 5.1.

Note that these and other results with Naive Bayes in this chapter are computed using a uniform Naive Bayes prior rather than the more standard prior based on the number of training documents in a cell. Experiments comparing the two did not yield benefits from the latter type of prior, and the results were in fact significantly worse for in-domain data (§5.2.2); in that case, at least, it appears that the use of such a prior swamps the likelihood term.

Both median and accuracy bounce around a good deal, and are somewhat uncorrelated ($r = -0.52$ when unrestricted, $r = -0.37$ when restricted). This may well be an effect of the small corpus. With just 82 data points in each of the dev and test sets, a 1% difference in accuracy corresponds to less than 1 data point. (This definitely suggests using cross-validation in general, which I did in producing learning curves for WOTR.)

#### Logistic regression

Initial experiments using flat and hierarchical logistic regression were not promising. Eventually I realized that the regression models were overfitting to Wikipedia and were highly sensitive to the number of BFGS passes. The number of passes used for Wikipedia and most other corpora above was 40, which worked well. However, the optimal number of passes for Beadle was less, and

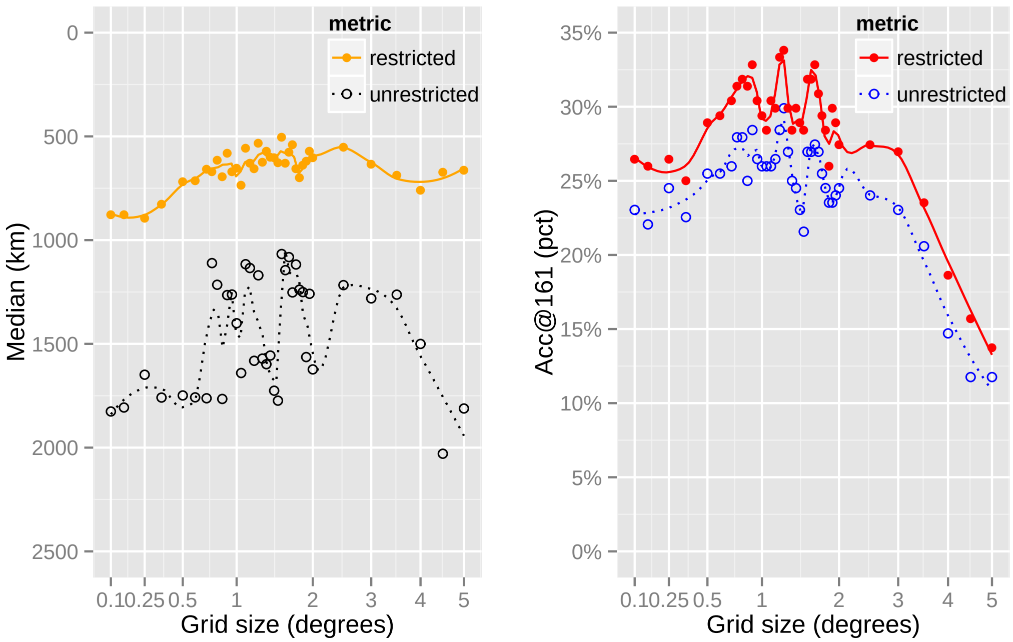


Figure 5.1: Plot of results of doing Naive Bayes on the dev set of BEADLE, training on ENWIKI13, with and without a restriction to the contiguous United States. Note that median error (in kilometers) is on a reversed scale, since smaller values are better.

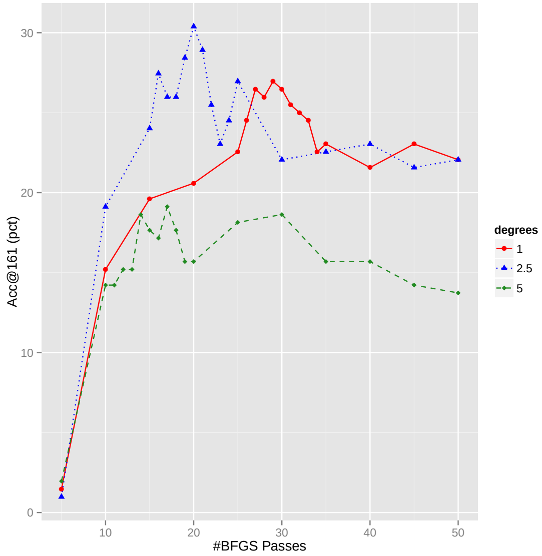


Figure 5.2: Plot of FLATLR accuracy as a function of grid size (degrees) and number of passes.

furthermore varied depending on the grid size, as shown in Figure 5.2. As can be seen, the optimal number of passes increases from 17 to 20 to 29 as the degree size decreases from 5 to 2.5 to 1. The best dev set result was slightly worse than Naive Bayes in accuracy but slightly better in median (see Table 5.1).

These runs were done without any $l_1$ or $l_2$ regularization. Varying $l_2$ parameters over a wide range from $[0.0001, 10]$ showed essentially no effect at $l_2 < 1$ and somewhat worse results at higher values. $l_1$ regularization is tricky to do using the particular logistic regression package and settings I used (BFGS under Vowpal Wabbit); when this was done for WOTR, there was no effect.

### Hierarchical classification

A similar procedure was used as in previous corpora to find optimal parameters for hierarchical logistic regression. A 3-level hierarchy was trained at 2.5°and 5°, the former because it produced close to the optimal results for FLATLR and the latter because previous experience showed that HIERLR often performed better given a coarser first-level grid size (§4.2.5). Consistent with previous experiments, subdivision factors of 2, 3 and 4 and beam sizes of 1, 2, 5 and 10 were tried. In addition, because of the clear dependency in FLATLR on the number of BFGS passes, the number of passes

| Method | Parameters | A@161 | Mean | Med. |
|---|---:|---:|---:|---:|
| RAND | 5° | 1.5 | 1963 | 1614 |
| PRIOR | 2.75° | 0.0 | 2599 | 2606 |
| ACP | 2.5° | 31.4 | 1006 | 568 |
| KL | 1° | 21.1 | 1157 | 1000 |
| NB | 1.25° | 33.8 | 905 | 532 |
| IGR | 2.5°, CU82% | **34.3** | 924 | 495 |
| FLATLR | 1.75°, 20 passes | 30.9 | 997 | 500 |
| HIERLR | 5°, SF3, BM10, 15 passes, 9 subpasses | 31.4 | **883** | **422** |

Table 5.1: Performance of cross-domain training (training on ENWIKI13, testing on BEADLE's dev set), uniform grid, for different methods and metrics.

used for lower-level classifiers was varied from the previously-determined optimal value of 12, with 6, 9, 12 and 15 passes tried. Best results turned out to come from 9 passes, although the effect of varying this parameter was less dramatic than the corresponding setting for the top-level classifier.

#### Results

Results for all ranking methods are shown in Table 5.1. In this case, IGR outperforms for acc@161, but HIERLR wins on mean and median, as it does with most other corpora. KL does quite poorly, for reasons not completely understood (unlike in previous experiments, where KL and Naive Bayes tended to perform similarly).

### 5.2.2 Within-domain geolocation

#### Results

Pure within-domain geolocation was also done on BEADLE, corresponding to Daumé's TGTONLY model. Similarly to above, when running logistic regression a search was done to find the optimal number of BFGS passes. The results are shown in Table 5.2.

It was suspected that results would be poor due to the small number of training instances (244), but in fact the results were significantly better than when training on Wikipedia, as can be seen in Table 5.2. It is clear that this is due to the relatively small number of distinct locations occurring in the data, and their frequent repetition. In essence, many locations in the test set have already been seen in the training set. This is made clear by the 11% accuracy figure of RAND and especially

| Method | Parameters | A@161 | Mean | Med. |
|---|---:|---:|---:|---:|
| RAND | 0.1° | 11.0 | 1235 | 962 |
| PRIOR | 2.5° | 15.8 | 1535 | 1428 |
| ACP | 0.5° | 54.9 | 561 | 111 |
| KL | 2.5° | 58.5 | 606 | 99 |
| NB | 1° | 54.9 | 552 | 108 |
| IGR | 2.5°, CU88% | 46.3 | 673 | 197 |
| FLATLR | 2.5°, 6 passes | 56.1 | **461** | 107 |
| HIERLR | 5°, SF2, BM1, 8 passes, 3 subpasses | **59.8** | 513 | **83** |

Table 5.2: Performance of within-domain training on BEADLE's dev set, uniform grid.

the 16% accuracy figure of PRIOR, which simply selects the most commonly-occurring cell. In this case, this cell is due to a stretch of 34 adjacent paragraphs, crossing two chapters, which I notated as "somewhere along a watershed divide near Taos Mountains" and annotated with the town of Eagle Nest, New Mexico. 24 of them appear in the training data, making up 10% of the total training data. The second most common location is Salt Lake City and environs, appearing 13 times in the training set, due to the author's frequent dealings with the Mormons.

In this case, unlike in the previous section, KL actually beats Naive Bayes and IGR does terribly. Best results come from HIERLR and FLATLR, with the latter having the best mean and the former the best acc@161 and median. Recall that the mean is penalized according to the magnitude of large errors while acc@161 and median are not; in this case, FLATLR is doing slightly better at reducing the error of inaccurate choices but slightly worse at making accurate choices.

### Word vs. location distributions

An important issue with document geolocation is what it means for training data to be "in-domain", i.e. drawn from the same probability distribution as the test data. In fact, there are at least two separate issues: that the words come from the same distribution and the locations come from the same distribution. For grid-cell geolocation, the distribution of locations influences the choices made in at least two ways, even without using a prior over locations (e.g. according to the number of documents in a cell). Only grid cells containing at least one training instance will be selected, which in the case of a sparse dataset such as BEADLE greatly restricts the set of possible choices, much beyond the simple bounding-box restriction to the contiguous United States that I use. In

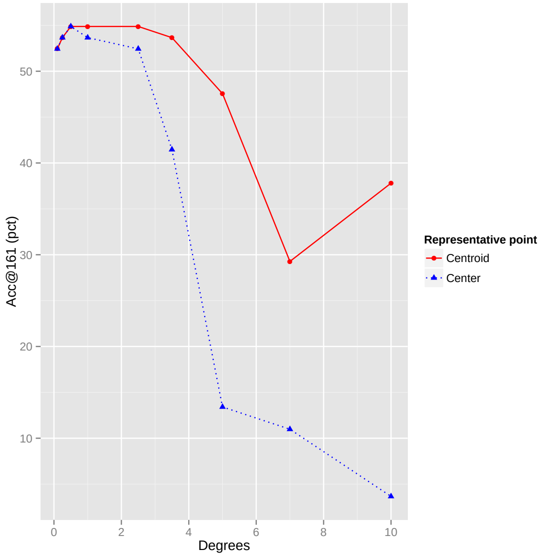


Figure 5.3: Plot of accuracy@161 as a function of grid size (degrees) for center vs. centroid.

addition, once a given grid cell is chosen, the actual point used as the predicted location (the cell's *representative point*) is based on the centroid of training points in the cell.

Figure 5.3 investigates the use of the centroid as representative point, rather than the cell's geographic center (§3.3). The choice of centroid has no effect at grid sizes of 0.5° or less, but has a progressively greater benefit at larger grid sizes, especially above 3.5°, where a huge drop-off in accuracy occurs when using the geographic center as representative point. This corresponds to the point at which a grid cell becomes significantly larger than a 161-km circle, meaning there is a high chance that, even if the correct cell is chosen, the correct point will lie more than 161 km from the center and will be considered a "miss". No such drop-off occurs when using the centroid method, and the accuracy actually *increases* going from 7° to 10° (at which point a grid cell at typical mid-latitudes is over 700 by 1,000 km in size). This shows the huge benefit gained by having the distribution of locations in the training data closely match that of the test data.

| Method | Parameters | A@161 | Mean | Med. |
|---|---|---|---|---|
| NB | 2.0° | 72.6 | 186.9 | 78.0 |
| FLATLR | 1°, 15 passes | 72.8 | 182.7 | 54.4 |
| HIERLR | 1°, SF2, BM2, 15 passes, 9 subpasses | **73.8** | **181.1** | **49.2** |

Table 5.3: Performance of within-domain training on WOTR's dev set, uniform grid.

## 5.3 Geolocation experiments on WOTR

### 5.3.1 Within-domain geolocation

As with BEADLE, within-domain geolocation was done using the standard methods developed in Chapter 3. Because of the larger number of annotated documents (5,010 vs. 408 for BEADLE), it was possible to do an 80/10/10 split. It was expected that accuracy would be greater for this corpus than for BEADLE due to the larger training set, and this is indeed borne out in Table 5.3. Here Naive Bayes performs fairly well, yielding an accuracy at 161km of nearly 73% and a median error of only 78 km. Further gains come from flat logistic regression and especially from hierarchical LR, which increases the accuracy by over a percentage point and drops the median error significantly, to below 50 km.

#### Learning curves

An important issue when considering the annotation of a large resource such as WOTR is the amount of annotation necessary to achieve a given result. This is due to the high cost of annotation, and applies especially to the digital humanities, where budgets are typically not as much as are available in the sciences, are are frequently shrinking. One way to quantify this is through *learning curves*, which show the tradeoff between amount of annotation and performance. Figure 5.4 shows such curves for median error and accuracy at 161km, for uniform and *k*-d grids and various grid sizes.

As can be expected, the steepest part of the curve is near the beginning, where there is the least amount of data; but performance continues to improve for quite a ways. Performance improvement in median error tapers off around 50% of the total training data (around 2,000 instances) and plateaus at around 75%, but accuracy starts to taper around 75% and continues to improve almost up to 100%.

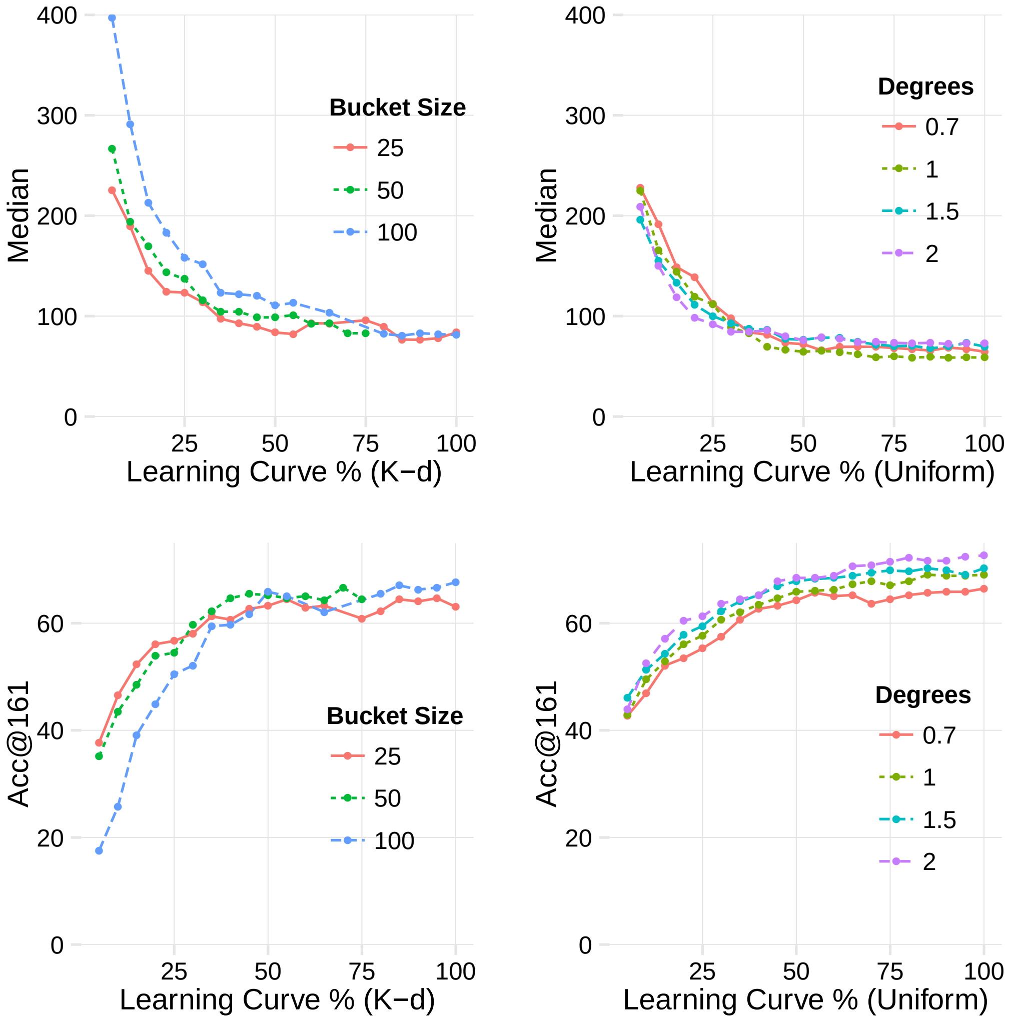


Figure 5.4: Graph of learning curves over annotated subset of WOTR, randomized 80/10/10 over individual data instances.

The point at which "enough" data produced and "sufficient" performance reached is of course subjective. It could be argued, for example, that the point at which the curve tapers off is the place to stop; however, this can't be determined without annotating significantly beyond the point at which tapering is first observed, to ensure that the tapering is a real phenomenon. Alternatively, one could pick a particular performance level and annotate up until that level is achieved. For example, it could be argued that a median error of 100km is sufficient for making reasonable conclusions given the scale of the Civil War, which covered an area of well over 1,000km by 2,000km.

In reality, however, the necessary accuracy depends on the application in question. For distinguishing North from South, a 100 km error might be more than accurate enough, but for distinguishing Washington, DC from Virginia, it is completely insufficient. In fact, if you need to distinguish both at the same time, then the entire concept of measuring error by distance may be inappropriate. (For this reason, the geographic topic models discussed in §5.5.1 make use of *k*-d trees, even though their performance in both median error and accuracy@161 is worse than uniform grids. *k*-d trees allow for an adaptive error rate that depends on the density of data in a particular place. The geographic topic models make use of hand-drawn regions whose most critical boundaries occur in areas of high document density, which correspond with areas where *k*-d trees will perform with significantly greater accuracy than uniform grids. )

It is also important to notice that the best grid size varies somewhat depending on the portion of the graph in question and on whether median or accuracy is considered. For the uniform-grid median, for example (lower-left graph in Figure 5.4), the coarsest grid size of 2° performs the best at points along the learning curve less than 50% of the total training data but the worst above this, while the pattern is reversed for the finest grid size of 0.7°. To the extent that a similar pattern can be observed for the *k*-d tree grid, it is much less consistent, if present at all.

The reason for this is the tradeoff in the uniform grid between the increased resolution that comes from a finer grid with smaller cells and the more accurate per-cell language models that come from a coarser grid with more training documents per cell. With comparatively little training data, the finer grids have too few training documents per grid cell and performance suffers, whereas with more training data the finer grid cells have sufficiently accurate language models that their smaller size ensures a lower median. The lack of a similar pattern in the *k*-d grid is due to the fact that this grid by its nature tries to keep the number of documents per cell relatively constant, instead scaling the cells as necessary.

The increased jumpiness of the *k*-d tree results presumably reflects the nature of the *k*-d division algorithm, which operates in a greedy fashion and divides a cell above the bucket size into two equal-sized smaller cells, which may have an unbalanced number of documents in them. (For example, the worst case for a bucket size of 25 is a division that leaves 24 documents in one subcell and 1 document in the other.) In general, these unbalanced cells will get balanced out by later points added to the cell, but this will not happen for the last few cell divisions. The additional points added when going from e.g. 25% to 30% of the total might by chance trigger a number of unbalanced cell divisions, which then get balanced when going from 30% to 35%; this would explain, for example, the temporary flattening in the median error at 30% along the learning curve for a bucket size of 100, followed by a sudden drop at 35% (upper left graph in Figure 5.4).

Note that my *k*-d code has an alternative splitting mechanism (termed *median* rather than *halfway*), which equalizes the number of documents per subcell rather than the size of the subcells when a cell division is made. However, this won't necessarily help the jumpiness issue, as it has its own worst case, where the two subcells are of extremely different sizes, with one taking up nearly all the area of the parent cell. As this cell has effectively the same area as its parent but half the number of documents, its language model is likely to be of lower quality and the overall performance will decrease, producing a spike or temporary flattening just as above. Indeed, exactly this pattern is in fact observed in Figure 5.5, and furthermore the overall performance is slightly worse, something that I observed in previous experiments as well.

An additional pattern of interest can be seen in the fact that the finest uniform grid size of 0.7°is tied for the best at the 100% point along the learning curve under median error, but consistently does the worst under accuracy at 161km, for reasons that are not completely clear.

**Split by volume** All of the above results were achieved with a split by individual document. Specifically, the full set of annotated articles was permuted randomly and then document assigned to training, dev and test in a round-robin fashion. This was the simplest way to perform the split.

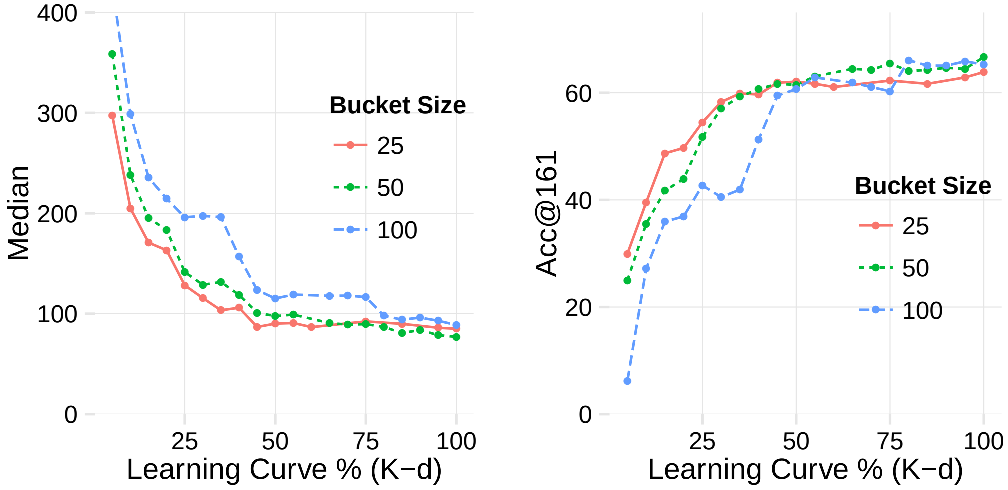


Figure 5.5: Graph of learning curves over annotated subset of WOTR, randomized 80/10/10 over individual data instances, for *k*-d trees using median cell-division method.

However, a potential criticism is that it is more realistic to keep complete volumes together when splitting. Each volume has somewhat different properties from the other, and thus a split by individual document is likely to perform significantly better than a split by volume. This is especially the case given that only a fraction of each volume was annotated, consisting of a set of sequential documents (typically stretching about 25 pages in length, for each of the 126 volumes). This means that the training data may be effectively representing the distribution of the test data but not necessarily that of the WOTR corpus as a whole, including the annotated data.

This especially applies when considering the perspective of using a trained model to do predictions on some of the remaining unannotated data, and also when deciding what volume to annotate next and whether enough data has been annotated.

To address this criticism, I did additional runs with a split by volume. Because these are different splits, direct comparison of results is tricky. Preliminary investigations revealed that there was significant variation in the results with different permutations, particularly with the splits by volume. As a result, both the split by document and split by volume were computed 10 times over 10 random permutations and averaged. A further comparison was done using a split by volume with the order of volumes along the learning curve following the actual order that data was received from

| Corpus | Uniform | | | | *k*-d | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Parameters | A@161 | Mean | Med. | Parameters | A@161 | Mean | Med. |
| NB | 2.25° | **51.3** | 438.0 | **148.8** | BK300 | **54.5** | 396.7 | **123.0** |
| FlatLR | 1.5°, 25 passes | 48.1 | 402.8 | 173.6 | BK1000, 10 passes | 48.7 | 380.0 | 169.0 |
| HierLR | 1.5°, SF2, BM1, 25 passes, 12 subpasses | 48.7 | **395.6** | 169.3 | BK1000, SF6, BM1, 10 passes, 5 subpasses | 48.7 | **376.3** | 165.1 |

Table 5.4: Performance of cross-domain training (training on EnWiki13, testing on WOTR's dev set), split by data instance.

the annotators. It was thought that this would produce better results than simply randomly choosing volumes, since the volumes were hand-selected to produce as wide-ranging a distribution of data as possible as early as possible (see §2.3.2). However, doing this ended up yielding no benefits compared with a random permutation of volumes. See Figure 5.6 for the results.

### 5.3.2 Cross-domain geolocation and domain adaptation

#### Domain adaptation

Interpolation between pure in-domain and pure out-of-domain models was performed. These results are for uniform Naive Bayes, where the uniform grid eliminates many issues that would otherwise appear when matching up the grid cells. Results are shown in Figure 5.7.

Note in particular how much gain is yielded from very little in-domain training data. In this case, with only 10% of the total training data, the accuracy already jumps up from 49% to 62% for interpolation factor 0.5, more than halfway to the maximum possible accuracy of 73%. Already, 70% accuracy is reached with only 30% of the total training data. Contrast the results from pure in-domain data, which needs 30% of the total just to reach 62% accuracy, and doesn't reach 70% accuracy until close to 65% of the total has been processed. (Note that e.g. an interpolation factor of 0.3 means a weight of 0.3 for Wikipedia, and a weight of 0.7 for the in-domain data).

Domain adaptation was also performed using various types of concatenation, including Daume's EasyAdapt (§5.1.2), with various weightings between the in-domain and out-of-domain data (i.e. Wikipedia). Results all along the learning curve were consistently disappointing, and were unable to exceed the values computed on pure in-domain data. (This result is puzzling.)

It is also interesting that domain adaptation was unable to exceed the best overall values reached using pure in-domain data when all the data was used. It is possible that this reflects the dif-

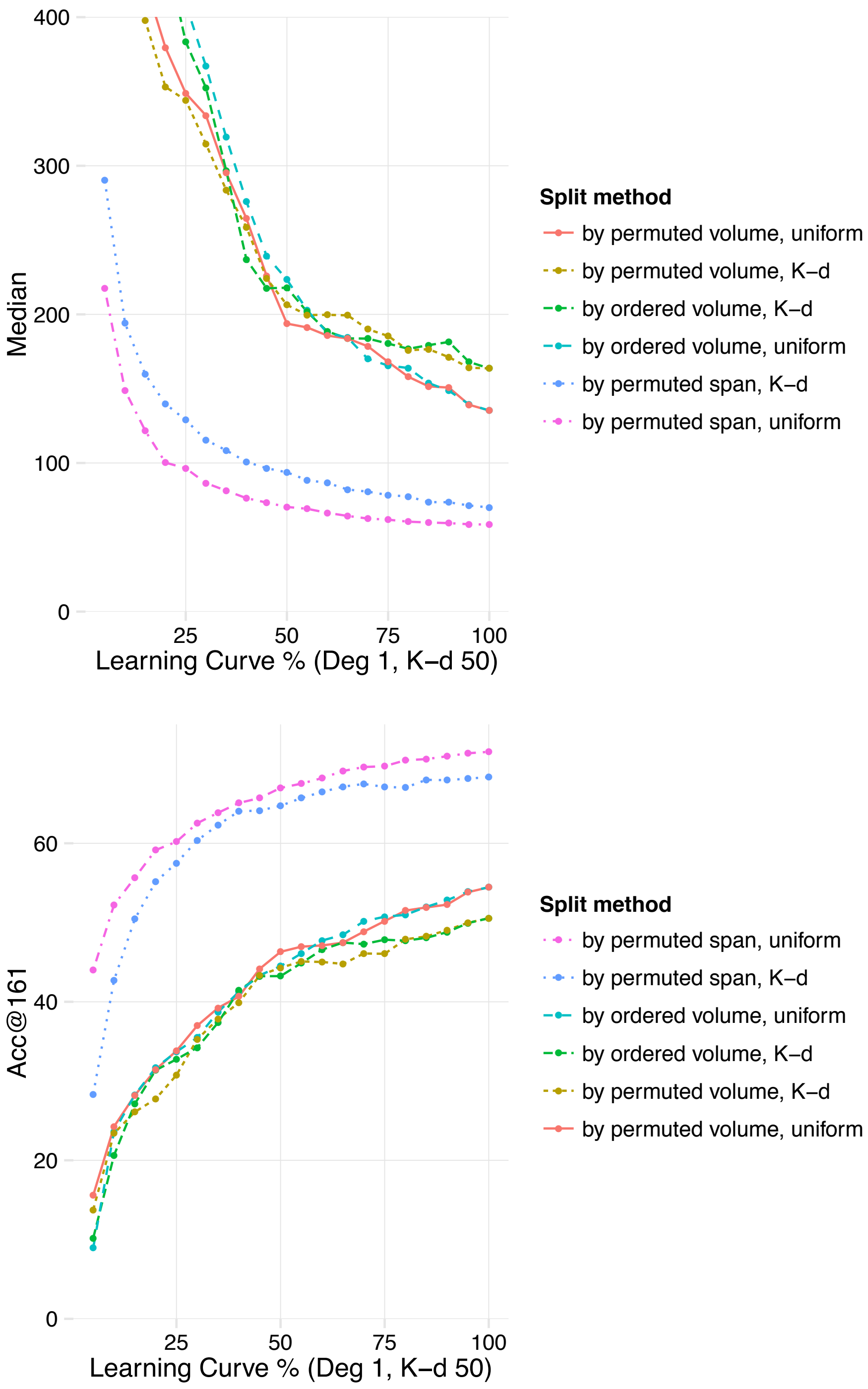


Figure 5.6: Graph of learning curves comparing different methods of creating the training/dev/test splits, averaged over ten permutations.

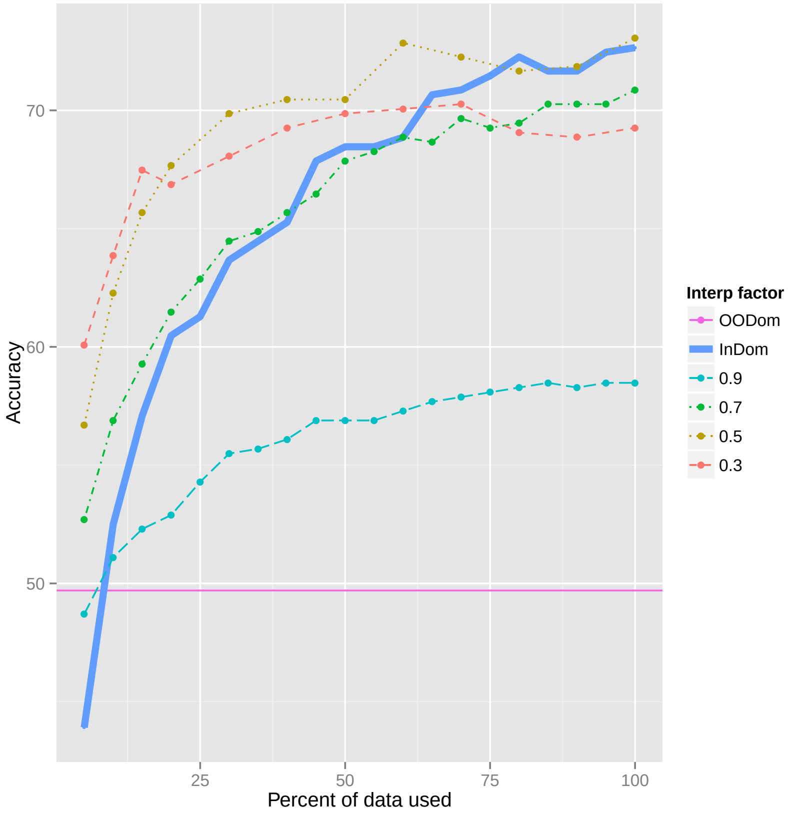


Figure 5.7: Interpolation results between Wikipedia and the in-domain training data of WOTR, applied to WOTR's dev set. *InDom* is pure in-domain, *OODom* is pure Wikipedia, and the interpolation factor is the weight assigned to in-domain data when interpolating.

ference in distribution between the in-domain and out-of-domain data, which is significant enough that it cancels out the boost that the extra information contained in the out-of-domain data would otherwise provide. This suggests that better results might be obtained, for example, by reweighting Wikipedia using the distribution of the in-domain data. This could be done purely using the annotated portion of the data. However, it is also possible to imagine using the unannotated portion of the data, in connection for example with the distribution of toponyms found in that portion of the data. §5.4, following, discusses using such distributions as proxies for the actual distribution of the unannotated portion of WOTR.

## 5.4 Investigation of the data in the *War of the Rebellion*

One of the main purposes of annotating WOTR is to be able to use the annotations to geolocate the remainder of the data and investigate the data to make conclusions applicable to the digital humanities.

The actual distribution of the full data of WOTR, including the unannotated portion, may not match the distribution of the data as predicted by a model trained on the annotated portion. In this section I investigate the distribution of the data as produced by this model. When plotted on a KML graph, it appears very similar to the distribution of the training data, as can be seen by comparing Figure 5.8 and Figure 5.9. This seems somewhat suspect, suggesting that the distribution of the training data is overly biasing the predictions made by the model. As a result, I look for other ways of deriving an approximate distribution of the full data. I propose the following:

1. One simple way is to do toponym resolution, and pick the first toponym that appears in the document; this is in many cases the location found in the header, which is often correct. (Actually I do something slightly more sophisticated, preferring cities over states and states over the toponym *Washington*. I justify this in §5.4.2.) For toponym resolution I use a simple but effective method using Wikipedia prominence. This serves a bit like a population baseline but is better because it reflects a broader concept of prominence than simply population at a particular point in time. (Future research could repeat these results using modern toponym

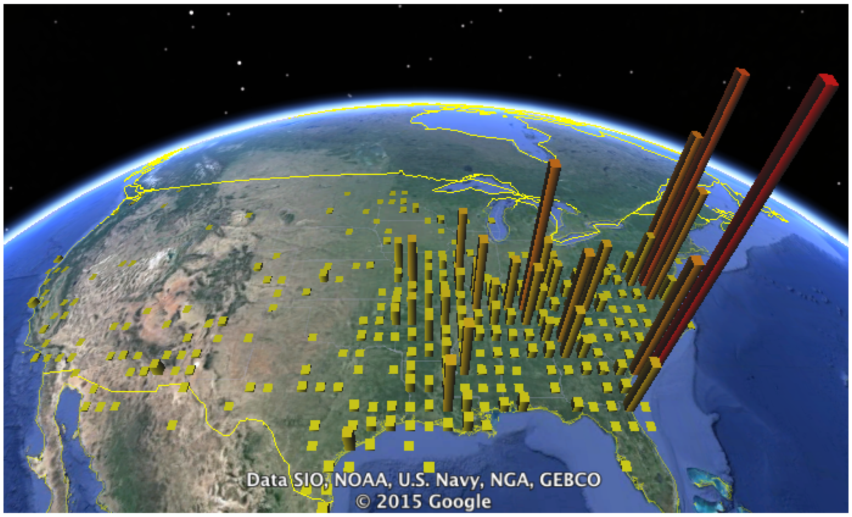


Figure 5.8: Graph of distribution of locations for annotated subset of WOTR.

resolution methods such as in (Speriosu, 2013; DeLozier et al., 2015).)

2. I also use another corpus of military actions developed by Scott Nesbit from *War of the Rebellion*. This is a different corpus but has the same source so in some way should reflect the distribution of the source.

I then compare these alternative distributions to the actual distribution derived, as a rough way of "evaluating" the accuracy of the distribution produced from the training-data model.

I discuss the possibility of then combining these distributions, especially the toponym-resolution one, with the training-data distribution, perhaps (e.g.) by using the toponym-resolution distribution as a Naive Bayes prior for the training-data model.

### 5.4.1 Evaluating the unannotated portion using the annotated portion

Figure 5.9 shows the geographic distribution of these documents as predicted using Naive Bayes run over 100% of the annotated material at 1°, with the height of a cell indicating the relative density of documents in that cell.

As mentioned above, the two distributions are quite similar, and both quite peaked. If anything, the distribution of predicted locations is even more peaked than that of the annotated locations.

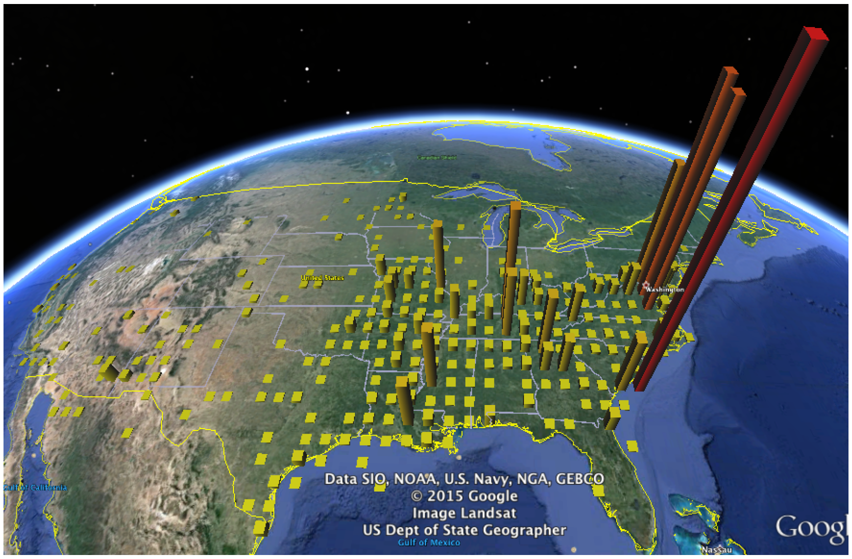


Figure 5.9: Graph of distribution of predicted locations for articles in WOTR.

(This phenomenon was also observed with Twitter, for example when investigating GEOTEXT, where logistic regression resulted in even more peaked distributions as compared with Naive Bayes.)

Another way to graph the distribution of locations in the annotated subset is to display the distribution of grid cells computed when using a *k*-d tree to do prediction. Figure 5.10 shows this distribution, and Figure 5.11 shows a zoomed portion of the distribution for the main area of fighting during the Civil War. The centroids of the cells are shown as blue dots, and the theaters of war that are used for constructing geographic topic models (§5.5.1) are shown with heavy red lines. Certain things can be seen in this view that are not apparent in the Google Earth distributions shown above. For example, the density of documents in parts of Virginia (but not other parts) can clearly be seen, along with Sherman's march across northwest Georgia and various major cities (e.g. Washington DC, Richmond, Charleston, Mobile, St. Louis).

## 5.4.2 Other ways of evaluating full data

### Corpus of military actions

One way to further investigate the predicted geographic distribution is to compare to an outside resource consisting of a list of military actions (skirmishes, battles, etc.) annotated with latitude and

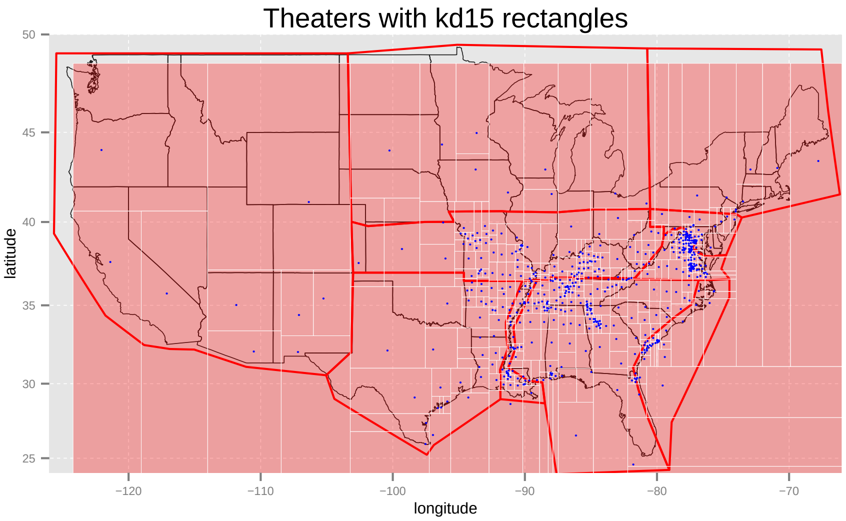


Figure 5.10: Graph of *k*-d tree grid cells derived from the annotated subset of WOTR with a bucket size of 15.

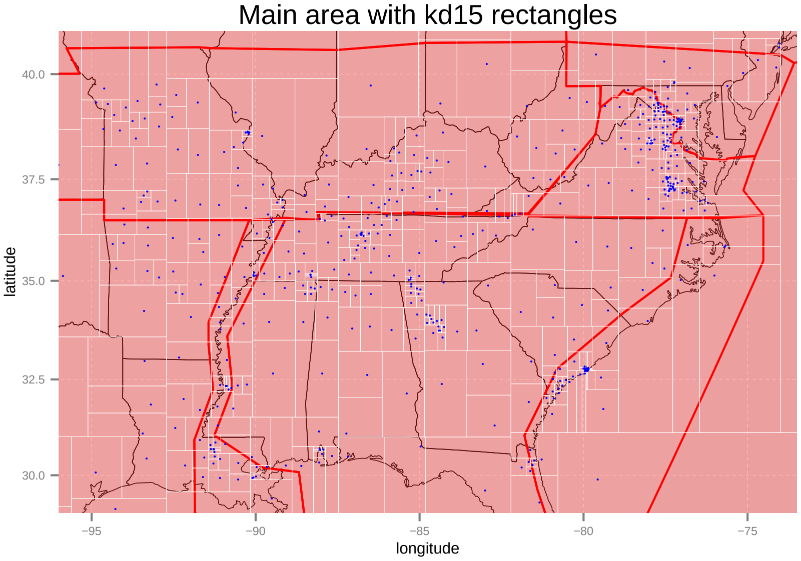


Figure 5.11: Graph of *k*-d tree grid cells derived from the annotated subset of WOTR with a bucket size of 15, zoomed in to cover the main area of fighting during the Civil War.

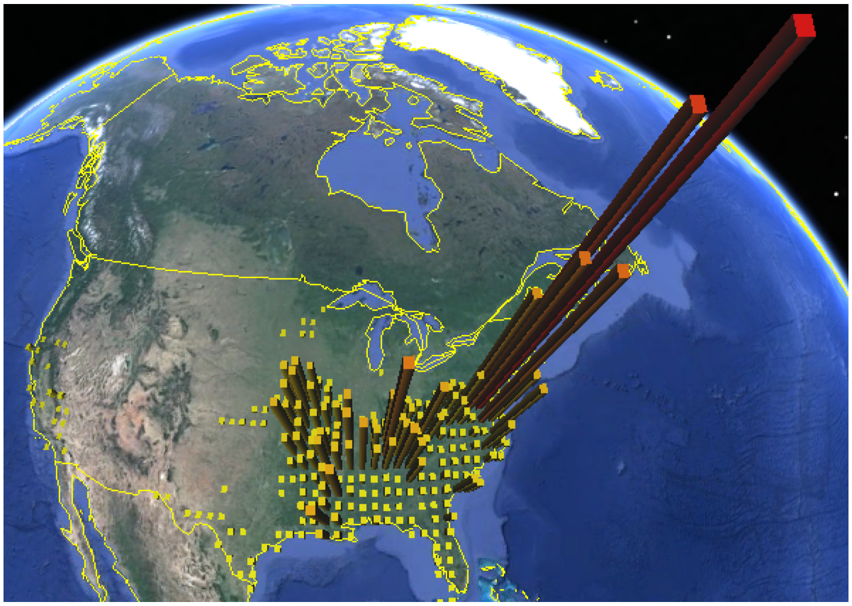


Figure 5.12: Graph of distribution of military actions in CWRED.

longitude. Such a corpus was provided to me by Professor Scott Nesbit of the University of Georgia, and was extracted from *War of the Rebellion* and from *Dyer's Compendium* (more properly, *A compendium of the War of the Rebellion*, written by Frederick H. Dyer and published in 1908). I term this corpus CWRED. The distribution is shown in Figure 5.12. As can be seen, the distributions are quite different, with much more sparseness in the predicted distribution of the WOTR text over much of the East (except for parts of Virginia, coastal South Carolina, and the Mississippi River), but with relatively more coverage of the West. This can be seen further in Figure 5.13, which plots a heatmap comparing the two distributions, with areas of higher density in CWRED indicated in green, while areas of higher density in WOTR appear in red. Those near white (and surrounded by a solid white border) are those with approximately equal density, while areas in gray without a solid white border lack any documents in either distribution.

### Toponym resolution as a proxy for document geolocation

Another way to get a different measure of the true distribution of locations in the WOTR documents is to use toponym resolution. Toponyms are often highly indicative of the location of a document,

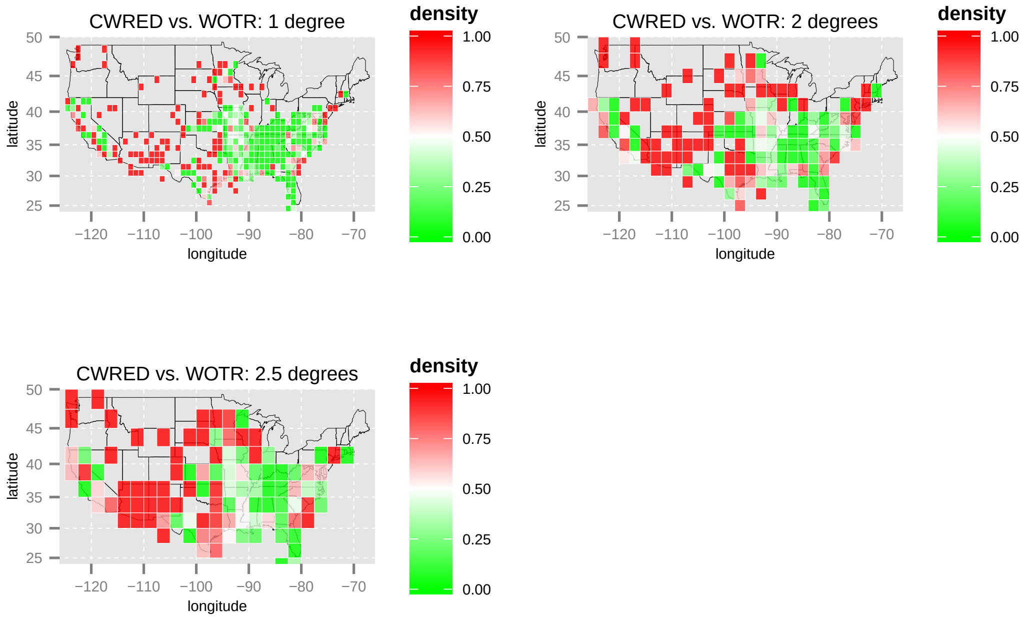


Figure 5.13: Heatmap comparing relative density of military actions in CWRED vs. geolocated documents in WOTR; green indicates areas with higher density in CWRED, and red indicates the same for WOTR.

and algorithms exist for finding toponyms in a document and resolving them to a latitude/longitude coordinate. I proceed as follows:

- I run a named entity recognizer (NER) on the volumes in WOTR.[1]
- I then use the *prominence* of locations in Wikipedia to resolve ambiguous toponyms. Prominence in Wikipedia is measured by counting the number of links from an ambiguous toponym such as *Springfield* to each distinct geolocated article; the article with the most number of such links is considered the resolution of that toponym. The set of allowable locations is restricted to a bounding box covering the U.S.—specifically, from $(25, -126)$ to $(49, -60)$, as is used elsewhere this dissertation. This eliminates many potential mappings likely to be spurious in a text focused on the U.S. (e.g. Cairo, Egypt in place of Cairo, Illinois or Paris, France in place of Paris, Texas).
- I then choose the first resolvable toponym in a given document, while dispreferring toponyms referring to states and dispreferring the toponym *Washington* even more. Specifically, the first resolvable toponym that is not the name of a U.S. state is chosen; if there is no such toponym, I fall back to the first name of a state, excluding the toponym *Washington*; finally, finally, *Washington* is used if it exists (and is manually mapped to Washington, D.C., since the highest-prominence Washington in Wikipedia is the U.S. state of Washington). The reason for choosing the first resolvable toponym is that the location of the sender is normally named in the header of each document, and this very often corresponds to the overall location of the document. The reason for dispreferring states is due both to their inherent inaccuracy when using a point-based mapping (typically states are "resolved" to their capitals) as well as the frequent mention of state-based militias (e.g. the 3rd Illinois or 20th Pennsylvania), which are often found serving in locations nowhere near their home state. The reason for especially dispreferring Washington is that a great deal of documents concern communication to or from the central government in Washington, D.C., which is frequently mentioned in the header or greeting of the document but is rarely the geographic subject of the document.

[1] Due to resource limitations, this was done only on pages 100–199 of each volume, and it was this subset that was used for these experiments.

| Location | Link Count |
|---|---|
| Springfield, Massachusetts | 1292 |
| Springfield, Illinois | 746 |
| Springfield, Missouri | 662 |
| Springfield, Ohio | 290 |
| Springfield, Oregon | 191 |
| Springfield, Virginia | 108 |
| Springfield, Kentucky | 101 |
| Springfield, Vermont | 87 |
| Springfield Township, Union County, New Jersey | 84 |
| Springfield, West Virginia | 82 |
| Springfield, IL | 63 |
| Springfield Township, Delaware County, Pennsylvania | 60 |
| Springfield, Tennessee | 59 |
| Springfield College | 45 |
| Springfield, Georgia | 40 |
| Battle of Springfield (1780) | 34 |
| Springfield, Colorado | 34 |
| Springfield, MO | 31 |
| Springfield, South Dakota | 31 |
| Springfield, New Hampshire | 30 |
| Springfield, New York | 29 |
| Springfield Township, Montgomery County, Pennsylvania | 27 |
| Springfield, Minnesota | 25 |
| Roman Catholic Diocese of Springfield in Massachusetts | 23 |
| ... | ... |

Table 5.5: Wikipedia link-based prominence of different resolutions of the toponym “Springfield”.

Using Wikipedia prominence as a toponym resolution technique is similar to the standard technique of using population to resolve locations but is potentially better. The idea behind using population figures is that it serves as a proxy for the likelihood that a given ambiguous toponym in an arbitrary text will resolve to a given location. This is similar to Wikipedia prominence, but has some well-known pitfalls, some of which are avoided when using prominence. For example, population figures may change significantly over time, and among a set of locations, the one with current highest population may be different from the one with the highest population in the 1860’s. Furthermore, some low-population places may have historical importance disproportionate to their population. Although it is true that the prominence of a location may also change over time, the presence of large amounts of history-related articles in Wikipedia is likely to mitigate this.

An example of the relative prominence of different resolutions of the toponym “Springfield” is shown in Table 5.5.

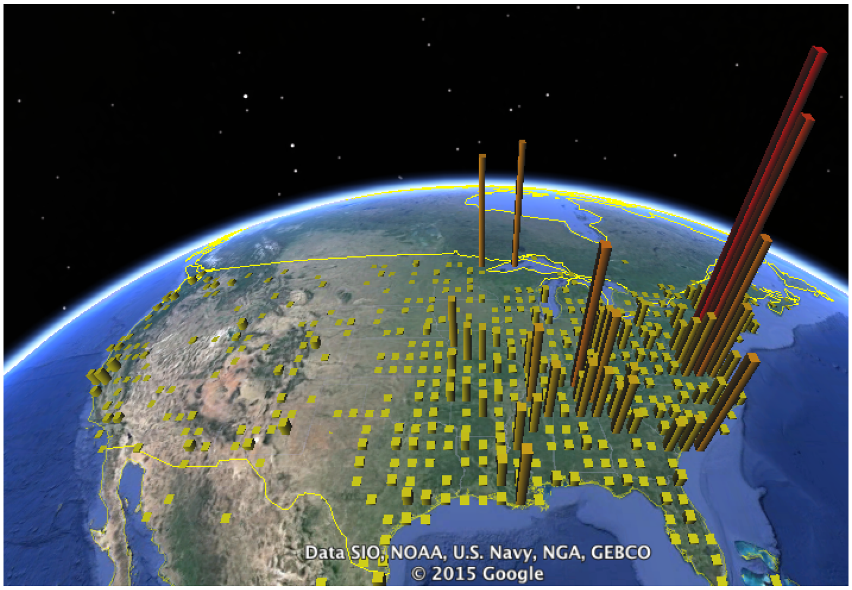


Figure 5.14: Graph of distribution of locations for documents in WOTR using toponym resolution as a proxy for document geolocation.

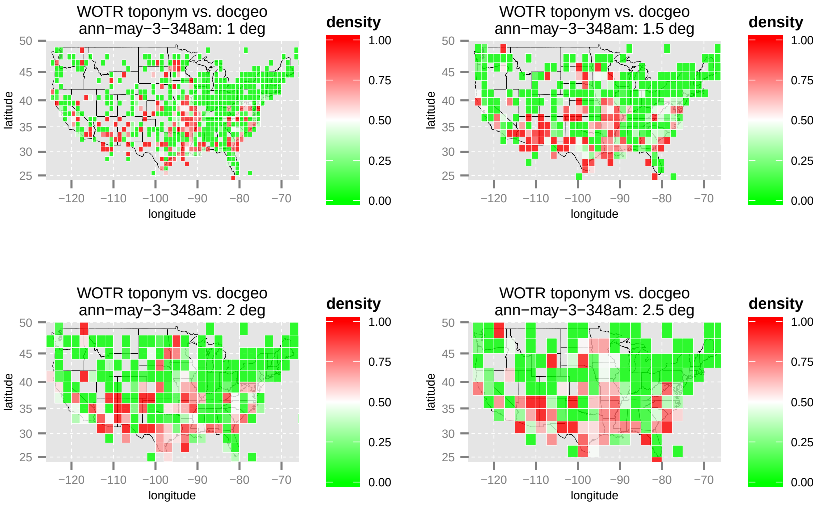


Figure 5.15: Heatmap comparing relative density of toponym resolution as proxy for document geolocation vs. actual document geolocation.

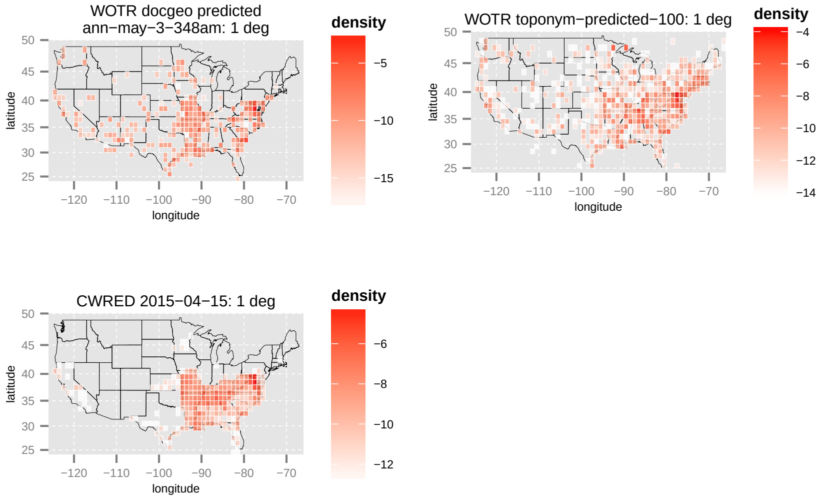


Figure 5.16: Heatmap showing the distributions of CWRED, toponym resolution as proxy for document geolocation, and actual document geolocation.

Figure 5.14 shows the distribution of locations using toponym resolution as a proxy for document geolocation, and Figure 5.15 is a subtractive heatmap comparing this distribution to the distribution of locations predicted directly with document geolocation. As can be seen by comparing the graphs in Figure 5.14 and Figure 5.8 and the subtractive heatmaps in Figure 5.15, the distributions are not all that different.

Examining the KML distribution of using toponym resolution as a proxy for document geolocation, some patterns can be noticed. The overall distribution with a big peak in eastern Virginia and Washington DC, and a smaller peak in Tennessee/Northern Georgia, with a few other peaks (e.g. near Fort Sumter), looks reasonable, as it tracks the known locations which saw heavy battle activity during the Civil War. However, the distribution across the Northern states is probably spurious (presumably due to Northern towns with higher prominence than same-named Southern towns), especially the two spikes in northern Minnesota and Lake Superior (see below). Possible reasons for spuriousness:

- non-toponym words getting treated as toponyms by the NER, which then are mapped to locations due to appearing in anchor text pointing to a geolocated Wikipedia article (e.g. words like “island”—I tried to exclude a lot of those manually, but I don’t think “island” was in my list);

- states getting resolved to their capitals;
- mention of state-based army divisions getting treated as toponyms (e.g. “Illinois” in “3rd Illinois”);
- confusion of river names and states (e.g. “Mississippi” marked as a toponym in the phrase “the Mississippi”);
- weirdness in the Wikipedia-assigned prominence (e.g. “Saint Louis” has “St. Louis County, Minnesota” as its most prominent entry, which probably explains the spike in Minnesota; the spike in Lake Superior is explained by “America” and “Cumberland” mapping to ships in Lake Superior for various weird and fixable reasons).

**Improving the prediction distribution over the unannotated corpus**

It should be possible to make use of these other distributions when doing document geolocation. One possibility is as a prior distribution, at least when doing Naive Bayes. This would serve to bias the document geolocation choice without completely forcing it to conform to the distribution of the toponym resolution proxy, which may in itself have errors. One potential issue is that this may overweight in favor of the prior distribution. This is what happens, for example, when you do normal document geolocation and use a standard prior based on the number of training documents in a cell, with the result that the prediction accuracy gets much worse than just using a uniform prior (see §5.2.2).

Another possibility that may work is simply to use a domain adaptation model, one that mixes (interpolates) the training and Wikipedia data sets. It would seem that this model might produce better accuracy on the full data set even if it doesn’t produce better accuracy on the test set because it will simply have more data available on more locations and people.

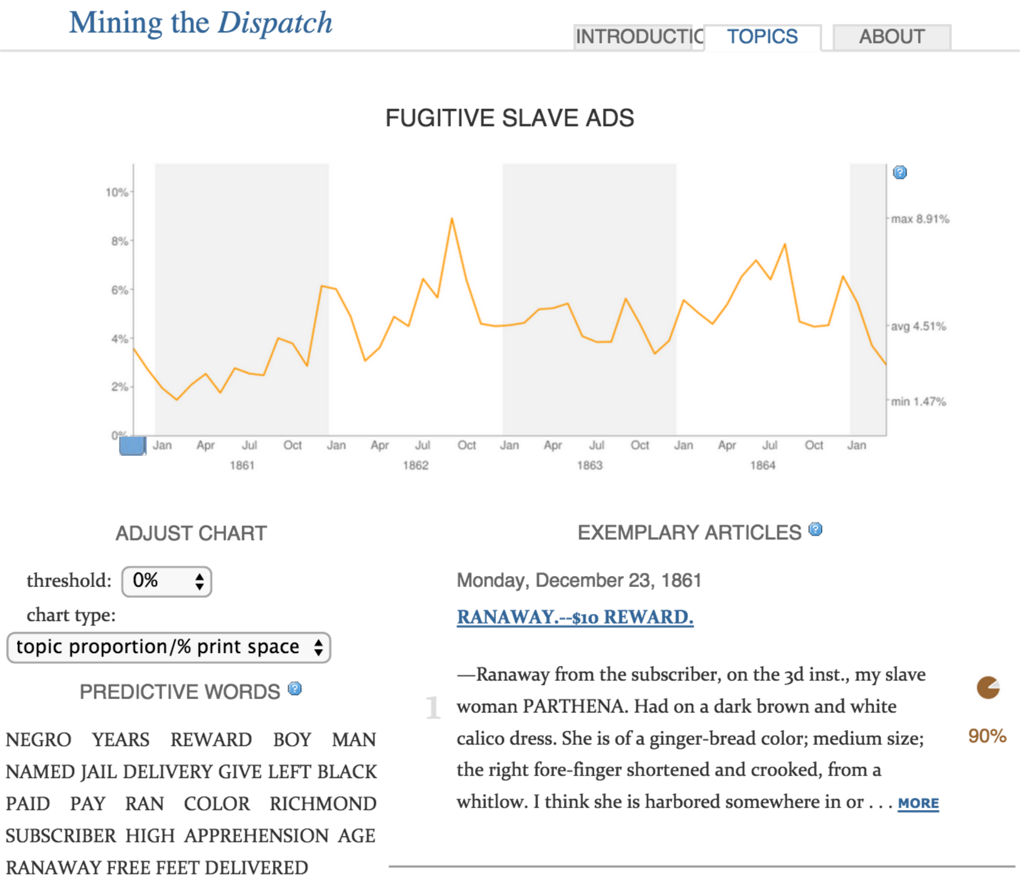
A basic issue here in determining the efficacy of these suggestions is that it is not easy to measure their predictive accuracy. The fundamental reason for doing them is due to the assumption that the distribution of the training data, due to its small size and possibly other biases, doesn't accurately reflect the actual distribution of the full data. Because the dev and test sets are drawn from the same set of data and produced using the same process as the training set, they are likely to share its distribution, and thus any attempt to "broaden" the predictive distribution beyond what is found in the training data is thus likely to cause results to be worse on the dev and test sets, or at least not to improve. (Compare the results in §5.3.2, where domain adaptation improved results along the learning curve but not when all data was used.) One possible exception might come from split-by-volume experiments (§5.3.1), which more closely mimic a situation where the training and test data are drawn from different distributions.

## 5.5 Application of topic models to *War of the Rebellion*

As noted in §1.6 and in the introduction to this chapter, one of the main concerns of the digital humanities is exploratory data analysis of large-scale textual datasets consisting of primary sources in the humanities. This allows for *distant reading*, the large-scale analysis of entire corpora in order to glean patterns inside them, and produces conclusions of a sort that are difficult or impossible to achieve with traditional techniques involving *close reading* and analysis of a small set of primary sources (Rhody, 2012). Among NLP tools, one of the most frequently used is topic modeling (Meeks and Weingart, 2012). Topic modeling in the digital humanities is primarily done using MALLET (McCallum, 2002) along with postprocessing tools such as Paper Machines (Johnson-Roberson and Guldi, 2014; Crymble, 2012).

Some subjects of digital humanities analyses done using topic models are figurative language in ekphrasis poems, i.e. poems about the visual arts (Rhody, 2012); the history of literary scholarship as seen in *PMLA*, the primary literary journal of the Modern Language Association of America [2] (Goldstone and Underwood, 2012); and the patterns of publication in a Virginia Civil War newspaper—the *Mining the Dispatch* project from the Digital Scholarship Lab at the University of Richmond (Nelson, 2015). This project applies topic modeling to the Civil War-era issues of the Richmond *Daily Dispatch* newspaper, in order to explore the political and social life of wartime

[2] https://www.mla.org/pmla



Mining the *Dispatch*

FUGITIVE SLAVE ADS

ADJUST CHART

threshold: 0%

chart type: topic proportion/% print space

PREDICTIVE WORDS

NEGRO YEARS REWARD BOY MAN NAMED JAIL DELIVERY GIVE LEFT BLACK PAID PAY RAN COLOR RICHMOND SUBSCRIBER HIGH APPREHENSION AGE RANAWAY FREE FEET DELIVERED

EXEMPLARY ARTICLES

Monday, December 23, 1861

RANAWAY.--$10 REWARD.

1 —Ranaway from the subscriber, on the 3d inst., my slave woman PARTHENA. Had on a dark brown and white calico dress. She is of a ginger-bread color; medium size; the right fore-finger shortened and crooked, from a whitlow. I think she is harbored somewhere in or . . . MORE

90%

Figure 5.17: Sample of the web site *Mining the Dispatch*, with a topic from a topic model applied to the Richmond *Daily Dispatch*, hand-labeled as *Fugitive Slave Ads*.

Richmond. Specifically, a topic model was applied to the newspaper articles in question, the topics were manually labeled, and a web application was created allowing the topics to be browsed, including the top words of each topic, the relative frequency over time of the topic among all the articles, and the most representative articles for the topic. Figure 5.17 shows a sample of the web application, for the topic *Fugitive Slave Ads*. As suggested by the example of this project, many of the analyses tend to be qualitative in nature.

Other related work has applied mixed-membership models that are analogues of the LDA (Latent Dirichlet Allocation) algorithm underlying topic models to digital humanities task as varied as reconstructing the contents of rooms in ancient Pompeian households (Mimno, 2009) and inferring social rank in Old Assyrian trade networks (Bamman et al., 2013).

Geography is an important part of the digital humanities, especially in the subfield known as the *spatial humanities* (see §1.6, where a number of large-scale mapping projects in the digital humanities are described). An example of the importance of geography in the context of the Civil War can be found in the *Valley of the Shadow* project (Ayers, 2007), which traces the daily life of two American communities, one in Pennsylvania (in the North) and one in Virginia (in the South), across the entire Civil War, including some years before and after. It includes an extensive series of maps, which help to contextualize the letters, diaries, newspaper articles, census and tax records, and other information contained in the project's online archive. One example is an animated map of battles shown in Figure 5.18. This map shows the movement over time of the Confederate 5th Virginia Infantry, from its formation in April 1861 in Augusta County, Virginia to its final engagement at Appomattox Court House in April 1865, where Robert E. Lee surrendered to Ulysses S. Grant. The particular snapshot shows a period in mid 1863, between the battles of Gettysburg and Culpeper Court House. Large yellow/red stars show previous and in-progress battles, while smaller stars show other military engagements.

From what I can tell, however, there has not been too much combination of topic models with geography in the digital humanities. The existing work on geographic topic models is primarily applied to social media (Hong et al., 2012; Ahmed et al., 2013) or news articles (Chang and Blei, 2010). Much of the NLP-related digital humanities work concerning geography has simply involved extracting and plotting toponym mentions in document corpora (Crymble, 2012), using a geoparsing tool such as `geodict` [3].

However, an innovative attempt to apply topic models directly to geographic references is found in Schmidt (2012). This work takes latitude/longitude coordinates of ship movements as found in the ICOADS Maury Collection (National Climatic Data Center, 1998) and treats them directly as "words", with e.g. a ship spending two days in Boston mapping to the two-"word" sequence *42.4,-72.1 42.4,-72.1* and a log of ship movements over time mapping to a "document". Transformed this way, the Maury data produced around 600,000 "words" across 11,000 "documents", which according to the author roughly follow a Zipfian distribution (Adamic, 2000). LDA topic modeling

[3] `https://github.com/petewarden/geodict`

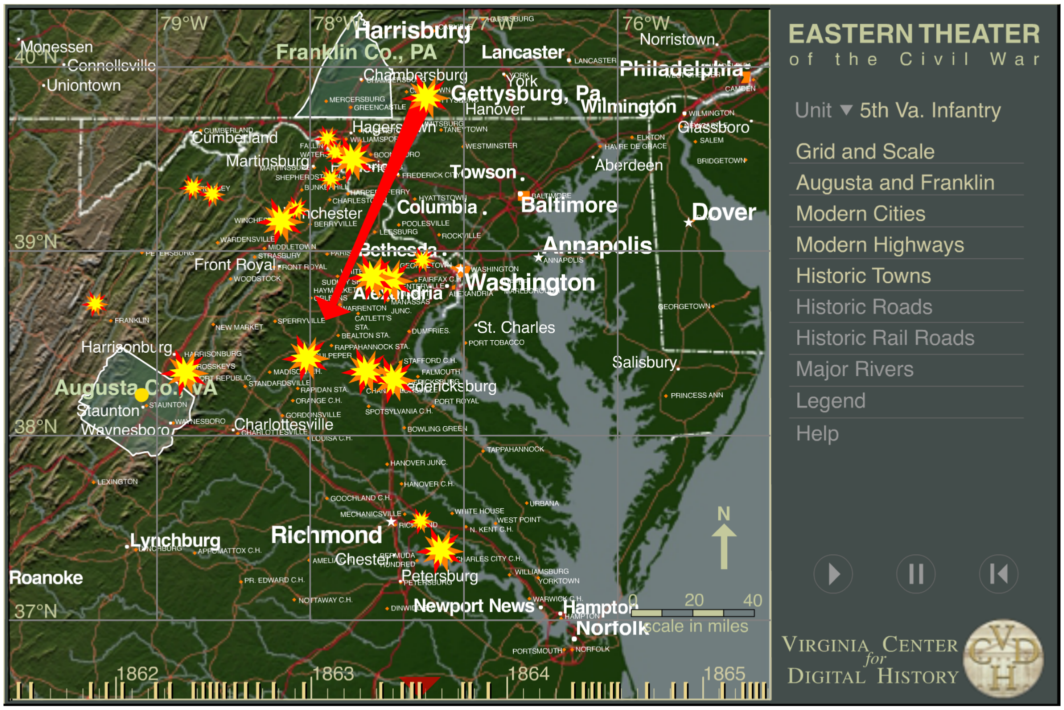


Figure 5.18: Snapshot of an animated battle map for the Confederate 5th Virginia Infantry, showing movement of the unit from the Battle of Gettysburg (July 1–3, 1863) to the Battle of Culpeper Court House (September 13, 1863). From *Valley of the Shadow*.

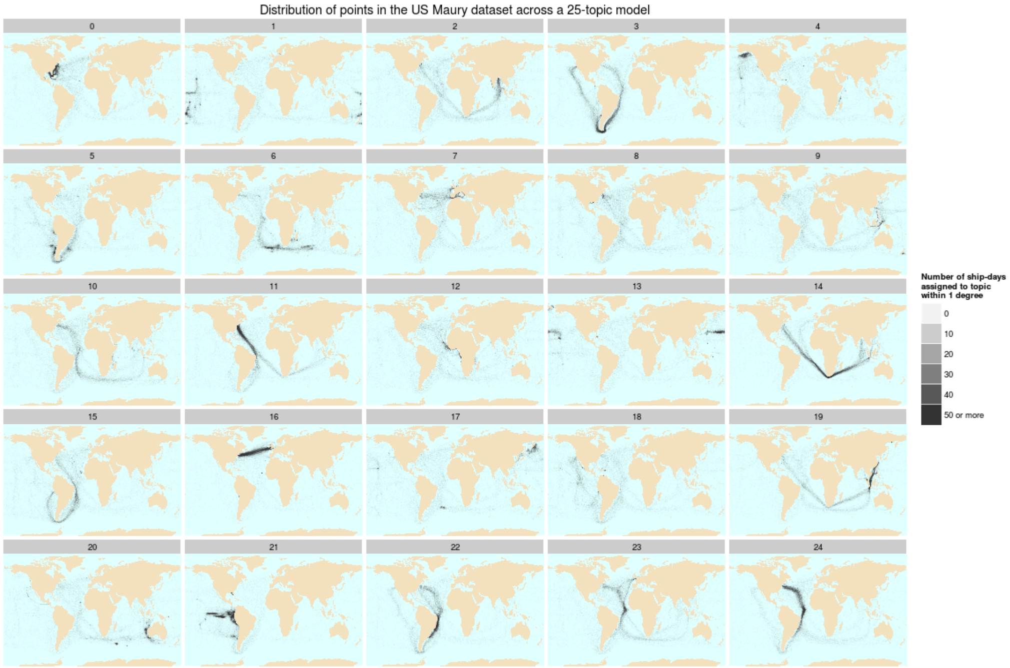

Distribution of points in the US Maury dataset across a 25-topic model
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
Number of ship-days assigned to topic within 1 degree
0
10
20
30
40
50 or more

Figure 5.19: Result of applying LDA to ship logs whose "words" consist of regularized latitude/longitude pairs, based on the ICOADS Maury Collection.

was applied directly to this data, with results shown in Figure 5.19. The resulting topics, colored according to the density of points within a 1-degree square, represent typical 19th-century ship trajectories, with some (according to the author) specifically identifiable as whaling paths, e.g. topics 1, 4 and 5. (Nevertheless, according to the author there are better and simpler techniques for this task, such as K-means clustering.)

In the following sections I show some applications of topic models in the context of WOTR. Because this dissertation focuses on geography, I mostly describe ways of geographically segmenting topic models so as to determine the different ways that people in different regions are talking about various topics. In §5.5.1 I demonstrate a relatively simple but revealing method of geographically tallying up the topics in a topic model to determine which topics are being talked about in which regions. In §5.5.2 I compute *dynamic topic models* both over time and space, showing how a conceptually unified topic can evolve over timeslices or regions, with some terms present only in some slices and other terms present in multiple slices but with different prominence. Expressed differently, §5.5.1 provides a way of determining *which* topics are talked about in which regions, and §5.5.2 shows how to determine *how* particular topics are being talked about in those same regions.

I compute topic models over two subsets of WOTR: One consisting of the full set of 255,000 documents, and another consisting of those documents that in some ways reference African-Americans (approximately 11,000 documents). This subset was chosen because it is one of the primary interests of Professor Scott Nesbit, who assisted me in interpreting the output of the topic models. This subset was determined by selecting documents containing the terms *colored*, *slave* (including terms such as *slavery*, *slaveholding*, etc.), *freedman*, *freedmen* and *contraband* (which in the context of the Civil War tends to refer to runaway slaves), along with carefully tailored expressions involving the word *black* (e.g. *blacks* plural as a whole word, *black man*, *black men*, *black soldier(s)*, *black population*, *black labor*, etc.).

### 5.5.1 Geographic distribution of topics in topic models

One question of interest when exploring the geography of the WOTR corpus is whether documents in different regions talk about different topics. To answer this question, I created topic models from the full set of documents in WOTR and divided up the documents according to the locations as predicted using Naive Bayes. The documents were assigned to regions reflecting differing theaters of war, approximately as shown in Figure 5.20. (The rest of the United States was divided up into regions as well, with one region covering the Pacific West, another the Union Midwest, and a third for the Union Northeast.)

Note that an alternative possibility is to jointly infer geographies and topics, as described in Hong et al. (2012) for single-level geographic regions and Ahmed et al. (2013) for hierarchical regions. I do not do this because I think the pre-specified theaters I use are more likely to be meaningful than any automatically-determined regions. Furthermore, the regions inferred by the methods of these papers are represented as multivariate Gaussians and thus will be approximately elliptical, which is insufficient to represent the irregularly-shaped theaters of war as seen in Figure 5.20. This is especially the case in theaters such as the Atlantic seaboard, the Mississippi River, and the Virginia/Maryland border, whose specific, manually-chosen boundaries are very important to ensure

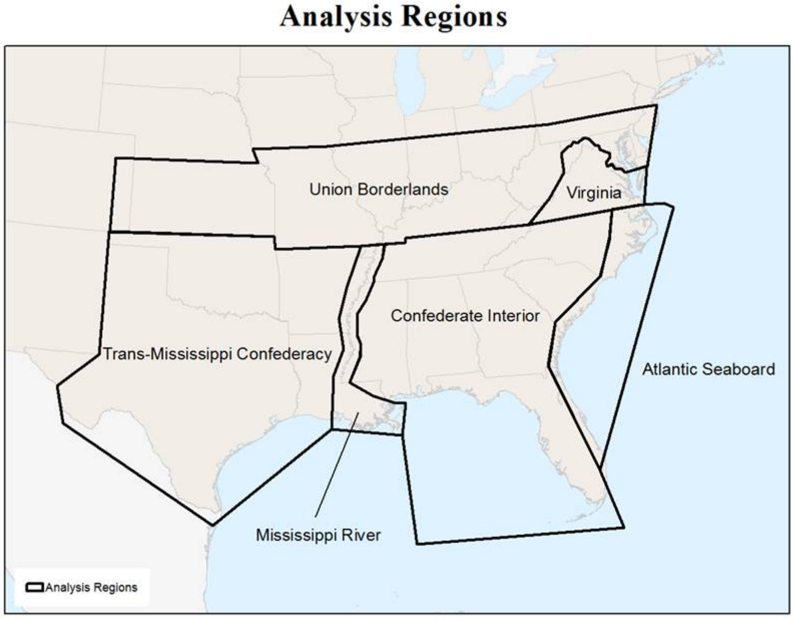


Figure 5.20: Approximate map of theaters of war used to create geographic-based topics in WOTR.

coherency of interpretation.

***k*-d vs. uniform grids** The best results were achieved in §5.3.1 using a uniform 1° grid. However, I choose instead to use a grid based on *k*-d trees (§3.2.2), using a bucket size of 20, which produces approximately the same total number of cells (371) as a 1° uniform grid does (316). This is because *k*-d trees have a number of advantages over uniform grids when using the map of theaters as shown. For one thing, *k*-d trees do not have problems with regions that are narrow and/or have carefully drawn boundaries that do not follow latitude/longitude parallels, such as the Mississippi River and Atlantic Seaboard regions. Provided that there is sufficient density of data within these narrow regions (which is to be expected, otherwise there would be no point in drawing these regions in this fashion), *k*-d trees will automatically shrink the size of their grid cells, allowing for cells that approximately track the border of these regions.

This advantage is even more pronounced in cases of important cities that occur right along a border, for example Washington, DC at the border of Virginia, with the boundary between the Virginia and Union Borderlands regions directly following the Virginia state border. Given the im-

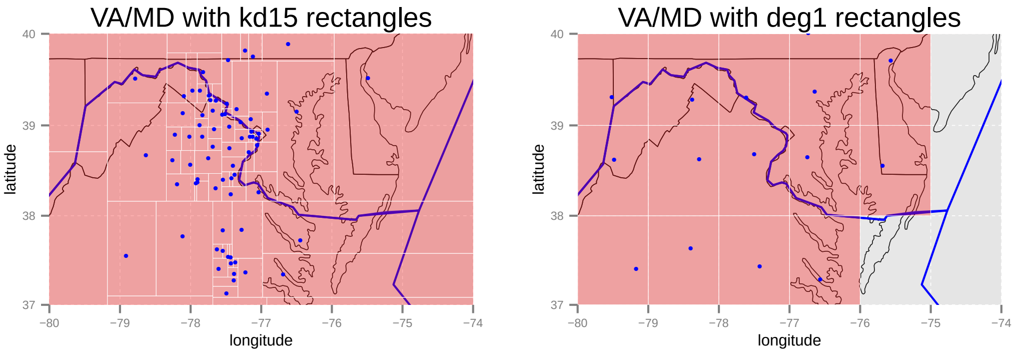


Figure 5.21: Comparison of two sets of grids for the Virginia/Maryland area, with theater divisions shown. Blue dots show the centroids of the rectangles.

portance of Washington, DC and Virginia in the Civil War, it is not at all surprising that there is quite a high density of documents in this region. This means that the *k*-d tree grid will correspondingly have quite small cells—small enough, in fact, that four or five cells are allocated for Washington, DC alone, and many more for nearby areas of Virginia. As a result, the grid is able to place documents in the vicinity accurately enough to ensure that they fall on the correct side of the theater region boundary.

Contrast this situation with that of the uniform grid, with fixed-size, relatively large 1° cells. The border is irregular and does not follow latitude/longitude parallels. Hence those cells along the border, including the one containing Washington, DC, are likely to include significant chunks of both regions. This means that they will be unable to place documents very accurately, with consequent smearing of the regions. In particular, Washington, DC, which is the most critical part of the Union Borderlands region, is likely to have all of its documents grouped in with the Virginia region.

This situation is clearly shown in Figure 5.21, which shows a comparison of the Virginia/Maryland theater for a *k*-d tree and a uniform grid. The graph in Figure 5.22 shows a similar comparison across the Southern United States. The blue dots in the graphs show the centroids of each of the cells, indicating where a test document geolocated to a given cell would be placed. As can be seen on the right side of Figure 5.21, Washington, DC is in fact placed in a uniform cell that is primarily located in Virginia and has its centroid in Virginia, whereas on the left side it can be seen that Washington, DC is divided up into multiple small *k*-d tree cells.

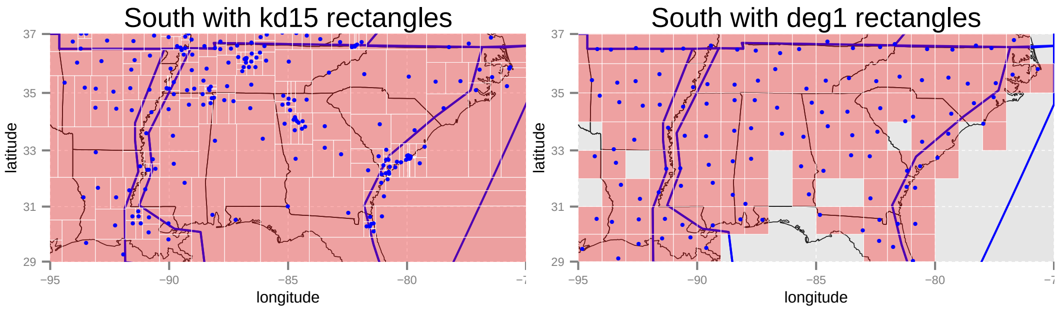


Figure 5.22: Comparison of two sets of grids for the Southern United States, with theater divisions shown. Blue dots show the centroids of the rectangles.

There are no easy ways to avoid this issue while maintaining the uniform grid. For example, one could imagine moving the entire grid slightly south and west so that Washington, DC ends up in the southwest corner of a cell; but doing this puts certain other cells, such as those overlapping the Mississippi River, in a worse position. (This is an instance of the *modifiable areal unit problem* (Gehlke and Biehl, 1934; Openshaw, 1983), where statistical bias is inevitably introduced by trying to divide a continuous space up into cells. The adaptive nature of the *k*-d grid means it is much less affected by this source of bias than a uniform grid.)

One could also imagine shrinking the size of uniform cells, but that would introduce sparsity issues. On the other hand, the *k*-d tree, by its construction, manages to have smaller cells where it counts while still avoiding problems with sparsity. (Recall that the *k*-d tree bucket size used for this work was chosen so that *k*-d tree cells on average have the same number of documents in them as uniform cells.) The way this is achieved is by having very large *k*-d cells in low-density areas, which are usually far away from theater boundaries. Even when such cells cross a theater boundary, this is not much of an issue because those cells by construction do not have very many documents in them.

**Computing topic distributions** Topic distributions were computed for each region by adding up the partial counts of topics for each document in the region and normalizing the resulting values. The top 4 topics are shown for each region in Table 5.6. A similar region-specific distribution of topics was computed for the African-American subset of documents, as shown in Table 5.7.

| Topic | Prop% | Top words |
|---|---|---|
| **Region: Confederate Interior (62299 spans, 25.89%)** | | |
| 15 | 4.06 | general tennessee kentucky nashville tenn chattanooga thomas cumberland ky ohio east morgan gap sherman louisville november knoxville major railroad |
| 12 | 3.23 | miles river road bridge creek camp march marched railroad crossed crossing roads night moved day side encamped cross distance |
| 1 | 3.22 | general mississippi major la west miss jackson vicksburg memphis orleans smith mobile river corinth sherman louisiana grant bayou ala |
| 16 | 3.08 | general commanding adjutant assistant brigadier major headquarters directs respectfully brevet desires acting wm november december wishes hdqrs february indorsement |
| **Region: Union Borderlands (60950 spans, 25.33%)** | | |
| 2 | 3.47 | missouri fort district saint mo indians louis post arkansas kansas militia indian california state price rock ark san curtis |
| 24 | 3.34 | regiment enemy wounded men fire left killed ordered field battle position back time line rear order front woods forward |
| 14 | 3.12 | war secretary department washington prisoners honorable exchange stanton office fort governor letter city parole sir commissary january exchanged edwin |
| 4 | 2.89 | general virginia ferry va brigadier major valley west baltimore winchester harper september md mountain maryland washington gap july jackson |
| **Region: Virginia (46838 spans, 19.47%)** | | |
| 23 | 4.25 | line brigade left enemy position front moved works night division morning rear day road advanced remained skirmishers ordered lines |
| 29 | 4.12 | road house general ford corps station railroad run court division left bridge morning church night junction side fredericksburg plank |
| 37 | 4.11 | general major chief army staff headquarters corps potomac halleck burnside meade geo humphreys warren commanding wright hancock washington grant |
| 24 | 3.45 | regiment enemy wounded men fire left killed ordered field battle position back time line rear order front woods forward |
| **Region: Atlantic Seaboard (20945 spans, 8.70%)** | | |
| 19 | 4.30 | states state united confederate president government military act war governor law excellency congress authority authorities power laws public executive |
| 9 | 4.09 | people country great government hope letter good dear feel make matter desire power policy men respect doubt confidence long |
| 8 | 3.73 | general north carolina south virginia georgia charleston brigadier major district florida anderson savannah beauregard wilmington island johnson james january |
| 27 | 3.62 | battery guns artillery batteries fire enemy pounder gun position inch captain firing pieces fort section ammunition fired shot opened |
| **Region: Trans-Mississippi Confederacy (20381 spans, 8.47%)** | | |
| 2 | 3.52 | missouri fort district saint mo indians louis post arkansas kansas militia indian california state price rock ark san curtis |
| 32 | 3.37 | orders command general adjutant numbers assistant department order special hdqrs duty report headquarters brigadier proceed assigned officer relieved dept |
| 1 | 3.37 | general mississippi major la west miss jackson vicksburg memphis orleans smith mobile river corinth sherman louisiana grant bayou ala |
| 36 | 3.32 | river boats fort boat island point steamer navy landing flag board vessels city land gunboats transports fleet port gun |
| **Region: Mississippi River (14707 spans, 6.11%)** | | |
| 1 | 5.32 | general mississippi major la west miss jackson vicksburg memphis orleans smith mobile river corinth sherman louisiana grant bayou ala |
| 32 | 3.67 | orders command general adjutant numbers assistant department order special hdqrs duty report headquarters brigadier proceed assigned officer relieved dept |
| 16 | 3.21 | general commanding adjutant assistant brigadier major headquarters directs respectfully brevet desires acting wm november december wishes hdqrs february indorsement |
| 36 | 3.01 | river boats fort boat island point steamer navy landing flag board vessels city land gunboats transports fleet port gun |

Table 5.6: Top topics for different regions in WOTR.

| Topic | Prop% | Top words |
|---|---|---|
| **Region: Union Borderlands (2725 spans, 24.45%)** | | |
| 36 | 3.91 | general orders department assistant command adjutant order duty officer officers headquarters commanding special quartermaster numbers provost hdqrs report district |
| 4 | 3.76 | war government military citizens property law united authorities authority loyal persons country rebellion civil soldiers laws acts protection army |
| 34 | 3.68 | service number men regiments officers state office draft recruiting total call recruits district states enrollment military organization july years |
| 3 | 3.58 | negroes slaves labor negro slave free white employed soldiers number owners children work plantations persons families service laborers population |
| **Region: Confederate Interior (2594 spans, 23.28%)** | | |
| 11 | 4.12 | miles road march marched river bridge camp creek railroad day roads night moved train encamped crossed crossing left distance |
| 7 | 4.04 | general tennessee railroad nashville tenn memphis forrest chattanooga river atlanta major cumberland east north thomas hood ala kentucky sherman |
| 6 | 3.78 | enemy cavalry command colonel force back miles artillery loss prisoners captured advance morning river infantry killed ordered moved creek |
| 1 | 3.51 | enemy line left position brigade front rear road ordered fire advance forward skirmishers moved advanced back regiment formed colonel |
| **Region: Virginia (1743 spans, 15.64%)** | | |
| 1 | 5.75 | enemy line left position brigade front rear road ordered fire advance forward skirmishers moved advanced back regiment formed colonel |
| 33 | 5.74 | corps june division army house line general july petersburg left brigade point road james works station moved night va |
| 19 | 5.31 | general cavalry enemy richmond virginia va lee potomac road court house yesterday left side hill information point force fredericksburg |
| 30 | 4.10 | colonel lieutenant captain major john st william brigade light brigadier general company companies george james charles colored henry thomas |
| **Region: Atlantic Seaboard (1440 spans, 12.92%)** | | |
| 22 | 5.63 | states state people government union united power constitution south president slavery rights peace federal congress war southern confederacy country |
| 27 | 5.38 | battery fort guns island enemy fire batteries day night pounder gun inch fired shells morris sumter shots shell rifled |
| 17 | 4.21 | state letter make hope governor give subject matter present desire great regard people excellency opinion leave feel attention respectfully |
| 4 | 4.05 | war government military citizens property law united authorities authority loyal persons country rebellion civil soldiers laws acts protection army |
| **Region: Trans-Mississippi Confederacy (1132 spans, 10.16%)** | | |
| 8 | 4.44 | mississippi la orleans river vicksburg bayou west louisiana mobile gulf april miss texas port jackson march major banks hudson |
| 36 | 3.76 | general orders department assistant command adjutant order duty officer officers headquarters commanding special quartermaster numbers provost hdqrs report district |
| 37 | 3.71 | force general enemy troops army forces movement attack river operations lines point hold command country position move movements time |
| 17 | 3.50 | state letter make hope governor give subject matter present desire great regard people excellency opinion leave feel attention respectfully |
| **Region: Mississippi River (1111 spans, 9.97%)** | | |
| 8 | 7.05 | mississippi la orleans river vicksburg bayou west louisiana mobile gulf april miss texas port jackson march major banks hudson |
| 36 | 4.82 | general orders department assistant command adjutant order duty officer officers headquarters commanding special quartermaster numbers provost hdqrs report district |
| 20 | 3.58 | respectfully servant obedient commanding honor report colonel instant headquarters adjutant sir negroes assistant hdqrs general negro submit information asst |
| 16 | 3.52 | men force horses rebel rebels command cavalry camp country citizens captured mounted information place miles guerrillas scout small returned |

Table 5.7: Top topics for different regions in WOTR, African-American documents only.

### 5.5.2 Dynamic topic models

*Dynamic topic models* (Blei and Lafferty, 2006) are variants of topic models designed for observing topics that change over time or over a similarly autocorrelated dimension. In such a model, the topic proportions are allowed to vary from one time slice to the next. Specifically, when the topic proportions are transformed into the natural parameters of the multinomial distribution, these natural parameters evolve using Gaussian noise, controlled by a *top-chain variance* parameter.

I computed dynamic topic models both across both time, in half-year intervals, and across regions in Figure 5.20, starting from the Trans-Mississippi Confederacy in the southwest and moving east and northeast through the Mississippi River, Confederate Interior, Atlantic Seaboard, Virginia, and Union Borderlands regions.

As above, runs were done using predicted coordinates computed using *k*-d trees with a bucket size of 20. For computational reasons it was impossible to run on the entire set; thus, it was run only on the subset of documents that mention African-Americans, according to the procedure described above. Words were lowercased and canonicalized to remove extraneous punctuation, and stopwords were ignored, along with words occurring fewer than five times in the African-American subset. Runs were done using 40 topics, which seemed to work well in producing topics that were fairly specific but without too much redundancy.

#### Region topics

Table 5.8 and Table 5.9 show the results for 6 of the region-based topics. Some words of interest are highlighted in blue.

- Topic 7 is a political topic, presumably reflecting high-level planning of war efforts, with words like *president*, *secretary*, *congress*, *confederate*, *united* and *states*. The word *slaves* here is of particular interest, increasing steadily across the regions with a particular jump in the last two regions, Virginia and the Union Borderlands. These two regions are where the respective seats of government of the North and South were located, and the high position of the word *slaves* may reflect the importance of slavery—its preservation or end—as a motivating factor for keeping the war going, and a frequent topic of political discussion in both governments.

| Topic 7 | | | | | |
|---|---|---|---|---|---|
| trans-miss. | miss. | conf. int. | atlantic | virginia | borderlands |
| war | states | states | states | states | ***slaves*** |
| states | war | state | state | act | act |
| state | state | act | act | state | states |
| act | act | war | president | ***slaves*** | service |
| secretary | president | president | confederate | war | state |
| service | secretary | congress | congress | service | war |
| congress | congress | confederate | united | confederate | persons |
| united | united | united | war | president | united |
| president | confederate | secretary | service | united | confederate |
| governor | service | service | secretary | congress | president |
| confederate | governor | governor | ***slaves*** | secretary | secretary |
| law | law | ***slaves*** | persons | persons | ***slave*** |
| persons | ***slaves*** | law | law | governor | congress |
| ***slaves*** | persons | persons | governor | law | person |
| office | impressment | military | military | military | military |
| military | military | african | sec | person | law |
| officers | office | court | court | sec | sec |
| impressment | officers | sec | officers | provided | governor |
| authorized | provided | office | office | section | section |
| department | authorized | officers | person | office | provided |

| Topic 19 | | | | | |
|---|---|---|---|---|---|
| trans-miss. | miss. | conf. int. | atlantic | virginia | borderlands |
| white | white | white | white | white | white |
| colored | colored | colored | colored | colored | number |
| black | black | black | population | population | population |
| population | population | population | black | number | colored |
| number | number | number | number | years | years |
| whites | ***slaves*** | ***slaves*** | ***slaves*** | ***slaves*** | cases |
| ***slaves*** | whites | whites | years | black | ***slaves*** |
| negro | negro | negro | whites | cases | slave |
| negroes | negroes | years | negro | slave | negroes |
| years | years | negroes | slave | negroes | cent |
| free | free | slave | negroes | negro | black |
| slave | slave | free | free | whites | ***disease*** |
| race | cases | cases | cases | free | negro |
| cases | race | race | race | cent | ***deaths*** |
| blacks | blacks | report | cent | report | free |
| report | report | blacks | report | race | report |
| average | average | cent | ***deaths*** | ***deaths*** | ***hospital*** |
| cent | cent | average | average | ***disease*** | ***diseases*** |
| year | great | ***deaths*** | blacks | total | whites |
| great | year | great | ***disease*** | ***hospital*** | total |

| Topic 26 | | | | | |
|---|---|---|---|---|---|
| trans-miss. | miss. | conf. int. | atlantic | virginia | borderlands |
| ***cotton*** | orders | property | property | property | orders |
| government | order | orders | military | orders | department |
| department | government | order | orders | military | persons |
| orders | property | military | order | department | officers |
| order | ***cotton*** | government | department | order | military |
| military | department | department | government | persons | property |
| general | military | general | persons | officers | order |
| property | general | persons | general | general | general |
| trade | persons | officers | officers | government | government |
| treasury | officers | ***cotton*** | officer | officer | commanding |
| persons | trade | officer | proper | commanding | officer |
| officers | treasury | trade | commanding | proper | proper |
| ***plantations*** | ***plantations*** | treasury | ***cotton*** | numbers | numbers |
| officer | officer | proper | authority | lines | labor |
| army | freedmen | commanding | treasury | authority | lines |
| lines | army | army | trade | labor | authority |
| commanding | commanding | ***plantations*** | army | army | made |
| articles | lines | lines | lines | headquarters | army |
| supplies | proper | authority | articles | made | headquarters |
| contraband | articles | articles | numbers | articles | employed |

Table 5.8: Top dynamic region topics in WOTR, African-American documents, topics 7, 19 and 26.

Topic 29

| trans-miss. | miss. | conf. int. | atlantic | virginia | borderlands |
|---|---|---|---|---|---|
| states | states | states | states | states | states |
| nation | state | state | state | state | government |
| confederate | confederate | government | government | government | state |
| state | people | people | people | people | people |
| people | government | united | ***war*** | ***war*** | ***war*** |
| united | nation | confederate | united | law | law |
| government | united | ***war*** | power | united | country |
| country | country | constitution | law | power | ***military*** |
| art | ***war*** | power | constitution | country | power |
| nations | union | union | union | ***military*** | united |
| ***treaty*** | power | country | country | union | congress |
| laws | constitution | law | public | congress | property |
| ***cherokee*** | laws | convention | great | constitution | laws |
| day | convention | made | rights | public | public |
| ***war*** | law | rights | ***military*** | great | union |
| made | made | nation | confederate | property | south |
| law | peace | peace | congress | laws | great |
| union | day | public | made | south | slavery |
| power | rights | laws | south | made | made |
| peace | time | south | laws | rights | present |

Topic 30

| trans-miss. | miss. | conf. int. | atlantic | virginia | borderlands |
|---|---|---|---|---|---|
| country | country | county | county | county | county |
| ***cotton*** | men | men | citizens | men | men |
| men | ***cotton*** | country | men | country | country |
| texas | citizens | citizens | country | citizens | citizens |
| citizens | county | people | people | people | counties |
| people | people | counties | counties | counties | people |
| county | counties | ***cotton*** | horses | union | union |
| brownsville | texas | horses | ***cotton*** | horses | ***guerrillas*** |
| horses | horses | negroes | union | soldiers | property |
| counties | negroes | union | property | property | horses |
| place | place | property | soldiers | negroes | soldiers |
| tex | soldiers | soldiers | negroes | ***guerrillas*** | good |
| soldiers | property | place | good | good | loyal |
| ***bales*** | good | good | home | town | negroes |
| negroes | brownsville | texas | ***guerrillas*** | loyal | state |
| property | union | home | town | home | town |
| bands | state | state | families | state | home |
| state | ***bales*** | town | state | ***cotton*** | protection |
| good | home | ***guerrillas*** | place | place | place |
| rio | bands | families | loyal | families | families |

Topic 38

| trans-miss. | miss. | conf. int. | atlantic | virginia | borderlands |
|---|---|---|---|---|---|
| river | river | river | river | river | river |
| ***bayou*** | ***bayou*** | boats | boats | boats | boat |
| boats | boats | ***bayou*** | point | point | boats |
| port | port | point | boat | boat | point |
| transports | point | boat | ***bayou*** | landing | board |
| ***la*** | hudson | landing | landing | board | landing |
| point | landing | port | board | ***bayou*** | shore |
| boat | ***rouge*** | transports | transports | steamer | steamer |
| ***orleans*** | ***baton*** | board | steamer | shore | ***bayou*** |
| landing | boat | gunboats | gunboats | transports | transports |
| hudson | transports | water | port | gunboats | gunboats |
| grand | ***vicksburg*** | enemy | water | land | land |
| gunboats | ***la*** | steamer | shore | gun-boats | gun-boats |
| steamer | gunboats | ***vicksburg*** | land | water | bank |
| ***vicksburg*** | enemy | land | enemy | bank | water |
| city | lake | general | gun-boats | port | enemy |
| general | general | lake | bank | enemy | port |
| water | city | gun-boats | steamers | steamers | steamers |
| alexandria | board | city | lake | transport | transport |
| enemy | water | shore | general | general | side |

Table 5.9: Top dynamic region topics in WOTR, African-American documents, topics 29, 30 and 38.

In addition, the Union Borderlands overall was an area where slavery was of great political importance, since these were areas that were part of the Union but where slavery was still legal; this may be the reason for its particularly high position here. It is also possible that the high position in Virginia and the Union Borderlands reflects the incidence of runaway slaves in these border areas; however, if this was the case the word *contraband* might be expected to appear more often, since this was the term used by the Union to describe runaway slaves.

- Topic 19 talks primarily about African-Americans. The high position of the word *white* here presumably reflects the frequent occurrence of the collocations *white and black* (and occasionally *black and white*) in sentences such as "Cause all women and children, both white and black, who had not their homes within our lines ... to be excluded therefrom" and "Hereafter proper medical attention will be given to all employees of this army, white and black ....". Note here that the word *slaves* does not show the same sort of jumps as in topic 7 above, which is perhaps to be expected in a topic that is so concerned with slaves as such; this demonstrates the importance of using topic models to identify and separate topics, and the insufficiency of simple word frequency counts in examining patterns of discussion. (Note that there is a moderate increase across the regions in the usage of the singular *slave*; it is unclear what to attribute this to.)

  The other words I have highlighted are those that concern death and disease, which appear more in the eastern theaters of the South than in the west, with a big jump in the Union Borderlands. Since this topic specifically concerns African-Americans, this may reflect the much greater incidence of colored troops serving on the side of the North and the consequent disease and death that was so much the lot of soldiers of the day. Colored troops served more during the later stages of the war, and in this period the war effort was focused more in the East, as the entire Mississippi had been captured by the North within the first two years of the war, cutting off the West.

  Some examples of this are (in a June 30, 1863 letter from the Surgeon General's office in Washington, D.C. to Secretary of War Stanton) "official information has been received at this office relative to the combative liability of white and colored troops to diseases of malarious origin"; (in a report dated June 30, 1865) "In the casualties among the colored troops the most striking circumstance is the enormous proportion of deaths by disease"; and (in a memorandum dated August 31, 1863) "twenty-three prisoners (one white officer and twenty-two colored and negro privates) were put to death in cold blood". (Numerous letters and reports note the increased susceptibility of colored troops to disease as compared with white troops, and often attribute this to physical or moral defects of African-Americans, when in reality this probably reflects the racism of the white officers, who undoubtedly treated and provisioned colored troops worse than white ones.)

- Topics 26 and 30 indicate the importance of cotton in the Southern war effort. The nature of these topics is less clear than the others. Topic 26 may partly be concerned with provisioning the war effort, with words such as *property*, *trade*, *supplies*, *articles*, and *treasury*. Topic 30 may be concerned with the effect of the war on civilians and their property, with words such as *property*, *citizens*, *people*, *home*, *town* and *families*. In both cases words related to cotton-based agriculture (*cotton*, *bales* and *plantations*) occur primarily in the more southern areas of the South, especially farther to the west, while hardly in Virginia and not at all in the North. This appears to reflect the distribution of cotton growing, as shown in Figure 5.23 (which reflects cotton-growing regions as of 2007, but which is probably not significantly different from the situation in the 1860's except perhaps for the areas in West Texas and farther west, which may not have been as developed at the time).

  Also interesting is the distribution of the term *guerrillas* in topic 30, which is opposite to that of *cotton*. Guerrilla fighting was particularly intense in Missouri, where most of the more than 1,000 battles fought by Northern troops were with guerrillas (Erwin, 2012). This is reflected in the high position of the term *guerrillas* in the Union Borderlands, which includes Missouri. Guerrillas also appeared more generally throughout the South in relatively unpopulated areas with Unionist sentiment, and in many areas towards the end of the war where civil authority had broken down (Bohannon, 2014). Sherman had many problems with guerrillas in northern Georgia in the Confederate Interior region (Bohannon, 2014), and significant guerrilla warfare also occurred in northwest Virginia (Sutherland, 2012). Fighting in the later stages of the war

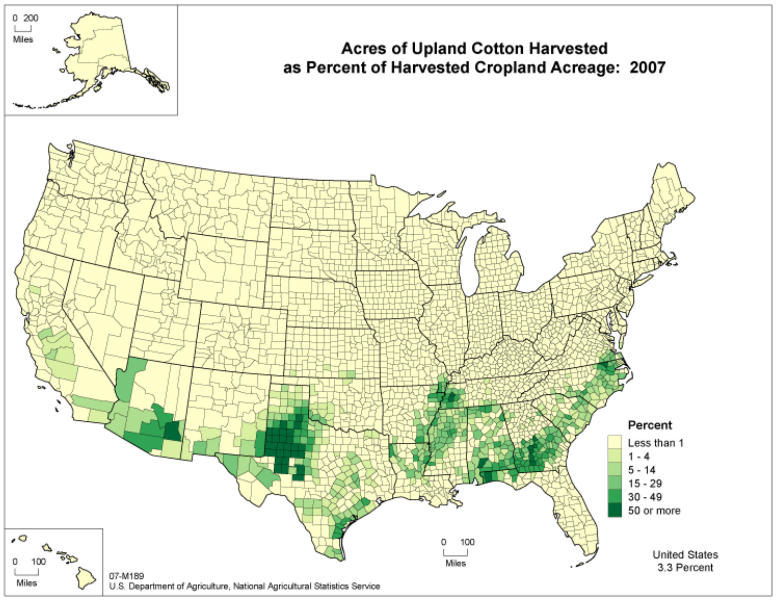

0 200
Miles
Acres of Upland Cotton Harvested
as Percent of Harvested Cropland Acreage: 2007
Percent
Less than 1
1 - 4
5 - 14
15 - 29
30 - 49
50 or more
0 100
Miles
0 100
Miles
United States
3.3 Percent
07-M189
U.S. Department of Agriculture, National Agricultural Statistics Service

Figure 5.23: Primary cotton-growing regions of the United States, as of 2007, from the U.S. Department of Agriculture.

was primarily in the East, which may be the reason why the term does not occur in the Western regions (Mississippi River and Trans-Mississippi).

- Topic 29 is, like topic 7, a political topic, but of a different nature, and less focused specifically on the war. Here, unlike in topic 7, the terms *war* and *military* are primarily limited to the more Eastern regions and to the Union Borderlands, which saw the bulk of the fighting. Particularly interesting is the occurrence of the terms *nations* (plural, as in phrasings such as “Indian nations” and “Choctaw and Chickasaw nations”), *treaty* and *cherokee* in the Trans-Mississippi, none of which occur in the top 20 terms of the topic among any of the other regions. This reflects the fact that dealings with Native Americans occurred significantly more in this region than elsewhere (indeed, Oklahoma was known at the time as the Indian Territory).

- Topic 38 clearly concerns water-based transport, with terms such as *river*, *port*, *bayou*, *boats*, *landing*, *gunboats*, *steamer*, etc. Not surprisingly, various cities where important naval battles took place are represented, namely Vicksburg, Baton Rouge and New Orleans. This topic also shows, as expected, that geographic-specific terms are generally represented in the areas in which they are located and to some extent in neighboring areas. The term *vicksburg*, for example, occurs in the Mississippi River region (the city of Vicksburg sits on the Mississippi), but also occurs at a lower position in the two neighboring regions. This may partly reflect the fact that the Mississippi River region is narrow, and the neighboring regions were involved in sending troops to fight at the Battle of Vicksburg. However, it may also simply reflect imprecise classification of documents, either due to limitations of the grid-based approach or misclassification for other reasons (e.g. mentioning of place names in regions other than that of the primary geographic topic of the document). In addition, it may partly be due to the nature of the dynamic topic model algorithm, which has a (controllable) parameter that determines how "jumpy" the topics are allowed to be. (I set this parameter so as to allow moderate but not extreme jumpiness, to ensure coherency of topics.) These latter two reaons may be part of the explanation for the distribution of the term *bayou*, which occurs at the highest position in the topics for the Mississippi and Trans-Mississippi areas, where bayous are normally found (primarily in Louisiana, southern Arkansas and eastern Texas), but also occurs fairly high in the topics for all the other regions. An example of why misclassification of this term may occur is a letter dated April 21, 1863 written from Opelousas, Louisiana, which mentions the term *bayou* 10 times but also speaks repeatedly of various militias stemming from faraway regions, such as the 4th Wisconsin, the 8th New Hampshire, the 21st Indiana, and the 173rd New York. Another such example, with six mentions of *bayou* (capitalized or not), is dated April 27, 1863 and identified in the header as being written "In camp on Bayou Boeuff, beyond Washington, La.", where the place name *Washington* will almost certainly bias the predicted location towards Washington, DC.

### Time-based topics

It's also possible to compute dynamic topics across time, as dynamic topic models were originally designed for. As described above, dynamic topic models were run on half-year intervals, using the same parameters as were used for region-based topic models, including using *k*-d tree predictions and producing 40 topics. Because of the highly uneven distribution of documents across the various half-year intervals, intervals with fewer than 50 documents were ignored. This had the effect of excluding intervals before 1861 and after 1865. (This was unnecessary in the case of regions, all of which had well over 50 documents.)

Because this dissertation is primarily concerned with geography, I focus less on time-based than region-based dynamic topic models. For time-based models, I focus primarily on the distribution of the term *colored*, showing the 7 out of 40 topics where this word occurred significantly, in Table 5.10, Table 5.11, and Table 5.12. All of the topics show a clear trend in that the word *colored* occurs more in the later stages of the war than the earlier stages. This appears to reflect the fact that mustering of colored troops primarily occurred later in the war, especially after the Emancipation Proclamation took effect in January of 1863. This is visible most clearly in topic 13, where *colored* appears around position 19 in the latter half of 1862, and then quickly jumps up starting in the first half of 1863.

Note that in the topics other than topic 13, the term *colored* tends to appear later, often not till 1864 or 1865. This may reflect that fact that topic 13 in particular, with terms like *department*, *headquarters*, *brigade*, *division*, *corps* and especially *adjutant-general*, may reflect primarily headings and salutations, such as "AR. C. BAILEY, / Captain, Commanding Eighth Regiment U. S. Colored Troops. / Lieutenant E. L. MOORE, / Acting Assistant Adjutant-General." and "GEORGE L. STEARNS, / Major and Assistant Adjutant-General, U. S. Volunteers, / Commissioned for Organization U. S. Colored Troops." As the latter example shows, some of these salutations refer to organizations for mustering colored troops, which would have been active before the actual troops themselves had been mustered. Topics 0, 16, 17 and 18 reflect a later time when such troops had in fact been mustered; the frequent reference to state names is indicative of collocations such as "Kansas Troops, 1st Regiment (Colored)", referring to state-based colored militias (note the word "militia" occurring prominently near the top of topic 18).

The topic with the latest instances of the term *colored* is topic 6; with terms like *expedition*, *captured*, *party*, *returned*, and *report*, this topic appears to indicate military expeditions. The late appearance of the term *colored* here appears to reflect the lag between when the colored troops were mustered and when they saw significant military action.

Finally, topic 27 shows a pattern different from all the others. It is similar to the others in that the word *colored* doesn't appear until 1863, but it peaks in position later in 1863 and then drops steadily. This topic, with prominent mention of *prisoners* and *war* along with *exchange* and *captured*, appears to reference prisoners of war and prisoner exchanges between the North and South. (Note also the occurrence of *united*, *states*, *confederate*, *washington*, *richmond* and *rebel*, which further suggest a discussion between the two sides.) These prisoner exchanges notably broke down in July 1863 due to the refusal of the Confederates to return colored soldiers captured as prisoners of war (or indeed, to treat such soldiers as prisoners of war at all) (Cloyd, 2010). Other terms in the topic are also consistent with the theme of prisoner exchanges. The terms *secretary* and *stanton* refer to U.S. Secretary of War Edwin Stanton. The term *butler* refers to Union general Benjamin Butler, who was well-known for initiating in 1861 the policy of treating runaway slaves as "contraband", meaning that they did not need to be returned to their owners, as the Fugitive Slave Act called for. He also served as the Union Commissioner of Exchange in charge of prisoner exchanges (National Park Service, 2015) starting in 1864.

Topic 0

| 01/01/1861 | 07/03/1861 | 01/01/1862 | 07/03/1862 | 01/01/1863 | 07/03/1863 | 01/01/1864 | 07/02/1864 | 01/01/1865 | 07/03/1865 |
|---|---|---|---|---|---|---|---|---|---|
| captain | captain | captain | captain | colonel | colonel | colonel | colonel | colonel | colonel |
| company | company | company | colonel | captain | lieutenant | lieutenant | lieutenant | lieutenant | lieutenant |
| colonel | colonel | colonel | company | lieutenant | captain | captain | captain | captain | captain |
| lieutenant | lieutenant | lieutenant | lieutenant | company | major | major | major | major | major |
| john | john | john | john | john | company | battery | battery | company | general |
| william | william | william | major | major | john | john | john | general | company |
| major | major | major | william | william | battery | william | company | artillery | artillery |
| companies | companies | companies | companies | battery | william | brigade | brigade | light | light |
| james | james | james | battery | companies | 1st | 1st | william | battery | brigadier |
| george | george | battery | james | 1st | brigade | general | artillery | john | ***colored*** |
| artillery | battery | george | 1st | artillery | artillery | company | light | brigade | battery |
| battery | artillery | artillery | artillery | james | general | light | 1st | ***colored*** | john |
| 1st | 1st | 1st | brigade | brigade | companies | artillery | general | william | brigade |
| charles | charles | general | general | general | james | brigadier | ***colored*** | 1st | william |
| light | general | brigade | george | light | light | york | york | brigadier | 1st |
| henry | light | charles | light | george | brigadier | james | brigadier | york | york |
| general | brigade | light | charles | 2nd | george | george | james | companies | companies |
| brigade | henry | henry | 2nd | brigadier | 2nd | companies | companies | james | charles |
| thomas | thomas | thomas | brigadier | charles | york | 2nd | george | charles | james |
| 2nd | 2nd | 2nd | thomas | henry | charles | charles | charles | george | george |

Topic 6

| 01/01/1861 | 07/03/1861 | 01/01/1862 | 07/03/1862 | 01/01/1863 | 07/03/1863 | 01/01/1864 | 07/02/1864 | 01/01/1865 | 07/03/1865 |
|---|---|---|---|---|---|---|---|---|---|
| men | men | men | men | men | men | men | men | men | men |
| captain | captain | captain | captain | captain | captain | captain | captain | captain | captain |
| river | river | river | river | lieutenant | river | river | river | lieutenant | lieutenant |
| lieutenant | lieutenant | lieutenant | lieutenant | river | lieutenant | lieutenant | lieutenant | river | river |
| night | boat | party | found | miles | miles | miles | miles | report | report |
| boat | night | found | miles | found | report | report | report | miles | miles |
| mr | mr | night | party | report | found | found | cavalry | bayou | bayou |
| party | party | miles | report | party | party | party | horses | cavalry | found |
| found | found | boat | horses | boat | morning | horses | found | found | cavalry |
| horses | house | horses | night | morning | horses | cavalry | morning | horses | horses |
| house | horses | mr | morning | horses | boat | morning | party | morning | boat |
| morning | miles | report | boat | night | cavalry | captured | command | boat | morning |
| miles | morning | house | expedition | expedition | captured | boat | captured | command | ***colored*** |
| place | report | morning | place | place | command | command | boat | ***colored*** | command |
| report | place | place | house | command | place | place | 1864 | party | party |
| o'clock | o'clock | proceeded | mr | cavalry | night | returned | place | captured | captured |
| left | proceeded | o'clock | proceeded | left | expedition | 1864 | bayou | place | place |
| proceeded | left | expedition | command | returned | returned | night | commanding | commanding | commanding |
| expedition | returned | left | left | captured | company | commanding | returned | returned | expedition |
| returned | expedition | returned | returned | company | camp | camp | rebel | expedition | returned |

Topic 13

| 01/01/1861 | 07/03/1861 | 01/01/1862 | 07/03/1862 | 01/01/1863 | 07/03/1863 | 01/01/1864 | 07/02/1864 | 01/01/1865 | 07/03/1865 |
|---|---|---|---|---|---|---|---|---|---|
| general | general | general | general | general | general | ***colored*** | ***colored*** | ***colored*** | ***colored*** |
| command | command | command | orders | orders | ***colored*** | general | general | general | general |
| orders | orders | orders | command | regiments | regiments | troops | troops | orders | 1865 |
| regiment | regiment | regiment | regiments | regiment | orders | volunteers | assistant | assistant | orders |
| army | army | army | regiment | command | regiment | command | commanding | command | department |
| regiments | regiments | regiments | army | ***colored*** | troops | regiments | command | 1865 | infantry |
| department | headquarters | commanding | commanding | department | command | orders | orders | infantry | commanding |
| volunteers | commanding | headquarters | adjutant-general | adjutant-general | volunteers | commanding | adjutant-general | commanding | assistant |
| headquarters | volunteers | division | department | volunteers | department | regiment | infantry | adjutant-general | command |
| commanding | division | adjutant-general | headquarters | troops | adjutant-general | assistant | 1864 | department | adjutant-general |
| adjutant-general | brigade | brigade | volunteers | commanding | commanding | adjutant-general | headquarters | headquarters | major |
| division | department | department | division | army | corps | department | regiments | troops | volunteers |
| brigade | adjutant-general | volunteers | assistant | corps | assistant | corps | department | major | troops |
| officers | officers | officers | officers | headquarters | colonel | colonel | regiment | volunteers | duty |
| colonel | assistant | assistant | brigade | assistant | headquarters | 1864 | volunteers | division | headquarters |
| assistant | colonel | colonel | troops | officers | army | headquarters | colonel | duty | division |
| troops | troops | troops | corps | colonel | officers | infantry | report | report | report |
| report | order | order | colonel | division | duty | report | division | numbers | numbers |
| order | report | corps | ***colored*** | brigade | report | army | corps | colonel | regiments |
| officer | officer | report | order | order | major | duty | major | regiments | brigadier |

Table 5.10: Top terms across time in WOTR, African-American documents only, dynamic topics 0, 6 and 13.

| Topic 16 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 01/01/1861 | 07/03/1861 | 01/01/1862 | 07/03/1862 | 01/01/1863 | 07/03/1863 | 01/01/1864 | 07/02/1864 | 01/01/1865 | 07/03/1865 |
| regiment | regiment | regiment | regiment | troops | troops | troops | troops | troops | troops |
| troops | troops | troops | troops | regiment | regiment | regiment | regiment | regiment | regiment |
| infantry | infantry | infantry | infantry | infantry | infantry | infantry | infantry | infantry | infantry |
| cavalry | cavalry | cavalry | cavalry | cavalry | cavalry | cavalry | cavalry | cavalry | cavalry |
| confederate | confederate | confederate | confederate | artillery | artillery | artillery | union | union | union |
| artillery | artillery | artillery | artillery | confederate | confederate | confederate | artillery | artillery | artillery |
| alabama | alabama | alabama | alabama | union | union | union | confederate | confederate | confederate |
| union | union | union | union | battery | battery | john | william | william | william |
| battery | battery | battery | battery | alabama | 1st | william | john | 1st | battery |
| battalion | john | john | john | 1st | john | battery | 1st | battery | 1st |
| john | battalion | 1st | 1st | john | william | 1st | battery | john | john |
| mississippi | 1st | battalion | mississippi | william | james | james | james | 2nd | battalion |
| 1st | mississippi | mississippi | william | mississippi | tennessee | 2nd | 2nd | james | 2nd |
| william | william | william | battalion | tennessee | mississippi | tennessee | george | ***colored*** | ***colored*** |
| tennessee | tennessee | tennessee | tennessee | james | 2nd | george | thomas | battalion | james |
| james | james | james | james | battalion | illinois | mississippi | ***colored*** | george | george |
| 2nd | 2nd | 2nd | 2nd | illinois | alabama | illinois | battalion | thomas | thomas |
| illinois | illinois | illinois | illinois | 2nd | battalion | battalion | charles | charles | charles |
| louisiana | louisiana | louisiana | louisiana | louisiana | george | alabama | illinois | 4th | alabama |
| thomas | thomas | thomas | thomas | george | 4th | thomas | 4th | alabama | 4th |

| Topic 17 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 01/01/1861 | 07/03/1861 | 01/01/1862 | 07/03/1862 | 01/01/1863 | 07/03/1863 | 01/01/1864 | 07/02/1864 | 01/01/1865 | 07/03/1865 |
| york | york | york | york | york | york | york | york | york | york |
| massachusetts | massachusetts | massachusetts | massachusetts | massachusetts | massachusetts | infantry | brigade | brigade | brigade |
| pennsylvania | pennsylvania | pennsylvania | pennsylvania | pennsylvania | pennsylvania | pennsylvania | pennsylvania | total | total |
| brigade | brigade | brigade | artillery | artillery | infantry | brigade | infantry | infantry | artillery |
| artillery | artillery | artillery | brigade | island | artillery | massachusetts | massachusetts | artillery | infantry |
| connecticut | connecticut | connecticut | connecticut | brigade | island | artillery | artillery | pennsylvania | pennsylvania |
| battery | island | island | island | infantry | brigade | total | total | massachusetts | massachusetts |
| island | battery | battery | battery | connecticut | connecticut | battery | battery | battery | connecticut |
| maine | maine | maine | infantry | battery | battery | connecticut | connecticut | men | battery |
| hampshire | hampshire | hampshire | hampshire | hampshire | hampshire | island | men | connecticut | men |
| men | jersey | rhode | maine | maine | men | men | island | cavalry | cavalry |
| jersey | men | men | rhode | rhode | maine | maine | maine | ***colored*** | ***colored*** |
| infantry | infantry | infantry | men | men | total | hampshire | cavalry | division | light |
| rhode | rhode | jersey | jersey | total | rhode | 1st | hampshire | island | island |
| 1st | 1st | 1st | 1st | 1st | 1st | heavy | ***colored*** | light | division |
| total | total | total | total | jersey | heavy | rhode | 1st | maine | maine |
| volunteers | volunteers | volunteers | volunteers | volunteers | volunteers | cavalry | light | heavy | heavy |
| cavalry | cavalry | cavalry | cavalry | cavalry | jersey | jersey | division | 1st | hampshire |
| officers | officers | officers | officers | heavy | cavalry | ***colored*** | heavy | hampshire | 1st |
| light | light | light | light | officers | ***colored*** | general | rhode | rhode | rhode |

| Topic 18 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 01/01/1861 | 07/03/1861 | 01/01/1862 | 07/03/1862 | 01/01/1863 | 07/03/1863 | 01/01/1864 | 07/02/1864 | 01/01/1865 | 07/03/1865 |
| militia | militia | militia | militia | militia | cavalry | cavalry | cavalry | cavalry | cavalry |
| texas | texas | texas | texas | missouri | missouri | missouri | missouri | missouri | missouri |
| missouri | missouri | missouri | missouri | cavalry | militia | militia | militia | militia | militia |
| troops | cavalry | cavalry | cavalry | texas | union | union | union | union | union |
| cavalry | troops | troops | troops | state | troops | troops | troops | troops | troops |
| state | state | state | state | troops | texas | regiment | regiment | 1st | 1st |
| kansas | kansas | kansas | kansas | union | state | state | texas | texas | texas |
| regiment | regiment | regiment | union | kansas | regiment | texas | kansas | regiment | 2nd |
| union | union | union | regiment | regiment | kansas | kansas | 1st | kansas | kansas |
| arkansas | iowa | iowa | iowa | iowa | 1st | 1st | state | 2nd | regiment |
| iowa | arkansas | arkansas | arkansas | arkansas | iowa | 2nd | 2nd | state | iowa |
| mexico | mexico | 1st | 1st | 1st | arkansas | arkansas | enrolled | iowa | state |
| 1st | 1st | 2nd | 2nd | 2nd | 2nd | iowa | arkansas | arkansas | arkansas |
| indian | 2nd | indian | enrolled | enrolled | enrolled | enrolled | iowa | enrolled | enrolled |
| 2nd | indian | mexico | indian | colorado | colorado | john | john | wisconsin | wisconsin |
| colorado | colorado | enrolled | colorado | indian | 3rd | 3rd | colorado | ***colored*** | ***colored*** |
| california | california | colorado | mexico | 3rd | john | colorado | 3rd | 3rd | colorado |
| enrolled | enrolled | california | 3rd | john | indian | ***colored*** | ***colored*** | colorado | 3rd |
| 3rd | 3rd | 3rd | john | wisconsin | wisconsin | wisconsin | illinois | john | john |
| john | john | john | california | california | provisional | provisional | wisconsin | illinois | illinois |

Table 5.11: Top terms across time in WOTR, African-American documents only, dynamic topics 16, 17 and 18.

| Topic 27 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 01/01/1861 | 07/03/1861 | 01/01/1862 | 07/03/1862 | 01/01/1863 | 07/03/1863 | 01/01/1864 | 07/02/1864 | 01/01/1865 | 07/03/1865 |
| war | war | war | war | war | war | prisoners | war | war | war |
| secretary | secretary | secretary | prisoners | prisoners | prisoners | war | prisoners | prisoners | prisoners |
| prisoners | prisoners | prisoners | secretary | officers | general | general | secretary | secretary | secretary |
| honorable | officers | officers | officers | secretary | officers | secretary | general | general | general |
| officers | honorable | honorable | general | captured | exchange | exchange | exchange | confederate | confederate |
| department | department | department | captured | general | secretary | officers | confederate | exchange | exchange |
| letter | letter | stanton | states | exchange | captured | confederate | officers | officers | officers |
| president | states | states | exchange | states | states | captured | states | states | states |
| washington | washington | general | department | stanton | stanton | states | soldiers | stanton | honorable |
| states | president | letter | stanton | honorable | ***colored*** | stanton | captured | honorable | stanton |
| captured | captured | captured | honorable | confederate | rebel | ***colored*** | honorable | captured | captured |
| general | general | exchange | confederate | department | confederate | soldiers | stanton | authorities | president |
| stanton | stanton | president | letter | rebel | honorable | honorable | ***colored*** | soldiers | authorities |
| confederate | exchange | washington | president | government | soldiers | made | authorities | president | soldiers |
| butler | confederate | confederate | government | authorities | authorities | rebel | made | richmond | richmond |
| exchange | butler | butler | rebel | letter | government | authorities | richmond | ***colored*** | rebel |
| united | government | government | butler | soldiers | mr | mr | rebel | mr | received |
| government | united | united | washington | ***colored*** | made | government | mr | rebel | ***colored*** |
| sir | sir | richmond | authorities | made | department | letter | president | received | mr |
| received | received | authorities | united | army | troops | richmond | government | made | washington |

Table 5.12: Top terms across time in WOTR, African-American documents only, dynamic topic 27.

# Chapter 6

# Document geolocation and toponym resolution

## 6.1 Introduction

This chapter discusses work that ties together document geolocation with toponym resolution, building on Speriosu (2013) and Speriosu and Baldridge (2013). This work stems from the observation that toponyms in a stretch of text are strong indicators of the overall location of that text. Document geolocation will naturally pick up on these toponyms by assigning high weight to the words comprising them. However, toponym resolution can still be of assistance, because

1. toponyms are potentially ambiguous, and document geolocation can benefit from working with resolved, rather than unresolved, toponyms;

2. some toponym resolution methods, such as SPIDER (§6.2), can do joint inference over the toponyms in an individual document or an entire corpus, taking advantage of dependence relations among different toponyms in a document;

3. toponym resolvers can make use of distinct knowledge sources from document geolocators, for example a gazetteer of known toponyms and their possible candidates.

Toponym resolution can be used to inform document geolocation in various ways. For example, a document geolocator could be constrained to making predictions that are near (within some threshold of) one of the toponyms in the text. Alternatively, toponyms in a text can be used as features, for example in a reranker (§7.2). Ideally, however, we would do joint inference over toponyms and document geolocations, allowing each to inform the other. In this chapter, I describe a method for doing such joint inference, using a variant of co-training (§6.3).

## 6.2 Toponym resolution techniques

Speriosu (2013) developed a number of methods for toponym resolution, some of which applied document geolocation as one component. It is important to note that these methods are *unsupervised* from the perspective of toponyms, in that they do not rely on an existing corpus marked up with disambiguated toponyms. (WISTR is a partial exception, in that it synthesizes toponym annotations from a document-geolocation corpus. WISTR$^*$ in §6.2.2 is even more of an exception.)

Some of these methods rely on outside knowledge:

- A corpus, such as Wikipedia, annotated with document-level geocoordinates.
- A document geolocator trained from such a corpus.
- A gazetteer listing known toponyms and possible candidates for their location.

The following are the methods relevant to this dissertation:

- TRIPDL directly uses a document geolocator to produce a probability distribution over a cell grid covering the Earth. Each of the possible candidates for a given toponym in the document is then assigned a probability in line with the document-geolocation probability of the cell containing the candidate, and the highest-probability candidate chosen for the toponym.
- WISTR is a stronger method that uses document geolocations of Wikipedia in an indirect fashion. A named-entity recognizer is run on a Wikipedia page, and the toponyms in the page containing candidates within 10km of the document's location are considered to resolve

to those candidates. The textual context around those toponyms is used as features to train classifiers to disambiguate those toponyms to their resolved candidates.

For example, the Wikipedia page on *Widgery Wharf* in Portland, Maine contains various mentions of the toponym *Portland*, and the gazetteer entry for Portland contains the candidate *Portland, Maine* whose location (presumably the city center) is within 10km of the location of Widgery Wharf. Thus, the text surrounding each mention of Portland in this article serves as classifier features to disambiguate a mention of Portland in some other article to Portland, Maine. Combined with appropriate features to identify other Portlands (for example, mentions of Portland in the article on the *Portland Youth Philharmonic* in Portland, Oregon), a strong classifier can be created.

- TRAWL interpolates between WISTR and TRIPDL, and weights the result by a factor that biases in favor of higher-level administrative entities when e.g. disambiguating between a city and a country of the same name.

- SPIDER is a weighted-minimum-distance resolver than seeks to implement the heuristics of *spatial minimality* (different toponyms in a text tend to be near each other) and *one sense per discourse* (multiple instances of a toponym in a text tend to refer to the same location). At its core is a basic minimum-distance resolver, which resolves each toponym to the candidate that is, on average, closest to all other toponyms. (More specifically, for each toponym, it chooses the candidate that minimizes the sum of the distances to the closest candidate of each other toponym.) This has the effect of clumping all toponyms in a document together.

  SPIDER builds on top of this basic resolver by attaching a weight to each candidate of a toponym, reflecting its prominence in the corpus. The minimum-distance algorithm is then modified so that all distances computed are divided by the weights of the candidates involved (since smaller distances are better). Furthermore, multiple iterations are run, and at the end of each iteration, the weights are recomputed, reflecting the proportion of times a given candidate has been resolved across the entire corpus.

  Unlike WISTR, TRIPDL and TRAWL, SPIDER does joint inference of toponyms both

| Geolocator⟶ | Naive Bayes uniform, 1° | | | Hierarchical *k*-d | | |
|---|---|---|---|---|---|---|
| Toponym resolver ↓ | Mean | Median | Precision | Mean | Median | Precision |
| RANDOM | 2397 | 933 | 23.4% | 2397 | 933 | 23.4% |
| TRIPDL | 1014 | 26 | 57.2% | 1235 | 38 | 51.7% |
| TRAWL | 1825 | 419 | 42.3% | 827 | 15 | 70.5% |
| WISTR | 665 | **0** | 74.5% | 665 | **0** | 74.5% |
| SPIDER | 675 | **0** | 74.7% | 675 | **0** | 74.7% |
| TRAWL+SPIDER | 673 | **0** | 74.8% | **243** | **0** | 82.0% |
| WISTR+SPIDER | **422** | **0** | **82.5%** | 422 | **0** | **82.5%** |

Table 6.1: Dev set performance on CWAR using various toponym resolution methods. Underlined values are those that have changed from left to right (the others remain the same because their method doesn't use a geolocator).

across an individual document and the entire corpus. (TRIPDL takes advantage of an entire document's context through the use of a document geolocator, but still resolves each toponym independently.)

- WISTR+SPIDER and TRAWL+SPIDER use WISTR and TRAWL, respectively, to initialize the weights of SPIDER. The underlying idea is that the weights in SPIDER can be viewed as set of prior distributions, one per toponym, over the candidates of that toponym. Both WISTR and TRAWL output probability distributions over the candidates of each toponym and use outside knowledge sources to do so, and thus can be used to more intelligently initialize SPIDER's weights than simply initializing them uniformly, as SPIDER does by itself. These combined methods are generally stronger than either of the component methods standing alone.

### 6.2.1 Baseline toponym resolution results

As described in §2.4, I redid the CWAR dataset to include coordinates for all of the 56,000+ distinct toponym types originally annotated in the corpus, as opposed to the only 2,000 or so types that were assigned coordinates in Speriosu (2013)'s work. I reran the methods described above, producing the updated results shown in Table 6.1.

In addition, I modified the code that implements these resolvers to allow for the use of the new document geolocation techniques described in this dissertation.[1] This allowed for new variants

[1] Speriosu's original system used the geolocation methods of Wing and Baldridge (2011), in particular KL Divergence

of TRIPDL, TRAWL and TRAWL+SPIDER, which were run with an underlying hierarchical *k*-d tree classifier geolocator trained on EnWiki13 using the optimal settings found in §4.2.3. These results are shown on the right half of Table 6.1. Using a hierarchical classifier does not help with TRIPDL (which is one of the weaker methods in any case), but definitely does with TRAWL and TRAWL+SPIDER, making the latter the strongest method for mean, and very nearly as strong for precision as WISTR+SPIDER. This suggests that a better geolocator can improve the performance of a geolocation-based toponym resolver, a result that is perhaps expected but nonetheless pleasing.

### 6.2.2 New method WISTR$^*$ (variant of WISTR)

WISTR, as described above, identifies toponyms in Wikipedia using a named entity recognizer (NER) and disambiguates them by looking for candidates that are very close (10km or closer) to the document's geolocation. This procedure would be unnecessary if Wikipedia were directly marked up with toponyms and their resolutions, and in fact we can synthesize exactly such toponyms by making use of the hyperlinks between Wikipedia articles. The idea is that

1. we can identify any stretch of text that is linked to a geolocated article and is also found in the gazetteer as a toponym;

2. we can resolve the toponym by finding the candidate in the gazetteer that is closest to the linked article's geocoordinate, provided the distance does not exceed a threshold (I use 100km, or 500km for candidates that are identified in the gazetteer as states or higher-level administrative entities due to potential disagreements between Wikipedia and the gazetteer in identifying the "representative point" of such a region);

3. we can identify further stretches of the same text in the article[2] as toponyms, with the same resolution (this is necessary because typically only the first mention of a given item in an article is linked).

By doing this procedure, we can identify toponyms both more precisely (since we eliminate NER errors) and in greater number (since we no longer rely on toponyms being close to the article's

and Naive Bayes.

[2] Or more precisely, until we find another occurrence of the same text with an attached link.

| Corpus | Source |
|---|---|
| CWarPortal | The Civil War Portal subsection of EnWiki13 |
| TopoWiki13 | All of EnWiki13 |

Table 6.2: New toponym resolution corpora for use with WISTR$^*$, derived from part or all of EnWiki13 using a new and better method to identify and resolve toponyms in Wikipedia.

| Method | Corpus | Mean (km) | Precision (%) |
|---|---|---|---|
| WISTR | EnWiki13 | 850 | 69.5 |
| WISTR+SPIDER | EnWiki13 | 107 | 89.5 |
| WISTR$^*$ | TopoWiki13 | 713 | 80.8 |
| WISTR$^*$+SPIDER | TopoWiki13 | 85 | 91.3 |
| WISTR$^*$ | CWarPortal | **183** | **86.8** |
| WISTR$^*$+SPIDER | CWarPortal | **61** | **92.0** |
| WISTR/WISTR$^*$ | EnWiki13+CWarPortal | 463 | 83.1 |
| WISTR/WISTR$^*$+SPIDER | EnWiki13+CWarPortal | 87 | 91.1 |

Table 6.3: Results for WISTR and WISTR$^*$ on CWar.

own geocoordinate). Finally, we can make use of articles that are not themselves geolocated, which comprise more than 80% of the total, and nearly 95% of those in the Civil War Portal subsection (§2.4).

Using this new procedure, I create two new toponym resolution corpora from EnWiki13 (see Table 6.2).

I then create a variant of WISTR, which I term WISTR$^*$, that directly relies on the toponyms in these corpora rather than finding toponyms in the former, more roundabout fashion. Table 6.3 shows the results of running on the CWar corpus. In addition, I investigate results using a combination of WISTR$^*$ features from CWarPortal, and WISTR features extracted from all of EnWiki13.

Interestingly, WISTR$^*$ results are noticeably better using only CWarPortal than TopoWiki13 (the entire Wikipedia). This demonstrates the importance of in-domain data.

### 6.2.3 Variants of SPIDER

I modified SPIDER to incorporate a document-level geotag when it is available. Such a geotag is typically available in the toponym-resolution portion of co-training (§6.3), as the toponym resolver is fed documents that have already been annotated by the document geolocator. There are two ways to do this:

**Weighted** This method sets the initial weights of each toponym to be inversely related to the distance from the document-level geotag.

**AddTopo** This method modifies SPIDER to add an additional toponym corresponding to the document-level geotag, effectively containing only one possible candidate, which resolves to the location of the document-level geotag. This biases SPIDER in favor of resolving other toponyms nearby, in order to satisfy the spatial minimality component of the algorithm.

AddTopo can be combined with any of the WISTR variants, but Weighted cannot, because its settings for the initial weights would conflict with the WISTR settings.

## 6.3 Co-training

### 6.3.1 Introduction

Co-training is a semi-supervised strategy introduced by Blum and Mitchell (1998) that allows for bootstrapping a classifier or other machine learning tool using only a small amount of labeled data and a large amount of unlabeled data. It depends on the existence of two different views of the data, described by two different feature sets that can be used to train distinct classifiers whose errors, in the ideal case, are uncorrelated. Co-training proceeds iteratively in an alternating fashion. Each classifier in turn is trained, beginning with the initial labeled data, and its most confident predictions on the unlabeled data are added to the set of training data and used to train the other classifier.

The initial task considered by Blum and Mitchell involved classifying web pages from computer science departments into one of four types, using the text of the pages themselves and the text of the links to the pages as the two different views. Co-training has since been applied to multiple domains, such as statistical machine translation (Callison-Burch, 2002), parsing (Sarkar, 2001), computer vision (Lu et al., 2011), and semantic role labeling (He and Gildea, 2006).

I propose here a variant of co-training that involves a combination of document geolocation and toponym resolution. This is a natural fit because these two types of geographic annotation complement each other: One provides an individual view of the distinct geographic place names in a text, and the other provides a holistic view of the overall geographic scope of the text.

The algorithm I describe here differs slightly from standard co-training, which uses the availability of two views onto a single corpus. I instead use two corpora, one of which has document geolocations while the other has toponym resolutions. One of the corpora is derived from the other, allowing information to flow between the two.

### 6.3.2 Basic algorithm

When co-training, I proceed as follows. The basic algorithm described here and below follows Abney (2007).

1. I begin with a corpus $U$ (such as CWAR) that is unlabeled but has the toponyms and their possible candidates identified. (Alternatively, such toponyms and candidates could be identified with the help of a gazetteer and named entity recognizer.) I also begin with a pre-trained document geolocator (e.g. one trained on ENWIKI13). The goal is to produce: (1) toponym annotations on the corpus; (2) a toponym resolver; (3) an improved document geolocator taking advantage of information contained in the corpus.

2. I create two initially empty corpora, $L$ (meant to contain documents labeled with both toponym-level and document-level annotations) and $L_w$ (meant to contain similarly labeled *document windows*, i.e. small segments of text surrounding a toponym in a given document, whose document-level annotation is derived from the toponym). $L_w$ is derived from the toponym annotations in $L$ and is used to train a document geolocator, while $L$ itself is used to train a toponym resolver such as WISTR$^*$.

3. The document geolocator makes predictions on $U$. Those documents whose location is predicted with relatively high probability are passed to a toponym resolution mechanism.

4. This mechanism accepts documents for which one of the toponyms in the document has a candidate that is close to the resolved document location. (This serves as one means to pass information from the document geolocator to the toponym resolver.) This produces an “accepted”

corpus $A$ which, in combination with $L$, is used to train a toponym resolver, if necessary (in particular, if we are dealing with a variant of WISTR). We then use the toponym resolver to resolve toponyms in $A$, in combination with $L$, if necessary (in particular, if we are dealing with a variant of SPIDER). Finally, the documents in $A$ are removed from $L$ and added to $U$.

5. I then generate smaller document windows of a certain size (e.g. 20 words) surrounding each toponym in $A$; see above. These are added to $L_w$, and used to train a document geolocator. This document geolocator is then interpolated with the pre-trained document geolocator (see above, and §6.3.3).

Figure 6.1 shows the basic algorithm. It has a number of subfunctions, whose purpose is as follows:

**TRAINDOCGEO** Train a document geolocator given a corpus of documents—in this case, small *pseudo-documents* that consist of a window (usually 20 words) surrounding a toponym, labeled with the coordinate of that toponym.

**INTERPOLATE** Interpolate between two document geolocators (§6.3.3).

**LABEL** Label a corpus using a document geolocator.

**CHOOSEBATCH** Select a subset of a corpus according to some criterion, e.g. the score exceeds some threshold (§6.3.4).

**FILTERCANDNEARLOC** Select a subset of a corpus, choosing documents that have a toponym with a candidate near the document geolocation, according to some threshold (§6.3.4).

**TRAINTOPRES** Train a toponym resolver. This is passed both $L$, which has toponym annotations, and $A$, which does not. This step is only necessary for one of the WISTR variants. For example, we may derive WISTR$^*$-style features from the toponyms in $L$ and WISTR features from the toponyms in $A$.

**RESOLVE** Resolve toponyms in a corpus. We are passed in both $L$, which has existing toponym annotations, and $A$, which does not. $L$ is only used in one of the SPIDER variants, which do

```
function Train($DG_w, U$)    ▷ On entry: Wikipedia document geolocator, unlabeled corpus
    $L \leftarrow$ new empty corpus    ▷ Labeled corpus
    $L_w \leftarrow$ new empty corpus    ▷ Labeled corpus of “document windows”
    $DG \leftarrow null$    ▷ Document geolocator that will be trained
    $TR \leftarrow null$    ▷ Toponym resolver that will be trained
    Loop
        $DG_0 \leftarrow$ TrainDocGeo($L_w$)
        $DG \leftarrow$ Interpolate($DG_w, DG_0$)
        $DG$.Label($U$)
        $C \leftarrow$ ChooseBatch($U$)    ▷ high-probability docs
        $A \leftarrow$ FilterCandNearLoc($C$)    ▷ docs “accepted” by toponym resolver
        if $A$ is empty then
            break
        end if
        $TR \leftarrow$ TrainTopRes($L, A$)
        $R \leftarrow TR$.Resolve($L, A$)    ▷ Resolved version of $A$
        $R_w \leftarrow$ GetDocWindows($R$)    ▷ Document windows derived from $R$
        for all $d \in R_w$ do
            $L_w$.Append($d$)
        end for
        for all $d \in R$ do
            $U$.Remove($d$)
            $L$.Append($d$)
        end for
        if $U$ is empty then
            break
        end if
    End Loop
    return $(DG, TR, L)$
end function
```

Figure 6.1: Basic co-training algorithm

joint annotation. We have a choice as to whether we re-resolve the toponyms in $L$ or hold them constant while resolving toponyms in $A$ (§6.3.4). The algorithm as described in Figure 6.1 assumes that $L$ is held constant.

**GETDOCWINDOWS** Create a corpus of document windows surrounding each resolved toponym in a corpus (see above).

**APPEND** Append an item to the end of a list.

**REMOVE** Remove an item from a list.

### 6.3.3 Interpolation

Interpolation between a *foreground document geolocator* $DG_0$ trained on $L_w$ (see above) and a large *background document geolocator* $DG_w$ pre-trained on Wikipedia is necessary because document geolocation often requires a fairly significant amount of textual data to produce reasonable results, and $L_w$ will often be too small.

However, interpolation is complicated by the necessity to interpolate between the individual cells of two sets of rankings over grids that may not be the same. There are two issues involved here:

- The grids will typically have different holes in them, due to the distinct datasets used to create them.
- The grids may have different shapes. This necessarily happens, for example, when *k*-d tree grids are used, although it can be avoided with uniform grids by using the same grid size for both.

The first issue is perhaps the trickier one. One possibility is to ignore cells that don't occur in both rankings, but this causes many problems (e.g. it is possible that no cells are common to both rankings). In reality it appears that it is necessary to hallucinate scores for cells that are lacking them, similar to language models. This suggests that $DG_w$ should be viewed as comparable to a global language model, which is either interpolated into the foreground document geolocator $DG_0$ or backed off to.

As with language models, a basic interpolation model can be defined as

$$DG = (1 - \lambda)DG_0 + \lambda DG_w$$

for a given cell. In the experiments I ran, I simply used a constant value of $\lambda$, comparable to Jelinek smoothing. However, a better idea would be the analogue of Dirichlet smoothing, where $\lambda$ varies according to the relative sizes of the corpora, e.g.

$$\lambda = \frac{|DG_w|}{|DG_w| + m|DG_0|}$$

where $|G|$ for some geolocator $G$ is the number of documents used to build the geolocator, and $m$ is an "importance factor" indicating how much larger $DG_0$ should be considered than its actual size. The effect here is that $\lambda$ gets smaller as $DG_0$ gets larger (is built on more documents).

The second issue above, that of differently-shaped grids, can be solved in various ways. One simple way is, for a given cell in $DG_0$, to find its centroid and locate the corresponding cell in $DG_w$ containing that centroid; this is what I currently do. More sophisticated solutions might involve choosing a number of (equally spaced?) points within the cell in $DG_0$, matching each one up to a cell in $DG_w$, interpolating between each pair, and averaging the results.

### 6.3.4 Additional considerations

**Batch sizes and toponym acceptance rates** As described above, the algorithm has two steps, ChooseBatch and FilterCandNearLoc, both of which winnow down the size of the corpus that is passed from the document geolocator to the toponym resolver.

The purpose of ChooseBatch is to eliminate those documents that the document geolocator is most unsure of, to provide a higher-quality set of documents to train the toponym resolver. The hope is that later iterations will train a document geolocator that is able to better geolocate these documents. In the most extreme case, exactly one new document is chosen each round (Nigam and Ghani, 2000). Abney (2007) suggests selecting those documents whose prediction probability is above some threshold, but it is unclear what threshold to use, and whether this threshold should change from round to round. A better idea is probably to select a given number of documents. This is what I have implemented. (In my experiments, I set this number to 1000, on the assumption that around this many documents would be necessary to train a reasonable geolocator. This number may be too high; the experiments in Chapter 5 show that reasonably accurate geolocators can be trained with many fewer documents.)

The motivation for FILTERCANDNEARLOC, which selects only documents containing a toponym with a candidate near the document geolocation, is similar; presumably, a document without toponyms that can be resolved near the document's location is one that is more likely to have an incorrect document location. This should especially help with the WEIGHTED and ADDTOPO variants of SPIDER, which directly use the document location as part of the resolution mechanism. (WISTR uses the document geolocation in its training mechanism, but already rejects toponyms without a candidate near the document's location.)

In my code, the threshold I use in FILTERCANDNEARLOC is 100km. Note that this means that the process eventually terminates having only processed some fraction of the total documents. One possibility I have implemented is an additional iterative loop, where the acceptance threshold start outs small, e.g. 10km, and eventually increases up to some maximum, e.g. 500km. In this fashion, the earlier runs are smaller, which should theoretically increase co-training accuracy, but ultimately a greater fraction of the total set of documents is processed. (However, this variant did not produce results better than simply using a fixed threshold.)

**Whether to re-resolve the entire corpus at each step** In the algorithm as I have described it above, once a given toponym has been resolved, its value never changes. (If this were not the case, for example, we would need to recompute $L_w$ from scratch each round, rather than appending to it.) This is a reasonable decision (Abney, 2007), and may lend the algorithm increased stability, but it is not the only possible one. SPIDER, for example, does joint resolution over all the toponyms it considers. It is for this reason that the RESOLVE step has both $L$ (the already-labeled documents) and $A$ (the not-yet-labeled documents) passed to it — so that SPIDER can make use of the toponyms in $L$ when resolving those in $A$. This assumes that SPIDER does not change the labels on $L$. However, it is quite possible that, given the additional set of toponyms in $A$ to resolve, it can do a better job

| Corpus fraction | Co-train-selected fraction | | Full corpus | |
|---|---|---|---|---|
| Method/Geolocator | Base | Co-trainer | Base | Co-trainer |
| SPIDER | 74% | 84% | 36% | 42% |
| SPIDER (WEIGHTED) | 79% | 88% | 37% | 41% |
| SPIDER (ADDTOPO) | 80% | 90% | 50% | 57% |
| WISTR | 76% | 86% | 54% | 60% |
| WISTR+SPIDER | 76% | 86% | 41% | 46% |
| WISTR+SPIDER (ADDTOPO) | 79% | 90% | 51% | 58% |

Table 6.4: Acc@161 for the CWAR dev set when evaluating using error distance to closest resolved toponym with toponyms resolved by a co-trained toponym resolver, comparing a base geolocator trained on Wikipedia and a co-trained geolocator.

relabeling $L$ than it did when labeling $L$ the first time around.

### 6.3.5 Results

One method for evaluating the performance of a document geolocator given a toponym-annotated and resolved corpus is to look at the nearest resolved toponym to the document-level annotation and treat the distance between the two as the error distance for the document. This allows for the use of the same metrics as are used elsewhere in this dissertation to evaluate document geolocations.

By this metric, co-training yields noticeable increases in accuracy compared with a Naive Bayes geolocator trained only on the base Wikipedia corpus, when both are evaluated on the CWAR corpus. This is shown in Table 6.4. Note that the WEIGHTED and ADDTOPO variants of SPIDER both perform better than standard SPIDER— a pleasing result.

# Chapter 7

# Conclusion

In this dissertation I investigated geolocation of a document (automatically identifying its location) solely using the document's text. The basic thesis that I maintained throughout the dissertation was that this was possible with sufficient accuracy to enable useful, real-world geographic investigations of textual corpora. Although metadata is plentiful in some types of corpora (e.g. Twitter, where user profiles provide home location, time zone, friends, followers, etc.), it is lacking in many corpora in many other domains, such as the digital humanities, a major target of my methods.

In the first part of the dissertation I developed methods for accurate text-based geolocation, based on both uniform and adaptive (*k*-d tree) grids over the Earth's surface and using various machine-learning methods. These culminate in a new technique based on hierarchical logistic regression, which achieves state of the art results on well-known corpora (Twitter user feeds, Wikipedia articles and Flickr image tags).

In the second part of the dissertation I applied these methods to the digital humanities. To my knowledge, I am the first to apply text-based document geolocation to historical text-only corpora, and to this end I developed a new NLP task for this purpose. Because there are no existing corpora to test on, I annotated two historical corpora of significantly different natures (BEADLE, a travel log, and WOTR, a large set of Civil War archives) and showed how my methods produce good geolocation accuracy even given the relatively small amount of annotated data available. I then used the predictions on the much larger unannotated portion of WOTR to generate and analyze geographic topic models, showing how they can be mined to produce interesting revelations concerning various Civil War-related subjects. Finally, I developed a new geolocation technique for text-only corpora involving co-training between document-geolocation and toponym-resolution models, using a gazetteer to inject additional information into the training process. To evaluate this technique I developed a new metric, the closest toponym error distance, on which I showed improvements compared with a baseline geolocator.

In the rest of this chapter I discuss avenues for further research and investigation.

## 7.1 Further corpus annotation

Both of the corpora I annotated are concentrated in the United States. It would be useful to annotate another corpus that covers a world-wide domain. This would add a good point of comparison, similar to the difference between the two large Twitter corpora I analyzed in the first part of my dissertation (§2.2.1)—one North-America-specific and the other worldwide. This would require a good deal of annotated material to get a reasonable distribution, perhaps significantly more than the approximately 5,000 annotated articles in WOTR. On the other hand, my experimental results with learning curves showed that the large majority of the benefit of the available annotated material was gained using 50-75% of the total in a pure in-domain setting (§5.3.1) and closer to 25% using domain adaptation with Wikipedia (§5.3.2), meaning that a corpus half or even a quarter the size might have sufficed nearly as well. This suggests that the proportionately larger version of this smaller corpus that would be required for worldwide scope might not need to be too much more than 5,000 articles in size. Further annotation efficiencies could probably be achieved using active learning, directing the annotator to the areas most in need of annotation material. (In fact, I employed a variant of this technique while overseeing the annotation of the Civil War material (§2.3.2), creating frequent KML maps of the corpus distribution so far and using these maps to inform my decisions about what needed to be annotated next.)

## 7.2 Improvements to the information-retrieval model

The information-retrieval model that underlies the methods described in this dissertation assume that what is in reality a continuous space of latitude/longitude coordinates can be partitioned into a grid of discrete cells, with independent language models constructed for each cell and the predictions of all points that geolocate to a given cell resolved to the same location. In this dissertation I explored various improvements to the basic uniform-grid Naive Bayes model that serves as the baseline, such as adaptive grids, cell centroids, and flat and hierarchical logistic regression. Adaptive grids such as *k*-d trees (§3.2.2) have the potential to improve results over uniform grids by efficiently adapting the grid structure to the non-uniform geographic distribution of documents, avoiding the wastage that comes from having some cells with very many documents and others with only a few. Using the centroid of a cell rather than its geographic center further accounts for the non-uniform distribution of documents within a cell and almost always yields improvements. It is well-known that logistic regression generally outperforms Naive Bayes and KL-Divergence. The hierarchical approach that I proposed solves a number of problems with the simpler flat grid approach, such as resolving what I term the *information/resolution tension*, i.e. the tension between the opposing goals of incorporating more information into each grid cell (through making larger grid cells with more training documents per cell) and increasing the resolving power of each cell (through shrinking the size of the cells, thereby reducing the minimum error distance).

Many further improvements are possible. In the following sections, I organize these improvements according to the algorithmic stage at which they apply: when creating the per-grid language models, when choosing a grid cell, and when choosing the representative point for the grid cell.

### 7.2.1 Improvements in language-model creation

At the language model creation level, one fairly simple method is to use a fine grid and smooth over nearby cells when creating the language model for a grid cell, i.e. create language models that interpolate in some fashion between the raw language model of the cell itself and those of its neighbors. This allows for the resolving power of fine grid cells to combine with the effective informing power of coarser grid cells. This was done, for example, in Serdyukov et al. (2009) and O'Hare and Murdock (2013), although it helped very little in their application to Flickr images, perhaps due to the simplistic method they employed, which used a uniform grid and defined "neighbors" using a small, fixed radius. In this case, smoothing with a wider cell radius and/or using *k*-d trees might provide greater benefits. Another, more sophisticated method I could imagine would start with a clustering step that would group together cells with similar distributions and use the resulting clusters to condition smoothing. This has the effect of increasing the information content of each cell while avoiding the diluting effect that comes from smoothing with dissimilar cells. The clustering and/or smoothing could also be done independently for each word, which might be a superior (if compute-intensive) approach as it would allow for finer-grained conditioning of the smoothing process. (This is somewhat similar to an approach followed by a colleague of mine, Grant DeLozier, for doing toponym resolution (DeLozier et al., 2015). In his case, he used per-word GI statistics, a measure of local spatial autocorrelation, to smooth the geographic distributions of individual words and derive effective clusters for the purpose of resolving toponyms without the need for a gazetteer.)

There is also the possibility of incorporating more sophisticated features than the unigram models I made use of. One of the simplest improvements is the use of bigram or higher n-gram features. In my preliminary experiments, I found little if any gain from bigram features, and did not pursue them further. However, van Laere et al. (2014), using a hybrid textual geolocator on Wikipedia, reported improvements from 67.05% to 69.71% on a particular accuracy measure using bigrams vs. unigrams, and much smaller improvements from higher n-grams, which suggests that more careful tuning could yield significant benefits.

Other more sophisticated features to consider are those based on morphological, part-of-speech (POS), syntactic or named-entity-recognition (NER) analysis of the text. I suspect that NER analysis may be the most fruitful, given its ability to recognize multi-word expressions with potential geographic scope (e.g. places, people). Twitter poses special problems, as linguistic analysis of Twitter is known to be difficult (Foster et al., 2011). However, Twitter-specific packages have been developed for some of these tasks, such as TweetNLP (Owoputi et al., 2013; Gimpel et al., 2011) (POS tagging) and TwitIE (Bontcheva et al., 2013) (POS tagging, NER). For Wikipedia and historical texts, standard off-the-shelf tools may be sufficient, either by themselves or in combination with domain-adaptation techniques such as self-training.

One experiment worth considering with WOTR is to do runs with and without the header, which frequently contains a location in it. Often, this location is the location of the overall document, although sometimes it is not. It would be interesting to see how well a document geolocator can do when the header is stripped out, and conversely, how well a simple toponym-based approach can do when applied only to the first location in the header.

### 7.2.2 Improvements in grid-cell selection

At the level of choosing a grid cell, the approach I follow simply ranks the cells by score and chooses the highest-scoring cell. There are other, potentially more sophisticated methods, such as the *mean shift algorithm* (Comaniciu and Meer, 2002). The idea is that, especially with a fine grid, if the topmost cell is geographically isolated and immediately below it in the ranking is a cluster of cells located somewhere else, that cluster may be more likely to represent the correct location than the top-ranked cell. The mean shift algorithm can be used to find this cluster. I implemented this algorithm, but preliminary experiments did not yield improvements. However, it is possible that careful tuning or or a more sophisticated implementation would work better.

Another approach is to *rerank* the topmost cells from the initial ranking on the theory that, when the topmost cell is wrong, the correct cell (or at least a better cell) is likely to be among the top few cells, and thus by reranking the cells with more information, a better ranking could be achieved. The idea is that, given the limited number of cells to be reranked, a better ranking algorithm with significantly more sophisticated features could be employed than was possible in the initial pass, with the need to potentially rank thousands of cells. Reranking has been successfully used for parsing (Collins and Koo, 2005), sentence boundary detection (Roark et al., 2006), grounded language learning (Kim and Mooney, 2013), machine translation (Olteanu et al., 2006), and other areas.

I spent a good deal of effort implementing reranking, using a logistic-regression reranker on top of an initial Naive-Bayes or KL-divergence ranker. I also evaluated, in place of logistic regression, a passive-aggressive perceptron. One of the nice features of this technique is the ability to introduce a cost function into the algorithm to directly model the cost of making a wrong choice rather than simply assuming all mistakes are equally bad. This allows the learner to directly target the metric. In this case, the metric is the error distance and I set the cost to relate to the distance between the correct and incorrect choice, reflecting the intuition that choosing an incorrect location that is 10,000 km from the correct one is much worse than choosing a location 5 km from the correct one. Unfortunately, I was consistently unable to beat the initial ranker. However, it is possible that more sophisticated features (§7.2.1) will yield better results.

Yet another possibility would be to implement a variant of the *k-nearest-neighbors* (kNN) algorithm (Altman, 1992). This is a nonparametric technique that chooses an output based on averaging or voting among the *k* nearest training examples to the test document, for some small *k*. In this case "nearest" is based on textual comparison, for example using the *Hamming distance* for comparing two vectors of categorical variables (i.e. only consider the presence or absence of words), or a measure such as KL-divergence. kNN can be used for both classification and regression, and in this case it seems to make more sense to view the the geolocation task as one of regression, despite the fact that I have generally set up geolocation as a classification problem. Geolocation in fact is more naturally conceived of as regression over a continuous two-dimensional space; the use of classification is for computational tractability, an issue that does not come up with kNN. Among the reasons why kNN might work well in this problem is that the local and nonparametric nature of kNN is well-suited for highly non-linear output spaces such as is the case with geolocation; on the other hand, careful tuning will be necessary to avoid overfitting the training set. The requirement to keep the entire training set around at evaluation time would not be a problem for the small historical data sets I consider, but might be prohibitive with a large data set such as Wikipedia. To deal with such a case, approximate nearest-neighbor techniques such as *locality-sensitive hashing* could be used (Andoni and Indyk, 2008).

### 7.2.3 Improvements in choosing a representative point in a cell

Finally, at the bottom level of choosing a representative location for a document given a grid cell, the approach I follow uses the centroid of the training documents. This is clearly superior to the approach used in Wing and Baldridge (2011), which used the geographic center of the cell, since it takes advantage of the potentially non-uniform distribution of documents in the cell, as was mentioned above. This is especially advantageous with large grid cells, such as might be encountered using *k*-d trees. In *k*-d tree distributions, cells over areas lacking training documents, such as the ocean, will be extremely large, and it is quite possible (indeed, common in the WOTR corpus) for such a cell to include a sliver of land in one corner containing most or all of the training documents. In such a scenario, use of the geographic center would lead to very large, avoidable errors. (An alternative to the centroid is the median location of the training documents, as used in Rahimi et al. (2015b). Based on experience in *k*-d trees, where an analogous choice comes up, I suspect that the difference will be small.)

Nonetheless, the use of the centroid still chooses a constant point for the cell, irrespective of the test document. It seems logical that improvements could result from tailoring the result to the particulars of the test document, e.g. through an additional search step to find the most similar training document. A similarity search could enable a fairly coarse grid to be used, with the two-step procedure of choosing a grid cell and then doing a similarity search within the cell analogous to a two-level version of the hierarchical approach described above, and finessing the information/resolution tension in the same fashion. This approach was in fact followed by van Laere et al. (2013) for Flickr, to great effect, and extended to Wikipedia in van Laere et al. (2014). Both papers used Jaccard similarity (Real and Vargas, 1996), although the second paper also considered the Apache Lucene library (McCandless et al., 2010), and reported mixed results when comparing the two, with neither obviously better than the other. Such an approach has the advantage of potentially reducing the error distance to a much smaller quantity than a centroid or median-selection algorithm. A potential disadvantage of such a method is that it is transductive and requires keeping the entire training set available at test time; this could result in very large models for large training sets such as Wikipedia, but should not be an issue for the relatively small annotated historical corpora I developed.

## 7.3 Grid partitioning

The issue of grid partitioning has already been discussed, in the context of uniform and *k*-d tree grids. As mentioned, *k*-d grids adapt to the distribution of the training data and can potentially make more efficient use of the typically non-uniform distribution of geographic information in real-world corpora. In Wikipedia and in social-media-based corpora, for example, urban areas are likely to be much better represented than rural areas, and in Civil War corpora such as WOTR, areas that saw heavy fighting such as Virginia and the lower Mississippi River have many more associated documents than areas with only sporadic fighting, such as the upper Midwest (see Figure 5.11).

Both grid-partitioning methods, however, still make use of rectangular grid cells in latitude/longitude space, which have the undesirable property (especially from the perspective of a uniform grid) that they become progressively less square and more elongated the closer the distance to the poles, due to the convergence of all lines of longitude at the north and south pole. A truly equal-area grid, the quaternary triangular mesh, does exist and was considered by Dias et al. (2012), although their results do not directly compare with ours due to differing experimental methodology. Grids in general, however, suffer from the *modifiable areal unit problem* (Gehlke and Biehl, 1934; Openshaw, 1983), the inevitable anomalies caused by drawing grids through areas that are continuous in distribution. A truly continuous distribution would alleviate that, and has been considered by Eisenstein et al. (2010) and a number of other papers continuing in this vein (e.g. Eisenstein et al. (2011b); Hong et al. (2012); Ahmed et al. (2013)), using complex Bayesian models generally learned through variational inference (see also §1.2). However, these papers generally only evaluated on the very small GEOTEXT corpus, and it's unclear they can be expanded to larger corpora or generalized beyond GEOTEXT to the digital humanities corpora I created.

Another alternative is to use clustering to create a perhaps more natural set of grid points. This was done by Han et al. (2014) for cities in connection with Twitter, using a gazetteer to group smaller cities with nearby larger cities occurring within the same administrative unit (e.g. state of the U.S.) and geolocating specifically to the center of these "city attractors" rather than to the centroid of any grid cells. This avoids various problems associated with grids but only works well if the data is primarily city-based. Even with Twitter, where this is indeed the case, they were unable to geolocate the 8% of Tweets which were not near a city, thereby significantly decreasing accuracy; for a data set like WOTR, detailing battles often fought nowhere near a city, such a restriction would be catastrophic. A better compromise is found in van Laere et al. (2013), who used mean shift and K-medoid clustering to create grids, comparing against uniform grids and finding K-medoid grids the best of all. Although they didn't specifically compare against *k*-d grids, the fact that their results were significantly better with K-medoid grids vs. uniform grids is promising, since *k*-d grids do not always beat uniform grids, and is the area I would explore first before considering the other possibilities mentioned above.

## 7.4 Geographic topic model improvements

The geographic topic models I applied to the Civil War archive shoehorn the dynamic topic models of Blei and Lafferty (2006), which were designed for a one-dimensional space such as time, into a two-dimensional geographic space. In order to derive various topics and study their distribution over a set of expert-defined geographic "theaters of war", it was necessary to linearize the theaters, which is problematic in that a given theater typically comes into contact with multiple other theaters. A better system would more directly model the connections between such regions, allowing a network of connected regions to be defined and propagating statistical similarities between region-specific variations of a given topic along those connections. Note that this model is general enough that it can directly model a dynamic topic model over combined geographic areas and timeslices, and thus simultaneously investigate variation over both time and space. It should be possible to extend Blei's variational inference algorithm to handle this more general connected network; it may also be possible to use another algorithm such as label propagation.

There are also other possible ways of defining geographic topic models. For example, Yin et al. (2011) define a number of such models, the most sophisticated of which simultaneously learns topics and coherent geographic regions. However, these topics differ from the topics of a dynamic topic model in that Yin's model finds individual topics that are regionally coherent, whereas the proposed extension of Blei's model finds general topics (e.g. "reconnaissance") and examines how those topics change over different geographic areas. The latter "variationist" approach may be more useful for the digital humanities researcher.

## 7.5 Statistical relations among document predictions

The basic mechanism I use to geolocate a document makes independent predictions for each document. This greatly simplifies the prediction process and makes sense for Wikipedia articles and Twitter users, which can be viewed as unordered collections of documents. However, in both of the historical corpora I considered, there are heavy statistical dependencies between adjacent documents. Paragraphs in BEADLE are organized by narrative structure, and letters/reports/etc. in WOTR are grouped by region and campaign and then ordered chronologically. This suggests that significant improvements might result from a sequence model, which would work somewhat analogously to a Hidden Markov Model (HMM, Ghahramani (2002)), with each grid cell corresponding to a state in the model. There are too many grid cells to use a classic HMM with the transition probability defined purely using a state table; instead, I could envision smoothing the state-based transition probabilities and incorporating a distance-based transition probability, or more properly an interpolation between a "dependent", distance-based probability of moving from the location of the previous article to the location of the current article and an "independent" probability of making a jump to a new location. The overall model would work similarly to e.g. Chen and Grauman (2011), which defines an HMM-based model over the locations of chronological sequences of tourist photos. Their model also implements "burstiness", assuming that there will be sequences of photos taken at the same location, and they show that this better models their data than not including burstiness. The analogous feature would probably be of benefit for both WOTR and BEADLE, and for WOTR the date of the article (which can usually be extracted from the article's text using pattern matching) could be used to help predict the extents of individual "bursts".

There are other ways of taking advantage of statistical relationships among documents' linear ordering. For example, if there are embedded links between documents that are likely to be statistically correlated, a network of such links can be created and techniques such as label propagation (Rahimi et al., 2015b; Jurgens, 2013)) or total variation minimization (Compton et al., 2014) used to statistically tie together documents that are nearby in the network. See also Jurgens et al. (2015) for a good recent overview of the state of the art in network-based geolocation. These network-based techniques have proven to be especially useful for Twitter, with users connected in friend/follower relationships (*ego networks*) and through textual references (*@-mentions*) to other users; these connections have been shown to correlate with the physical proximity of the users (Takhteyev et al., 2011). Additional such relationships can be found in Wikipedia (hyperlinks between documents) and in image corpora such as Flickr (relationships between images authored by the same user).

For pure-text corpora such as the historical corpora considered here, the metadata necessary to create such links is not available. However, it is possible to imagine creating links directly based on words in the text. This could be especially useful in conjunction with semi-supervised techniques that are able to learn from unannotated as well as annotated data. For example, it may be possible to create a label propagation network for names, places and other terms, in conjunction with a named-entity recognizer, toponym resolver, or similar resource making use of outside knowledge. Contrast this with the information-retrieval model I use, which considers individual words independently when finding their geographic distribution and is unable to learn from unannotated text. As a motivating example, imagine we have two names, a common name A and a relatively rare name B. Only A occurs often enough among our annotated documents to derive a reasonable statistical distribution of associated locations, but we can observe from our unannotated documents that B frequently co-occurs with A. (Possible values for A and B in WOTR are General William T. Sherman and the name of one of the officers who served in his army.) We should then be able to make conclusions about the location of a new document that mentions B but not A, and a technique such as label propagation should make such indirect inferences possible.

## 7.6 Error analysis and significance testing

**Error analysis** More work could be done on error analysis. I did compute heatmaps of various sorts, e.g. comparisons of the correct and predicted locations and the locations with the maximum errors. I also generated lists of such locations and examined them. However, it was often difficult to detect patterns in the errors. There is much that remains to be done. One important area is to investigate what sorts of different errors occur in hierarchical logistic regression vs. plain logistic regression or Naive Bayes—does hierarchical LR do better in areas with more training data, for example?

**Significance testing and averaging** So far I have done no significance testing on the various accuracy, mean error and median error figures I have produced. Thus, I do not know for sure which differences are statistically significant and which ones aren't. This is important given that some of the differences between geolocation methods are quite small. Furthermore, from the experiments in §5.3.1 that involved averaging over several randomized permutations of training and test data, it is clear that the variance of individual results is high. This suggests that results should be recalculated using multiple permutations and/or 10-fold cross validation, which would have the effect of decreasing the variance and thereby increasing the confidence that a given difference is statistically significant.

## 7.7 Remaining issues and final thoughts

There are many other directions that this research could be taken. Here I address a few points that I haven't previously touched upon.

**Regions vs. points** The data contained in the WOTR corpus has polygons as well as points, and has multiple points per document. To make use of these annotations in my geolocation methods, which were designed for single-point corpora such as Wikipedia and twitter, I reduce them to a point reference by taking the centroid of the points and polygons in question. This works fairly well for small polygons and for nearby points, but produces distortions when applied to larger regions and widely-separated points. It is far from obvious, for example, that a state or country, which may be hundreds or thousands of miles wide, can reasonably be approximated by a point in the middle. Similarly, using the centroid of a set of widely separated points results in a location that may be nowhere near any of those points, again far from optimal. Conceptually, a textual span annotated with a large region or set of widely separated points has a vague reference, and can be equally well represented by any location within the region, or by any one of the points; choosing a single "correct" point and using a point-based distance metric introduces a large source of error that is difficult or impossible to avoid even with a perfect geolocation algorithm. (Note that the common practice in Wikipedia of identifying a region by its capital city will result in even worse errors than using the centroid. In the case of the United States, for example, the capital city of Washington, D.C. is near the East Coast, introducing a built-in error nearly twice what would result from using the proverbial "point in Kansas", i.e. close to the country's geographic center.) This built-in error may cause a metric such as *acc@161* (§4.1.2) to register an accuracy failure regardless of how accurate the geolocation algorithm is, and will introduce significant distortions into mean and median error.

It is an open question how to properly handle polygons and point sets in a grid-based geolocation system. At training time, for example, the probability mass associated with a given instance could be spread over all the grid cells covered by a polygon. (Whether this would generalize to multiple points is unclear.) It seems clearer what to do at evaluation time: use as the error distance the smallest distance between the predicted point and any point in the polygon. This captures the intuition that a textual span geolocated to a large region such as a state has a correspondingly vague reference, so that any point within the region should be considered correct.

**Further use of probabilities** Methods such as Naive Bayes and logistic regression generate actual probabilities. So far I use them only as a score, choosing the highest-scoring cell. However, it's possible to use the actual values, e.g. to estimate the confidence of the prediction. It would be interesting, for example, to see which predictions are the most confident, and whether that correlates with the presence of certain toponyms in the text or other identifiable features. (Keep in mind, however, that the probability estimates from Naive Bayes at least are notoriously unreliable.)

**North vs. South** WOTR contains primary sources from both sides of the Civil War. It may be useful to separate the sources in this fashion. This would allow, for example, creating topic models to investigate differences in how the North and South viewed major battles between them. Whether a given document was written by a Northerner or Southerner could also serve as a very useful feature during geolocation. A classifier could be constructed to separate North from South based on features such as language. It should be possible to quickly seed such a classifier with training data by using the fact that many of the volumes group the North-authored and South-authored documents together.

**Tracking war participants** Through a combination of geolocation, date parsing, and named entity recognition, it should be possible to track the movements through time and space of individual Civil War participants (e.g. soldiers, generals, politicians) and of larger military units. This would allow, for example, the construction of maps that enable digital humanities researchers to get a better high-level perspective of the progress of the war.

In conclusion, I have shown through my research how methods that are conceptually relatively simple (such as hierarchical logistic regression) can yield innovative solutions to difficult problems. I have also demonstrated how restricting the type of information that one can make use of—in this case, text-only geolocation, ignoring metadata—can open up new avenues for NLP applications in unexpected domains such as the digital humanities. Those researchers who make use of metadata such as is provided by Twitter tend to focus only on Twitter—they have their hammer, so to speak, and they ignore things that don't look like nails. I hope that my research, and the discussion in this final chapter, will enable others to take up where I have left off and continue to improve the state of the art.

# Bibliography

## References

Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC Computer Science & Data Analysis. CRC Press, Boca Raton, FL, USA.

Lada A. Adamic. 2000. Zipf, power-laws, and pareto—a ranking tutorial. Xerox Palo Alto Research Center, Palo Alto, CA. Online at `http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html`. [Accessed 16-Jun-2015].

Benjamin Adams and Krzysztof Janowicz. 2012. On the geo-indicativeness of non-georeferenced text. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM'12: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. 2014. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15:1111–1133.

Amr Ahmed, Liangjie Hong, and Alexander J. Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 25–36, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

N. S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Alexandr Andoni and Piotr Indyk. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122, January.

Faiz Anuar and Ulrike Gretzel. 2011. Privacy concerns in the context of location-based services for tourism. In *ENTER 2011 Conference*.

Edward L. Ayers. 2007. Valley of the shadow. Virginia Center for Digital History and University of Virginia Library. Online at `http://valley.lib.virginia.edu/`. [Accessed 9-Jun-2015].

Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. 2008. Spatial variation in search engine queries. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 357–366, New York, NY, USA. ACM.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 61–70, New York, NY, USA. ACM.

Jason Baldridge, Matthew Lease, and Katrin Erk. 2012. Spatial and temporal analysis of multilingual texts. A Proposal to the New York Community Trust. Edition of April 6, 2012.

David Bamman, Adam Anderson, and Noah Smith. 2013. Inferring social rank in an Old Assyrian trade network. In *Digital Humanities*, Lincoln, NE, USA.

Paul N. Bennett and Nam Nguyen. 2009. Refined experts: improving classification in large taxonomies. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *SIGIR*, pages 11–18. ACM.

Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

David Blei and John Lafferty. 2009. Topic models. In Ashok N. Srivastava and Mehran Sahami, editors, *Text Mining: Classification, Clustering, and Applications*, pages 71–94. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Boca Raton, FL, USA.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA. ACM.

David J. Bodenhamer, John Corrigan, and Trevor M. Harris, editors. 2010. *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Indiana University Press, Bloomington, IN, USA.

Keith S. Bohannon. 2014. Guerrilla warfare during the Civil War. In *New Georgia Encyclopedia*. Georgia Humanities Council and University of Georgia Press. [Online; accessed 5-Jun-2015].

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.

Letizia Bollini, Rinaldo De Palma, and Rossella Nota. 2013. Walking into the past: Design mobile app for the geo-referred and the multimodal user experience in the context of cultural heritage. In *Computational Science and Its Applications–ICCSA 2013*, pages 481–492. Springer.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. 2013. TwitIE: An open-source information extraction pipeline for microblog text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.

Charles Booth. 1902. *Life and Labour of the People in London, Volume 1*. Macmillan, London.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187.

Olha Buchel and Linda L. Hill. 2009. Treatment of georeferencing in knowledge organization systems: North American contributions to integrated georeferencing. In *North American Symposium on Knowledge Organization*, Syracuse, NY, USA, June 18–19.

Anne Burdick, Peter Lunenfeld, Johanna Drucker, Todd Presner, and Jeffrey Schnapp. 2012. *Digital humanities*. MIT Press, Cambridge, MA.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

G.J. Houben C. Hauff. 2012. Geo-location estimation of Flickr images: Social web based enrichment. In *ECIR 2012, Proceedings of the 34th European Conference on Information Retrieval*, pages p. 85–96. Springer LNCS 7224, April 1-5.

Chris Callison-Burch. 2002. Co-training for statistical machine translation. Technical report, In Proc. of the 6th Annual CLUK Research Colloquium.

Jonathan Chang and David M. Blei. 2010. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4(1).

Hau-Wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. 2012. @phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 111–118. IEEE Computer Society.

Rachel Chasin, Daryl Woodward, Jeremy Witmer, and Jugal Kalita. 2013. Extracting and displaying temporal and geospatial entities from articles on historical events. *The Computer Journal*, page bxt112.

Chao-Yeh Chen and Kristen Grauman. 2011. Clues from the beaten path: Location estimation with bursty sequences of tourist photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1569–1576.

Minmin Chen, Kilian Q. Weinberger, and John Blitzer. 2011. Co-training for domain adaptation. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2456–2464. Curran Associates, Inc.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2013. A content-driven framework for geolocating microblog users. *ACM Trans. Intell. Syst. Technol.*, 4(1):2:1–2:27, February.

Pilar Chias and Tomas Abad. 2009. Geolocating and georeferencing: Gis tools for ancient maps visualisation. In *Information Visualisation, 2009 13th International Conference*, pages 529–538. IEEE.

Benjamin G. Cloyd. 2010. *Haunted by Atrocity: Civil War Prisons in American Memory*. Making the Modern South. Louisiana State University Press.

Patricia Cohen. 2011. Digital maps are giving scholars the historical lay of the land. *The New York Times*, July 26.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Comput. Linguist.*, 31(1):25–70, March.

Dorin Comaniciu and Peter Meer. 2002. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619.

Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging one hundred million Twitter accounts with total variation minimization. *CoRR*, abs/1404.7152.

Gregory Crane. 2012. The Perseus project. In William Sims Bainbridge, editor, *Leadership in Science and Technology: A Reference Handbook*, pages 644–653. SAGE Publications, Inc.

Yves Croissant, 2013. *mlogit: multinomial logit model*. R package version 0.2-4.

Adam Crymble. 2012. Review of paper machines, produced by chris johnson-roberson and jo guldi. *Journal of Digital Humanities*, 2(1). Online at `http://journalofdigitalhumanities.org/2-1/review-papermachines-by-adam-crymble/` [accessed 16-Jun-2015].

Geoff Cunfer. 2008. Scaling the Dust Bowl. In Amy Hillier and Anne Kelly Knowles, editors, *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*, pages 95–121. ESRI Press, Redlands, CA, USA.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 53–59, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2382–2388.

Inderjit S. Dhillon and Suvrit Sra. 2003. Modeling data using directional distributions. Technical Report TR-03-06, The University of Texas at Austin, January.

Duarte Dias, Ivo Anastácio, and Bruno Martins. 2012. A language modeling approach for georeferencing textual documents. In *Proceedings of the Spanish Conference in Information Retrieval*.

Junyan Ding, Luis Gravano, and Narayanan Shivakumar. 2000. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 545–556, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

G. Dutton. 1996. Encoding and handling geospatial data with hierarchical triangular meshes. In M.J. Kraak and M. Molenaar, editors, *Advances in GIS Research II*, pages 505–518, London. Taylor and Francis.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA, October. Association for Computational Linguistics.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011a. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1365–1374, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacon Eisenstein, Ahmed Ahmed, and Eric P. Xing. 2011b. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1041–1048.

Katrin Erk. 2013. Towards a semantics for distributional representations. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 95–106, Potsdam, Germany, March. Association for Computational Linguistics.

James W. Erwin. 2012. *Guerrillas in Civil War Missouri*. The History Press, Charleston, SC, USA.

Alessandra Ferrighi. 2015. Cities over space and time: Historical gis for urban history. *Handbook of Research on Emerging Digital Tools for Architectural Surveying, Modeling, and Representation*, page 425.

Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. #hardtoparse: POS tagging and parsing the Twitterverse. In *Analyzing Microtext*, volume WS-11-05 of *AAAI Workshops*. AAAI.

Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3:209–226.

C. E. Gehlke and Katherine Biehl. 1934. Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185):169–170.

Zoubin Ghahramani. 2002. An introduction to hidden markov models and bayesian networks. In *Hidden Markov Models*, pages 9–42. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the*

*49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Goldstone and Ted Underwood. 2012. What can topic models of PMLA teach us about the history of literary scholarship? *Journal of Digital Humanities*, 2(1). Online at `http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/` [accessed 16-Jun-2015].

Siddharth Gopal, Yiming Yang, Bing Bai, and Alexandru Niculescu-Mizil. 2012. Bayesian models for large-scale hierarchical classification. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2420–2428.

Mark Graham, Scott A. Hale, and Devin Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.

Ian Gregory, Alistair Baron, Patricia Murrieta-Flores, Andrew Hardie, and Paul Rayson. 2013. Geographical text analysis: Mapping and spatially analysing corpora. In *Corpus Linguistics*, pages 105–108, Lancaster, July. UCREL.

Jo Guldi. 2009. What is the spatial turn? In *Spatial Humanities: A Project of the Institute for Enabling Geospatial Scholarship*. Online at `http://spatial.scholarslab.org/spatial-turn/what-is-the-spatial-turn/`.

Bo Han, Paul Cook, and Tim Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *International Conference on Computational Linguistics (COLING)*, page 17, Mumbai, India, December.

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

Qiang Hao, Rui Cai, Changhu Wang, Rong Xiao, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. 2010. Equip tourists with knowledge mined from travelogues. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 401–410, New York, NY, USA. ACM.

Shan He and Daniel Gildea. 2006. Self-training and co-training for semantic role labeling: Primary report. Technical report, TR 891, Department of Computer Science, University of Rochester.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA. ACM.

Linda L. Hill. 2006. *Georeferencing: The Geographic Associations of Information*. The MIT Press, Cambridge, Massachusetts, September.

Livia Hollenstein and Ross Purves. 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. *J. Spatial Information Science*, 1(1):21–48.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 769–778, New York, NY, USA. ACM.

David W. Hosmer Jr. and Stanley Lemeshow. 2004. *Applied logistic regression (Wiley series in probability and statistics)*. John Wiley & Sons, 2 edition.

Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. Kernel density estimation for text-based geolocation. In *Proceedings of AAAI 2015*, pages 1390–1395, Austin, TX, USA, January. The AAAI Press.

ISACA. 2011. Geolocation: Risk, issues and strategies. White paper, ISACA.

Chris Johnson-Roberson and Jo Guldi. 2014. Paper Machines. `https://github.com/papermachines/papermachines`. [Version 0.4.9, accessed 16-Jun-2015].

I. A. Junglas and C. Spitzmuller. 2005. A research model for studying privacy concerns pertaining to location-based services. In *Proceedings of the 38th Hawaii International Conference On System Sciences*, January 3–6.

David Jurgens, Tyler Finethy, James McCorriston, Yi Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *International AAAI Conference on Web and Social Media*.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*.

Joohyun Kim and Raymond Mooney. 2013. Adapting discriminative reranking to grounded language learning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 218–227, Sofia, Bulgaria, August. Association for Computational Linguistics.

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pages 61–68.

Anne Kelly Knowles. 2013. A cutting-edge second look at the Battle of Gettysburg. *Smithsonian.com*, June 27. Online at `http://www.smithsonianmag.com/history/A-Cutting-Edge-Second-Look-at-the-Battle-of-Gettysburg-1-180947921` [accessed 02-Feb-2015].

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Vasileios Lampos, Tijl De Bie, and Nello Cristianini. 2010. Flu detector: Tracking epidemics on Twitter. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, pages 599–602, Berlin, Heidelberg. Springer-Verlag.

Lewis Lancaster. 2014. Atlas of maritime Buddhism. In *Electronic Cultural Atlas Initiative*. Online at `http://ecai.org/projects/MaritimeBuddhism.html`. [Accessed 02-Feb-2015].

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5).

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

David D. Lewis. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 4–15, London, UK. Springer-Verlag.

Moshe Lichman and Padhraic Smyth. 2014. Modeling human location data with mixtures of kernel densities. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 35–44, New York, NY, USA. ACM.

M. D. Lieberman and J. Lin. 2009. You are where you edit: Locating Wikipedia users through edit histories. In *ICWSM'09: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, pages 106–113, San Jose, CA, May.

T.Y. Liu. 2011. *Learning to Rank for Information Retrieval*. Springer.

Huchuan Lu, Qiuhong Zhou, Dong Wang, and Ruan Xiang. 2011. A co-training framework for visual tracking with multiple instance learning. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 539–544. IEEE.

Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2012. Where is this tweet from? Inferring home locations of Twitter users. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM'12: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

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

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. 2010. *Lucene in Action: Covers Apache Lucene 3.0*. Manning Publications Co.

Jeffrey McGee, James A. Caverlee, and Zhiyuan Cheng. 2011. A geographic study of tie strength in social media. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2333–2336, New York, NY, USA. ACM.

Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 459–468, New York, NY, USA. ACM.

Thomas McMullan. 2014. Developers must address the ethics of using location data. *The Guardian*, October 20.

Elijah Meeks and Scott B. Weingart. 2012. The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1). Online at `http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/` [accessed 16-Jun-2015].

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 889–892, New York, NY, USA. ACM.

M. G. Michael, Sarah Jean Fusco, and Katina Michael. 2008. A research note on ethics in the emerging age of "uberveillance". *Computer Communications*, 31(6):1192–1199, April.

David M. Mimno. 2009. Reconstructing Pompeian households. In *Applications of Topic Models Workshop, NIPS*, Whistler, BC, Canada.

Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE*, 8(4):e61981, 04.

Patricia Murrieta-Flores, Alistair Baron, Ian Gregory, Andrew Hardie, and Paul Rayson. in press. Automatically analysing large texts in a GIS environment: The Registrar General's reports and cholera in the nineteenth century. *Transactions in GIS*.

National Climatic Data Center. 1998. The Maury Collection: Global ship observations, 1792–1910. Online at `http://icoads.noaa.gov/maury.html`; originally on CD-ROM, Version 1.0, February 1998. [Accessed 16-Jun-2015].

National Park Service. 2015. Myth: General Ulysses S. Grant stopped the prisoner exchange, and is thus responsible for all of the suffering in Civil War prisons on both sides. `http://www.nps.gov/ande/learn/historyculture/grant-and-the-prisoner-exchange.htm`. [Accessed 5-Jun-2015].

Robert K. Nelson. 2015. Mining the dispatch. Digital Scholarship Lab, the University of Richmond. Online at `http://dsl.richmond.edu/dispatch/pages/home`. [Accessed 9-Jun-2015].

Kamal Nigam and Rayid Ghani. 2000. Understanding the behavior of co-training. In *Proceedings of KDD-2000 Workshop on Text Mining*.

Jorge Nocedal. 1980. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782.

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

Marian Olteanu, Pasin Suriyentrakorn, and Dan Moldovan. 2006. Language models and reranking for machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 150–153, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stan Openshaw. 1983. *The Modifiable Areal Unit Problem*. Geo Books.

David O'Sullivan and David J. Unwin. 2010. Point pattern analysis. In *Geographic Information Analysis*, pages 121–155. John Wiley & Sons, Inc.

Simon Overell. 2009. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. Ph.D. thesis, Imperial College London.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June. Association for Computational Linguistics.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged twitter data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, Morristown, NJ, USA. Association for Computational Linguistics.

Vivien Petras. 2004. Statistical analysis of geographic and language clues in the MARC record. Technical report for the "going places in the catalog: Improved geographical access" project, University of California at Berkeley, December 8.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

Daniele Quercia, Neal Lathia, Francesco Calabrese, Giusy Di Lorenzo, and Jon Crowcroft. 2010. Recommending social events from mobile phone location data. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 971–976, Washington, DC, USA. IEEE Computer Society.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015a. Twitter user geolocation using a unified text and network prediction model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 630–636, Beijing, China, July. Association for Computational Linguistics.

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015b. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1362–1367, Denver, Colorado, May–June. Association for Computational Linguistics.

Benjamin Ray. 2002. Salem Witch Trials documentary archive and transcription project. `http://salem.lib.virginia.edu/home.html`. [Accessed 02-Feb-2015].

Raimundo Real and Juan M Vargas. 1996. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, pages 380–385.

Lisa M. Rhody. 2012. Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1). Online at `http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/` [accessed 16-Jun-2015].

Randy Ridley, John-Henry Gross, and John Frank. 2005. Can geography rescue text search? *ArcUser*, April. Online at `http://www.esri.com/news/arcuser/0405/textsearch1of2.html`.

B. Roark, Yang Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung. 2006. Reranking for sentence boundary detection in conversational speech. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages 545–548.

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.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, November.

C. J. Rupp, Paul Rayson, Alistair Baron, Christopher Donaldson, Ian N. Gregory, Andrew Hardie, and Patricia Murrieta-Flores. 2013. Customising geoparsing and georeferencing for historical texts. In Xiaohua Hu, Tsau Young Lin, Vijay Raghavan, Benjamin W. Wah, Ricardo A. Baeza-Yates, Geoffrey Fox, Cyrus Shahabi, Matthew Smith, Qiang Yang 0001, Rayid Ghani, Wei Fan, Ronny Lempel, and Raghunath Nambiar, editors, *BigData Conference*, pages 59–62. IEEE.

Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. 2012. Finding your friends and following them to where you are. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pages 723–732.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Bjorn Sandvik. 2008. Using KML for thematic mapping. Master's thesis, The University of Edinburgh.

Anoop Sarkar. 2001. Applying co-training methods to statistical parsing. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Benjamin M. Schmidt. 2012. Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1). Online at `http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/` [accessed 16-Jun-2015].

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.

Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. 2009. Placing Flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 484–491, New York, NY, USA. ACM.

C.N. Silla Jr. and A.A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):182–196, January.

David A. Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, pages 127–136, London, UK. Springer-Verlag.

H.J. Smith, S.J. Milberg, and S.J. Burke. 1996. Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, 20(2):167–196.

Mark D. Smucker and James Allan. 2006. An investigation of Dirichlet prior smoothing's performance advantage. Technical report, University of Massachusetts, Amherst.

Michael Speriosu and Jason Baldridge. 2013. Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1466–1476. The Association for Computer Linguistics.

Michael Speriosu. 2013. *Methods and Applications of Text-Driven Toponym Resolution with Indirect Supervision*. Ph.D. thesis, University of Texas at Austin, August.

Daniel E. Sutherland. 2012. Guerrilla warfare in Virginia during the Civil War. In *Encyclopedia Virginia*. Virginia Foundation for the Humanities. [Online; accessed 5-Jun-2015].

Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. 2011. Geography of Twitter networks. *Social Networks*, August.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD '09, pages 442–457, Berlin, Heidelberg. Springer-Verlag.

Dennis Thom, Harald Bosch, and Thomas Ertl. 2012. Inverse document density: A smooth measure for location-dependent term irregularities. In *Proceedings of COLING 2012*, pages 2603–2618, Mumbai, India, December. The COLING 2012 Organizing Committee.

Edward R. Tufte. 1986. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185.

Olivier van Laere, Steven Schockaert, and Bart Dhoedt. 2013. Georeferencing flickr resources based on textual meta-data. *Information Sciences*, 238:52–74.

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.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. 2009. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1113–1120, New York, NY, USA. ACM.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 955–964, Portland, Oregon, USA, June. Association for Computational Linguistics.

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. 2011. Data-rich document geotagging using geodesic grids. Master's thesis, University of Texas at Austin.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196, Morristown, NJ, USA. Association for Computational Linguistics.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 247–256, New York, NY, USA. ACM.

May Yuan. 2010. Mapping text. In David J. Bodenhamer, John Corrigan, and Trevor M. Harris, editors, *Spatial Humanities*, pages 109–123. Indiana University Press, Bloomington, IN, USA.

Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA. ACM.

Claude Zurcher. 2013. www.notrehistoire.ch: Building a collective audiovisual memory. In *Digital Heritage International Congress (DigitalHeritage), 2013*, volume 1, pages 445–445. IEEE.