The Thesis Committee for Grant Hollis DeLozier
certifies that this is the approved version of the following thesis:

Data and Methods for Gazetteer Independent Toponym Resolution

APPROVED BY

SUPERVISING COMMITTEE:

Jason Baldridge, Supervisor

Katrin Erk

Data and Methods for Gazetteer Independent Toponym Resolution

by

Grant Hollis DeLozier, B.A., B.A.Geo.Inf.

Thesis

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

Master of Arts

The University of Texas at Austin

May 2016

Acknowledgments

This thesis would not have been possible without the generous support of numerous people. First and foremost, I thank my advisor Jason Baldridge, who had the patience and curiosity to accept a student with little to no background in his field. Whenever roadblocks were encountered in research and study, Jason could always be counted on to help me find a way out. Furthermore, I forever adopt your wise advice "Do the dumb thing first" as my own mantra. Second, I thank Ben Wing, who is both the most generous and talented programmer I've ever met. Our late nights in the lab were among the most memorable and helpful experiences in all of graduate school; none of this work would have been possible without your help. Third, I thank Katrin Erk, whose opinions and work in computational semantics inspired in ways that she doesn't know.

On a personal level, I was helped immeasurably by my parents, Danny and Holly DeLozier. Their constant support during the ups and downs of my study made everything possible. Also I thank my best friend Chase Brown, whose intellectual companionship and constant curiosity helped make the whole of graduate school feel less like work and more like fun. Lastly I thank my brother Dirk DeLozier and sister Noel DeLozier, who have always supported and entertained me with my various professional diversions.

Grant Hollis DeLozier

The University of Texas at Austin
May 2016

Abstract

Data and Methods for Gazetteer Independent Toponym Resolution

Grant Hollis DeLozier, M.A.

The University of Texas at Austin, 2016

Supervisor: Jason Baldridge

This thesis looks at the computational task of Toponym Resolution from multiple perspectives. In its common form the task requires transforming a place name–e.g. *Washington*–into some grounded representation of that place, typically a point (latitude, longitude) geometry. In recent years Toponym Resolution (TR) systems have advanced beyond heuristic techniques into more complex machine learned classifiers and impressive gains have been made. Despite these advances, a number of issues remain with the task. This thesis looks at aspects of typical TR approaches in a critical light and proposes solutions and new methods. In particular, I'm critical of

the dependence of existing approaches on gazetteer matching and under-utilization of complex geometric data types. I also outline some of the shortcomings in existing toponym corpora and detail a new corpus and annotation tool which I helped to develop.

In earlier work I explored whether TR systems could be built without dependencies on gazetteer lookups. That work, which I expand and review in this thesis, showed that competitive accuracies can be achieved without using these human curated resources. Additionally, I demonstrate through error analysis that the largest advantage of a gazetteer matching component is with ontology correction and matching, and not with disambiguation or grounding.

These new approaches are tested on pre-existing TR corpora, as well as a new corpus in a novel domain. In the process of detailing the new corpus, I remark on many challenges and design decisions that must be made in Toponym Resolution and propose a new evaluation metric.

Table of Contents

Chapter 1

Introduction

Geospatial information abounds in Natural Language, from geographically sensitive phonetic and lexical alternations to more obvious semantic contributions of place names and spatial prepositions. The sources of this geographic information are both latent and overt. Some of the information is latent: regional dialects and accents, alternating synonymous lexical items, and slang. Overt geographic information, such as place names and geospatial prepositional phrases, reference geographies directly and frequently appear in language. This thesis focuses primarily on automated methods for extracting geographic information from one form of overt geographic language: place names, aka toponyms, though other latent geographic information can help to ground more overt language. To better understand the challenges entailed by the task I begin by reviewing some theoretical aspects of toponyms before moving on to the existing computational methods for the task and its close relatives.

1.1 Semantic Characteristics of Toponyms

Toponyms are named geographic entities. The entities they denote exist at a variety of geographic scales, from the largest–Earth–down to much smaller entities–bus stops, intersections, and buildings. To date, academic literature on automated toponym resolution does not go to scales smaller than neighborhood, though commercial geocoders (e.g. Google Maps) attempt this to some degree. Because of difficulties obtaining annotated data for small scale entities, most of the examples and literature in this paper focus on entities at the level of neighborhood or higher,

though recent work has attempted to build such small scale corpora (Matsuda et al., 2015).

Within formal semantics, the meaning of place names has been subsumed into a general theory of proper names. Drawing chiefly from Kripke (1980) it is common for semanticists to argue that named entities are not descriptions of entities but rather they function as a direct link between a name and its particular in-the-world referent (Kamp and Reyle, 1993). In this theory outlined by Kripke, there is said to be some initial 'naming event' wherein an entity is given a proper name. All subsequent uses then chain back to this original event, wherein the entity was directly referred. This story is significant to note for TR work because it points to an important property of toponyms, namely that they exist as fiat objects: the only limiting factor on the number of unique toponyms is our willingness to name them (or perhaps our *need* to name them).

It should come as no surprise then that there is an extremely large number of unique toponyms that systems can be asked to resolve. This nearly limitless capacity for naming places however does not mesh well with how most predominant systems resolve the meaning of toponyms. All named entity disambiguation systems rely on curated knowledge bases, which range in their degree of structure. At a lower level these resources take the form of a semi-structured encyclopedic repository (e.g. Wikipedia). However, resources such a wikipedia lack enough structure for classical entity classification procedures, thus researchers commonly rely on entity gazetteers or databases such as DBPedia and YAGO (Hoffart et al., 2011). All of these entity gazetteers source their geographic information from GeoNames, a resource which dictates only 9 million unique geographic entities. In addition to lacking in number, its place names are heavily biased towards North America and Europe (Graham and De Sabbata, 2016).

1.1.1 Ambiguity and Toponyms

Given the limitations of geographic gazetteers it is difficult to truly know how ambiguous toponym strings are. Nevertheless, GeoNames gives the opportunity to estimate this, at least in some limited fashion. GeoNames contains 8,943,067 unique geographic entities which are represented in language by 11,171,842 unique strings. The average entity ambiguity of these strings is 1.98. This at first seems low, how-

ever prominent place names tend to be much more ambiguous in the gazetteer (e.g. 'Washington' is ambiguous over 68 entities).

1.1.2 Representing Geographic Reference

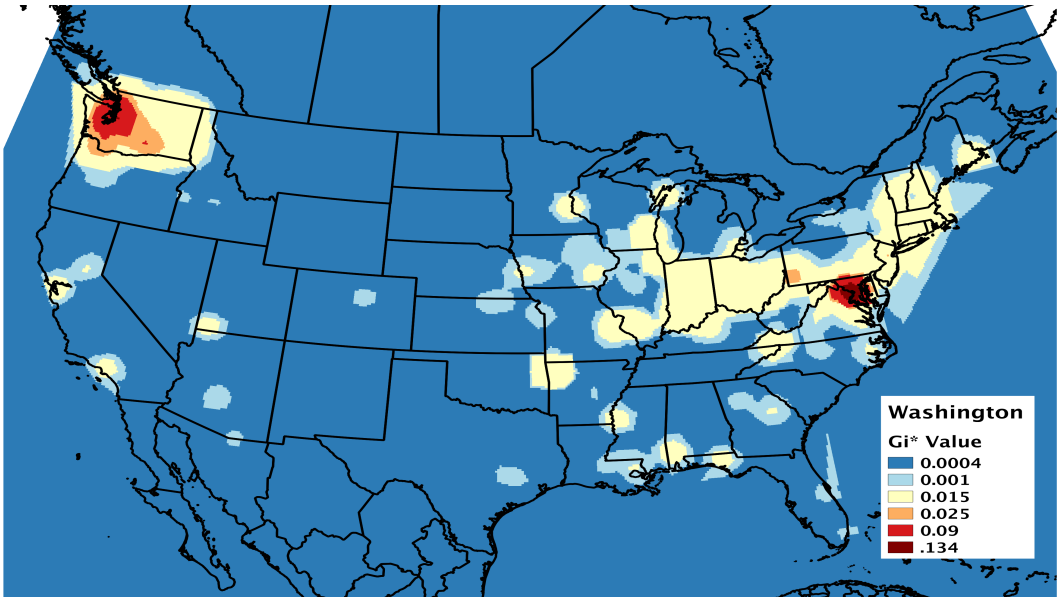
It is difficult to develop geographic representations of toponyms, principally due to the issue of spatial vagueness. Most toponyms do not have formal boundaries that describe them, and those which do have a formalized legal boundary (city limit boundaries, county, and state boundaries) often have socially realized boundaries that are different from their legal boundaries. For example, consider a representation of Austin, TX. While most Austinites will agree on the core of what this refers to in some abstract sense, people will disagree as to the specific borders (e.g. I've heard on multiple occasions that 'Austin' is north of Ben White and south of Highway 183, while the city boundaries go significantly further north and south). Crucial then to developing representations of geographic entities is dealing with this inherent fuzziness of reference (Montello et al., 2003).

In practice, work in toponym resolution has opted for highly simplified point based representations of toponyms. This tendency has had noticeable downsides, particularly with respect to evaluation processes concerning natively discontinuous polygon geometries and line geometries. For example, in this framework countries like the United Kingdom become represented by a point in the Irish Sea and rivers typically become represented by a point at a mouth or head of the river. The use of such representations, particularly in evaluation corpora, becomes problematic as systems can produce geometrically correct disambiguations yet fail to achieve a 'correct' response.

One way of dealing with the spatial vagueness of geographic entities is to opt for a distributional representation. Distributional geographic representations, which I (DeLozier et al., 2015) and others (Wing and Baldridge, 2014) explore, describe the meaning of spatial entities to be distributed over the space which they are observed. Practically, the geographic meaning of words becomes approximated with the distribution of the usage of the words over geographic space. For a visual rendition of 'Washington' in this framework, see Figure 1.1.

In nice ways this geographically distributional approach to representing place names coheres with earlier observations made by formal semanticists, particularly

Figure 1.1: Local G_i* values for *Washington*.



in its ability to fiat new names into existence. In other ways it struggles to cohere; Kripke's theory seems to demand that a particular named entity's referent not change once it is willed into existence, while distributional approaches suggest constant fluctuations in the geographic meaning of toponyms (e.g. *Austin* would not be observed in the same places in 1990 as 2016). While it doesn't naturally adhere to the abstractly anchored referent suggested by semanticists, there may be means of integrating the concepts; later in this thesis I utilize a 'snapping' methodology to try and integrate my distributional approach with an anchored ontology.

1.2 Tasks involving Geographic Information and Text

Most work involving Geographic Information and Text can be grouped within two domains: Toponym Resolution and Document Geolocation. Toponym resolution involves assigning geographic reference to individual, potentially ambiguous toponyms, while document geolocation assigns geographic reference to larger spans of text (documents, broadly construed).

1.2.1 Toponym Resolution

There are a number of difficulties associated with toponym resolution, but they key difficulties involve ambiguity and paucity of training corpora. As an example of the issues surrounding ambiguity, consider the toponym *Springfield*: dominant place name gazetteers dictate at least 236 unique senses of the term (and these underestimate the true total), with possible references spanning the globe. TR systems must choose referents in these highly ambiguous scenarios, even when correct referents are not listed in gazetteers.

The second key difficulty—paucity of training corpora—is one that is shared among document geolocation and toponym resolution. All existing training corpora in both domains fixate around very narrow ranges of geographic entities. One major corpus used in toponym resolution for example, TR-CoNLL Leidner (2008), has only 800 unique toponym referents while gazetteers such as GeoNames list over 8 million places (and these resources greatly underestimate the true number of toponyms). Such mismatches do more than underscore the need for larger and domain diverse corpora, they point to fundamental issues associated with learning to resolve geographies from language. Dealing with this paucity is a challenge for all geolocation systems and many systems attempt to alleviate it by splicing corpora with latent annotations inferred from a more general resource like Wikipedia (Speriosu and Baldridge, 2013; Santos et al., 2014; DeLozier et al., 2015)

Many other issues associated with how one defines Toponym Resolution as a task can affect how one creates a corpus. Metonymy–the ability of a place name to refer to something closely related to a place (e.g. a government)–is one such issue. All existing TR corpora include metonymic uses of place names. Demonymy–names for the people who inhabit an area (e.g. Americans)–is another such issue. The Local Global Lexicon (LGL) corpus (Lieberman and Samet, 2012) includes such terms as toponyms and georeferences them, while all other corpora do not. An additional issue pertains to the range of entity types a system is expected to resolve; many corpora limit their expectations to larger entities (e.g. TR-CoNLL is limited to cities, states, and countries), while others focus more on highly local entities (e.g. bus stops) (Matsuda et al., 2015). The last issue relates to whether systems ought resolve places which are embedded inside other named entities. For example, the LGL corpus expects *New York* in the expression *New York Times*, to be resolved

to the state of New York.

1.2.2 Document Geolocation

The first Geolocation tasks to use machine learning arose in the related task of document geolocation (Backstrom et al., 2008). The conceptual goal of the task is not straightforwardly defined, rather it depends on the specific application. For example, in Twitter-based Document Geolocation tasks, the goal is typically to geolocate a twitter user, given a sampling of their messages (Wing and Baldridge, 2011). When geolocating images, the goal is to find the location at which an image was taken using accompanying text as aids (O'Hare and Murdock, 2013; van Laere et al., 2013). However in tasks such as geolocating Wikipedia Pages and Civil War Correspondences (Cite Ben's dissertation), the goal is often more abstract–to represent the geographic topic or summary of some text.

The goal of the task is thus somewhat different than Toponym Resolution in that it is not directed at any linguistically definable unit of speech. This yields both advantages and disadvantages for researchers seeking to enrich their texts with geographic information. On the one hand such loosely defined systems can quickly generate geographic information in a wider variety of contexts, even in contexts where place names are not given at all. On the other hand it is less clear how such information actually connects with the particular utterances and content of the text. For example, a single news article may discuss events that are widely dispersed in geographic space and cover several semantic topics. The inability of document geolocation systems to situate their geo-references in a specific syntactic and semantic organizational scheme can make untying geographically complicated documents very difficult.

1.3 Applications

Toponym resolution has far reaching applications, with uses in question answering and information retrieval tasks Leidner (2008); Daoud and Huang (2013), automated geographic wayfinding, and social and historical research Smith and Crane (2001); Grover et al. (2010); Nesbit (2013). Commercial Toponym Resolvers such

as GeoParser, Clavin, and FinchText utilize the technology to aid in geographic summarization of news content and defense-oriented intelligence gathering.

Chapter 2

Data

In this chapter I describe existing corpora for Toponym Resolution before detailing work I did to create a novel toponym resolution corpus in a historical domain.

To date very few corpora exist for text geolocation tasks, and those which do exist have flaws or are very small in size. This is particularly true for tasks seeking to do geolocation work with historical texts. In the realm of document geolocation, there exist no historical corpora whatsoever; in the realm of toponym resolution historical corpora exist, but are flawed in important respects (Speriosu and Baldridge, 2013; DeLozier et al., 2015).

2.1 Existing TR Data

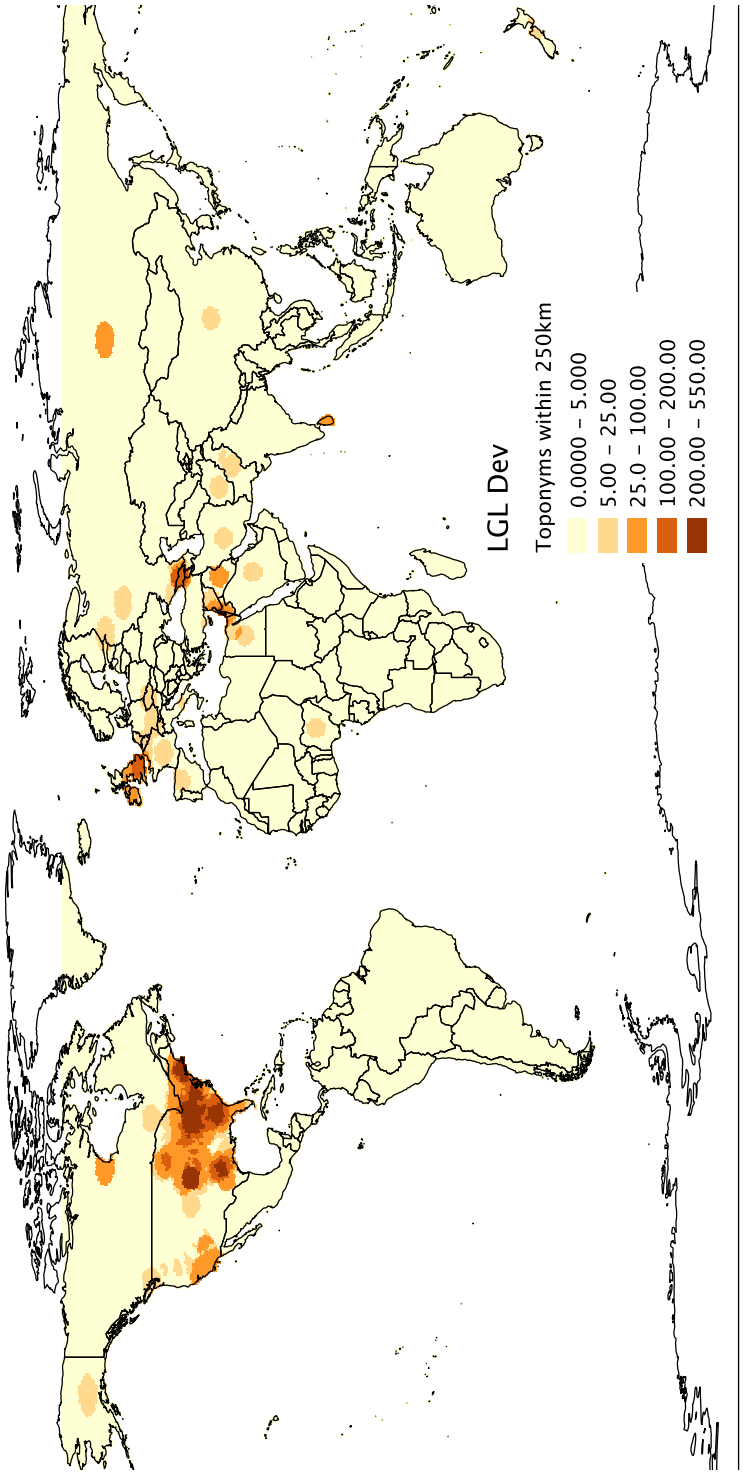
Many of the characteristics of existing TR corpora are summarized in Table 2.1.

2.1.1 LGL

The Local-Global Lexicon corpus (**LGL**) was developed by Lieberman et al. (2010) to evaluate TR systems on geographically localized text domains. It is widely used in current toponym resolution research (Lieberman and Samet, 2012; Santos et al., 2014; DeLozier et al., 2015). LGL consists of 588 news articles across 78 sources. The sources were selected purposefully to highlight less dominant senses of common places names; e.g., some articles are from the Paris News (Texas) and the Paris Post-Intelligencer (Tennesee). LGL contains 5,088 toponyms, which are mostly small populated places, although a significant number of locales, counties, states, and

countries appear. LGL also has important differences in how annotations were done compared to related datasets. Demonyms (e.g. *Canadian* and *Iranian*) are marked as toponyms and annotated with latitude-longitude pairs throughout the corpus. Also, organization names that contain place names are marked solely as toponyms; e.g., *Woodstock* is marked as a toponym even when it is in the larger phrase *Woodstock General Hospital* and *London* is marked as a toponym in *Financial Times of London*. While nested named entities have been recognized as an important problem in NER system design and evaluation (Finkel and Manning, 2009), using inner-most entities is unconventional in the context of other Toponym Resolution work. The geographic spread of the toponyms in LGL dev can be seen in the density map in Figure 2.1.

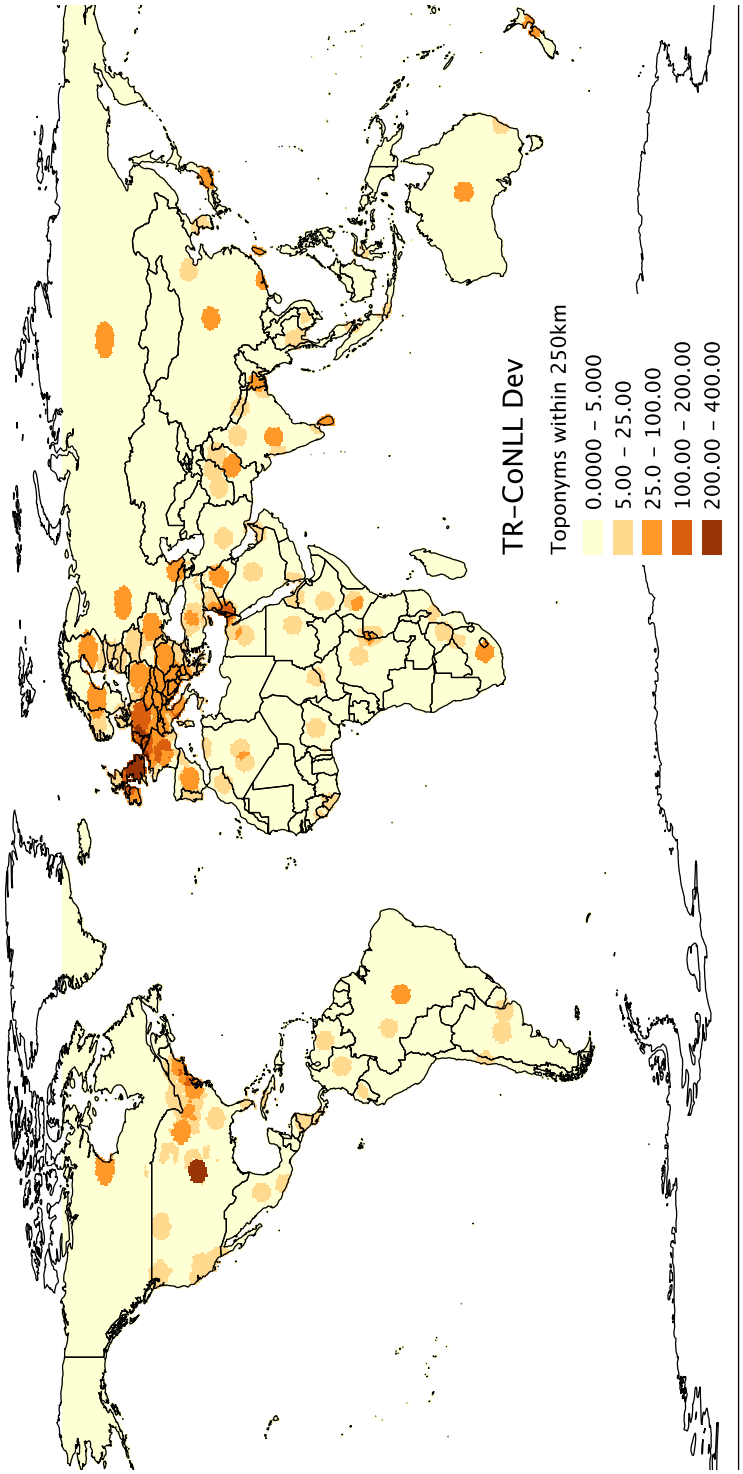
Figure 2.1: LGL Dev Toponym Density



2.1.2 TR-CoNLL

TR-CoNLL has been used in several papers, and was the first proper toponym resolution corpus devloped Leidner (2008); Speriosu and Baldridge (2013); DeLozier et al. (2015). TR-CoNLL was constituted from the CoNLL-2003 (Conference on Natural Language Learning, Shared task 2003 on Named Entity Recognition) and consists of roughly 1,000 Reuter's international news articles and identifies 6000 toponyms in 200,000 tokens. Place names in the dataset were hand-annotated with latitude-longitude coordinates. The feature types in this corpus are: countries, states, and cities. Reference is represented as a point, and is not linked to a knowledge base or structured ontology. The domain is about 1000 Reuters international news articles. This is one of the easier toponym corpora in that a large proportion of the toponyms it contains are that of countries and prominent places. Within the corpus there is also a low level of ambiguity among toponym strings (hence population heuristics perform very well Speriosu and Baldridge (2013)). TR-CoNLL was split by Speriosu and Baldridge (2013) into a dev (4,356 Toponyms) and a held-out test set (1,903 Toponyms).The geographic spread of the toponyms can be seen from the density map in Figure 2.2. Speriosu (2013) observed many annotation errors in the corpus and attempted to provide corrections prior to his evaluations. Informally, I utilized this 'fixed' verson of TR-CoNLL in my own experiments but found a large number of lingering annotation errors. Some of these errors I detail in chapter 4.

Figure 2.2: TR-CoNLL Dev Toponym Density



2.1.3 CWar

CWar is the Perseus Civil War and 19th Century American Collection, which consists of 341 books (58 million tokens) printed around the time of the United States Civil War. Tuft's university digitized and OCR corrected much of the collection, then place names were annotated with single latitude-longitude pairs using a combination of manual annotation with off-the-shelf toponym resolvers and named entity recognizers. We use the same split of CWar as Speriosu and Baldridge (2013): dev (157,000 toponyms) and test (85,000 toponyms). It is an interesting dataset for TR evaluation because it is a different domain than contemporary news articles. It also contains a larger proportion of more localized (less populous) place names and is much less geographically dispersed than TR-ConLL. Unfortunately, numerous issues exist with the named entity annotations in the corpus (Speriosu (2013) gives details) so it is only appropriate for evaluating using oracle toponyms, but not those identified by a named entity recognizer. The basic problem with using CWar is that only a subsection of place names were annotated (i.e. it lacks gold NER annotations for place names). Because of this, it is impossible to do a proper evaluation of TR systems as they would appear in the wild. Later, when we evaluate a number of baseline systems on this corpus we purposefully only include systems evaluated using 'oracle' NER. In addition to problems with the range of place names annotated, the toponym annotations themselves were done using a computer assisted approach. I speculate that this is how the corpus obtains its size, and it does so at the cost of quality. Toponyms which lack gazetteer entries are simply not annotated, greatly restricting the scope of the dataset. This undermines one of the core reasons for using the dataset in the first place (frequency of less popular, 'small' places).

2.1.4 LRE

The Location Referring Expression (LRE) corpus developed by (Matsuda et al., 2015) is the newest of the preexisting toponym corpora available. Due to its age, and perhaps its language domain (Japanese), it has not yet been utilized in the toponym resolution literature. They develop the corpus with two main goals in mind: (1) annotate highly local geographic features which they term 'facilities' and (2) do so within the social media domain of Twitter. The corpus consists of 951 toponyms across 10,000 tweets. LRE adheres to a point based reference ontology

Table 2.1: Toponym Corpora

Corpus	Domain	Entity Types	Reference Types	Metonyms	Demonyms	Nested NE	Toponyms
TR-CoNLL	Contemporary International News	Cities,States, Countries	Point only	Yes	No	Most Encompassing NE	5000
LGL	Contemporary Local Newspapers	Few Locales, cities, states, countries	Point only	Yes	Yes	Annotates Embedded Places	5088
LRE	Tweets from Japan	Highly local 'facilities' and above	Point only	?	No	?	951
WOTR	US Civil War Letters + Reports	Locales, Cities, and States	Point and Polygon	No	No	Most Encompassing NE	10380

which they hand build, with most geographic entities being highly local features (e.g. bus stops). The corpus is novel in that it contains a wide range of geographic entities, but it is very small compared to related corpora.

2.2 War of the Rebellion Corpus

In order to address some of the shortcomings of existing TR corpora, I helped build a novel TR corpus called WoTR-Topo (War of the Rebellion - Toponyms). The corpus was designed to be novel in its (1) domain, (2) size, (3) richness and depth of geographic annotation. This section describes the annotation procedure that was followed, as well as summarizes aspects of the data.

The War of the Rebellion corpus is a large set of United States Civil War archives, published in 128 books (which were broken into 70 volumes and 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.[1] The Ohio State University version of the corpus which we utilize was digitized and OCR corrected, making

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

	Topo subset	Full data
Total tokens	447,703	57,557,037
# volumes	15	126
# documents	1,644	254,744
Avg. tokens/document	272.32	225.94
Avg. toponyms/document	7.17	NA
Toponyms	11,795	NA
Toponyms with geometries	10,380 (88%)	NA
Toponyms with points	8,130 (69%)	NA
Toponyms with polygons	2,296 (19%)	NA
People	7,994	NA
Organizations	2,591	NA

Table 2.2: Statistics on WOTR, annotated subset and full data.

it a high quality digital version of the archive `http://ehistory.osu.edu/books/official-records`. This section describes the process of annotating 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).

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

The scope of the annotation process is given in Table 2.2. The toponym annotation process, which spanned 4 months, resulted in the annotation of 11,795 toponyms spanning 1,644 annotated documents across 15 volumes. Originally all annotations were done by a single annotator. After this process all of the original annotations were reviewed by a second team of three annotators. These annotators

were asked to correct a number of problems with the annotations that were not realized until after the initial annotation process had finalized. Mostly these corrections were focused on (1) conjunctive toponyms, (2) possessive toponyms, (3) difficult to find toponyms, (4) geo-reference for rivers, (5) vague region toponyms, (6) gray areas between proper names and referring expressions, (7) gray areas regarding nested named entities. Also in this process of correction, reviewers discovered a large number of errors in the NER which were not properly identified and corrected by the initial annotator. The review process resulted in about a 20% increase in the total number of toponyms in the corpus. Details on the challenges of annotation in this dataset are given later.

2.2.1 Toponym annotation guidelines and challenges

Annotators were asked to quickly scan the documents and look for place names. Place names which were not detected by Stanford NER were asked to be added, and other entities which were incorrectly classified as places were deleted. These annotations were done inside of a HTML/JavaScript interface which allowed annotators to highlight spans of text, then add reference directly through a simple web map. We directed annotators to include point, multi-point, polygon, and multi-polygon geometries where appropriate by drawing the reference directly onto a Google Streets web map embedded into the interface via the JavaScript library OpenLayers`http://openlayers.org/`. A screen shot of the tool is shown in Figure 2.3.

They key guidelines annotators were given for the task concerned three aspects of toponyms: metonymy, demonymy, and nested named entities.

Metonymy refers to a phenomenon wherein a place name comes to refer to a non-place entity that is closely associated with a place (e.g. a government). *Washington* in the sentence *Washington passed several laws this term* is an example of a metonymic place name. Metonyms are seen most commonly in WOTR as country and state names. Annotators were asked to exclude metonymic names from annotation.

Demonyms are terms that refer to the people who live or come from a certain area (e.g. *Iranians*). They are typically not annotated as toponyms, except for in LGL. We follow most of the toponym literature in asking our annotators to not

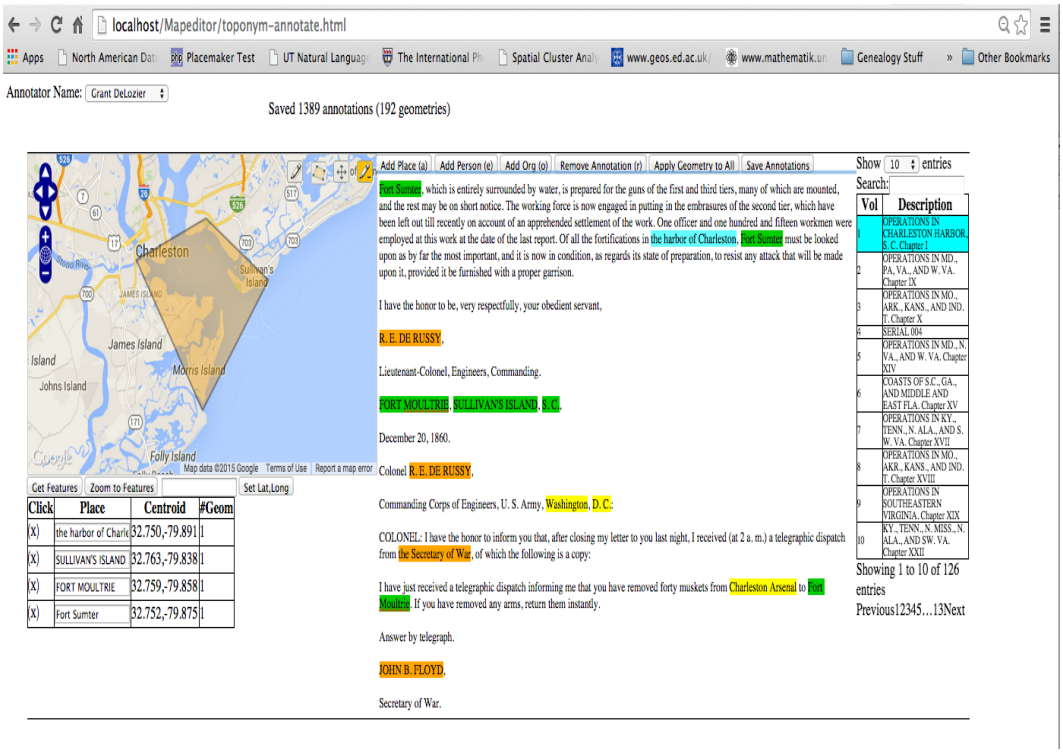


Figure 2.3: Screen shot of the toponym annotation tool. Place names highlighted in yellow, place names with geoemetries in green.

annotate and reference these entities.

Nested named entities were addressed by Finkel and Manning (2009). Researchers have typically adopted the stance of annotating the most encompassing named entity, though there are exceptions to this trend as is the case in the LGL corpus. We ask annotators to only mark toponyms which constitute the most encompassing named entity (e.g. in *44th Virginia Cavalry*, Virginia is not marked, as the larger encompassing entity is an organization). Not included among nested named entities are toponym hierarchies, or disambiguators such as in the phrase *Richmond, VA, CSA*. In these cases each toponym is annotated with separate reference.

To find the reference of places, annotators were allowed access to Internet search. Annotators were encouraged to look up troublesome toponyms with helpful relevant keywords, such as *Civil War* or the region or commander mentioned of the larger document context. Post-hoc review of the resources annotators used revealed

that Wikipedia was the most dominant resource used, though Google Maps and niche US Civil War websites were used as well.

A number of challenges were encountered during the annotation process:

(1) conjunctive toponyms, (2) possessive toponyms, (3) difficult to find toponyms, (4) geo-reference for rivers, (5) vague region toponyms, (6) gray areas between proper names and referring expressions, (7) gray areas regarding nested named entities.

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

2. Possessive toponyms, or toponyms partially constituted of a person's name, appeared in the corpus (e.g. *Widow Harrow's house*). Originally, we asked annotators to avoid annotating these as toponyms, and instead merely annotate the embedded person as a person. This guideline was complicated in many examples such as *Lovejoy's Station [railroad station]* or *Varnell's Station* where the possessed enity was also capitalized. We amended our guidelines to ask annotators to mark fully capitalized, possessed entities as toponyms.

3. Difficult Toponyms, or toponyms that could not be geographically referenced, made up about 12% of the overall toponyms in Wotr-Topo. This was typical of toponyms that described the locations of ferries, bridges, railroads, and mills. These features usually no longer exist, so discovering their exact reference even with access to Google is very difficult. These appear in the corpus as toponym entities without geographic reference.

4. Rivers, and physical features more generally, are difficult to reference geographically because their geometric definitions are often highly complex, vague, and poorly defined in gazetteers. Rather than ask annotators to annotate the full extent of rivers, we asked them to mark a point on the river that

they felt was most relevant to the context. Annotators tended however to opt for whichever point the river's Wikipedia page indicated, though this was not always the case.

5. Geographically vague toponym regions appear in the texts. Some of the common examples appearing in the text are *the North*, *the South*, *the West*, and *Northern Mississippi*. We asked annotators to mark these as toponyms, and attempt to draw their reference given the context. Mostly, annotators did not feel comfortable annotating such entities, so they appear as toponyms without defined geographic references.

6. We asked annotators not to annotate referring expressions (e.g. *the stone bridge*), yet we failed to anticipate referring expressions which were partially constituted of place names (e.g. *the Dalton road*). Given that these expressions contain proper place names, and are places themselves, we decided to ask them to try and annotate them as toponyms. Such expressions were common ways of indicating the names of roads, prior to more well developed highway systems (i.e. they describe the road that takes one from the current place to the noted place. Sometimes this appears as two places as in *the Decatur and Atlanta road*). Annotators tended to mark the location of such entities as a point near one of the embedded city toponyms.

7. We gave our annotators a rule to only annotate the entity type of the *most-encompassing* named entity. Using this rule expressions like *44th Virginia Cavalry* became annotated as one single organization, rather than a place inside an organization. We did not anticipate however the range of semantically equivalent expressions such as *44th Cavalry of Virginia* or *44th Cavalry from Virginia*.

2.2.2 WoTR Summary

Summary of the important aspects of the corpus I helped to create are given in Table 2.1 and Table 2.2. From the perspective of toponym resolution, the War of the Rebellion Corpus is innovative in many respects: richness of geometric annotation (annotations with multi-point, polygon geometries), corpus size (wotr-topo

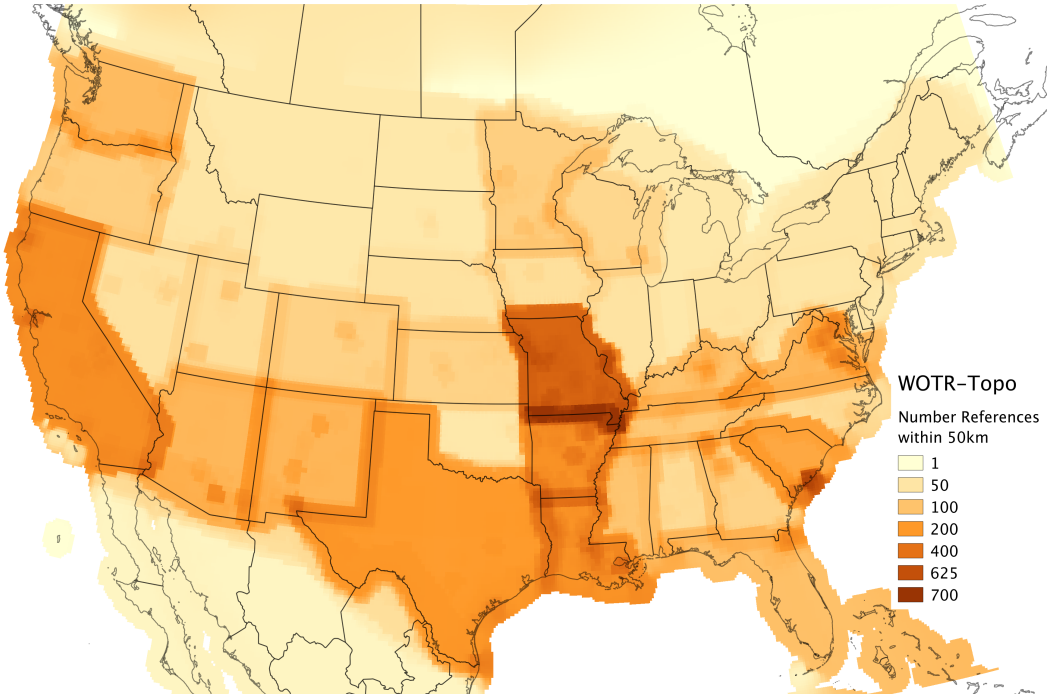


Figure 2.4: Distribution of Toponyms in WOTR-Topo

has roughly 2X the toponyms of other corpora), and place names not in gazetteers. Baseline system resolution results (given in chapter 3) indicate that the corpus is the most difficult of the corpora surveyed, with Accuracy @ 161 km scores–and especially NER inclusive scores–being significantly lower than the next most difficult coprus, LGL.

The geographic distribution of the annotations is given in Figure 2.4. As can be seen in the map, the annotated volumes skew towards western theatres of the war, though events in Virginia and South Carolina are healthily represented. State names are among the most frequent toponyms in the dataset, so states tend to be outlined in the distribution.

Chapter 3

Gazetteer Independent Toponym Resolution

This chapter addresses a weakness of prior toponym resolution work: explicit reliance on curated knowledge resources such as gazetteers. These are highly incomplete resources that depict only narrow portions of the total set of place names. To reduce dependence on them, we rely on recent advances in the related task of document geolocation, where the goal is to predict the geographic context of a much larger span of text. Much of this work has been directed at guessing social media users' locations given only their observed language (Cheng et al., 2010; Eisenstein et al., 2010; Wing and Baldridge, 2011; Roller et al., 2012; Wing and Baldridge, 2014). The success of these approaches is generally on a much coarser geographic scale than is required by TR systems, but the approaches used are applicable to TR. Crucially important to our work are geographically situated language models. Geographic language models describe the probability of observing certain words at different places on the earth and capture not just explicitly geographic words like *Philadelphia* and *Midwest*, but also over latently geographic words such as *y'all* and *hockey*. In the gazetteer-independent method I've helped to develop, we take these geographic language models, spatially smooth and geographically cluster them, and use them to form the core of a toponym resolution process.

Only limited attempts have been made to use local geographic clustering techniques in the context of text-based geographic disambiguation. Cheng et al. (2010) derive information analogous to local geographic clusters for words to geo-

reference Twitter users. Following work by Backstrom et al. (2008) on determining the geographic focus of queries, they identify a subset of words with a prominent geographic center (characterized by large probability of the word occurring at a location) and steep decay of the probability over distance from that center. This approach does find many geographically indicative words, but it makes assumptions about their distributions that are not ideal for toponym resolution. In particular, they assume that geographic words have well-defined centers and highly peaked distributions. Many toponyms–which intuitively should be the most helpful words for geographic disambiguation–lack such distributions. Instead, many toponyms are widely dispersed over distance (e.g. toponyms that describe large geographic spaces like countries lack steeply peaked centers) or have multiple prominent geographic centers. Figure 3.1 gives an example of how the toponym *Washington* is characterized via multiple prominent geographic clusters.

In Topocluster DeLozier et al. (2015) use the profiles of these local clusters to build a system that grounds toponyms by finding areas of overlap in the distributions of toponyms and other words in a toponym's context. We also demonstrate that such a system can operate well without the aid of gazetteer resources and extensive metadata, and as a result, it performs better than gazetteer-bound methods on toponyms found by a named-entity recognizer.

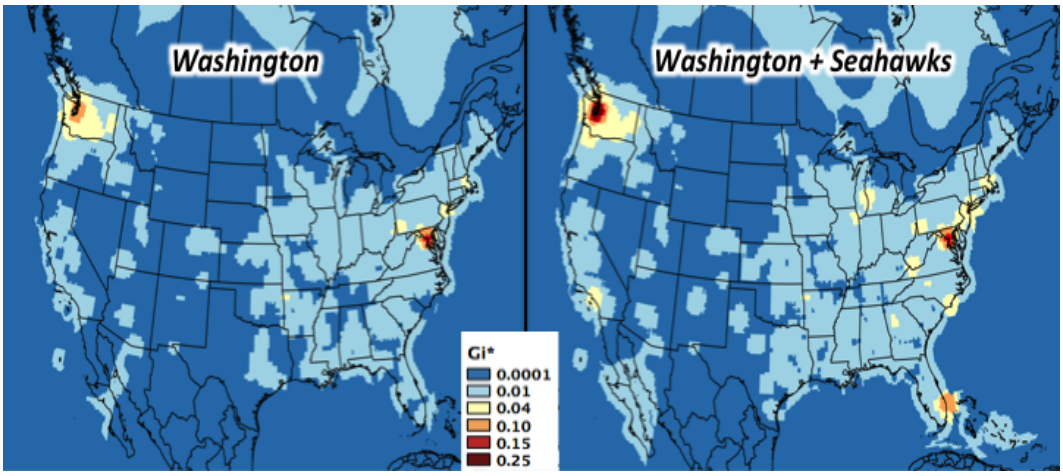
3.1 TopoCluster

The key insight of language modeling approaches to geolocation is that many words are strong indicators of location, and these tend to surface in regionally specific models. However, rarely is any attempt made to determine the specific spatial strength of a given word. Our approach, TopoCluster, derives a smoothed geographic likelihood for each word in the vocabulary and then finds points of strongest overlap for a toponym and the words surrounding it—effectively merging the shared geographic preferences of all words in the context of a toponym, including the toponym itself.

Consider a very ambiguous toponym like *Hyde Park*. The standard view asks what the probability of a given location is given the context, using a set of models per location. Various models of this kind have been proposed, including generative models for geolocation, possibly with feature selection Speriosu and Baldridge (2013). TopoCluster in contrast employs an indirect relation between a target and

its context by appealing to a shared relation in geographic space. Crucially, that geographic space is defined by how tightly all the words in the vocabulary tie themselves to local regions—effectively doing selection of geographically releveant features in the determination of a given location. One effect of this is that even in situations where *Hyde Park* does not appear at all in training, our system can guess a referent given the geographic clusters associated with known context words like *Austin* or *Texas*.

Figure 3.1: Left: Local G_i* values for *Washington*. Right: interpolated G_i* values for *Washington + Seahawks*.



The above motivation requires identifying geographic clusters for every word. We derive these by applying local spatial statistics over large numbers of georeferenced language models. Disambiguation is performed by overlaying the geographic clusters for all words in a toponym's context and selecting the strongest overlapping point in the distribution. Gazetteer matching can optionally be done by finding the gazetteer entry that is closest to the most overlapped point and matches the toponym string being evaluated.

Local spatial statistics have long been used to derive hot spots in geographic distributions of variables. TopoCluster uses the Local Getis-Ord G_i* statistic to measure the strength of association between words and a geographic space Ord and Getis (1995). Local G_i* measures the global proportion of an attribute that is observed in a local kernel.

$$G_i{}^*(x) = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{n} x_j} \tag{3.1}$$

Each x_j is a measure of the strength of the variable x at location j and w_{ij} is a kernel defining the weight (similarity) between locations i and j. For x_j, we use an unsmoothed local language model as the strength of a word x in each geolocated document D. In addition to single-token words being in the unigram model, multi-token named entities are included. These were derived from Stanford NER's 3-class CRF model Finkel et al. (2005).

We use an Epanichnikov kernel Ord and Getis (1995); O'Sullivan and Unwin (2010) and a distance threshold of 100 km to define the weight w_{ij} between a grid point i and a document location j.

$$w_{ij} = .75(1 - (\frac{dist(i,j)}{100km})^2)_{\{dist(i,j) \leq 100km\}} \tag{3.2}$$

This weights the importance of particular documents at locations j to their near grid points i. When used in the $G_i*$of equation 3.1, the kernel has the effect of smoothing the contributions from each document according to their nearness to i, the current cell under consideration.

The output of the local statistic calculations is a matrix of statistics with grid cells as columns and each word as a row vector $\vec{g^*}(x)$. The G_i* statistic serves primarily to create a geographically aggregated and smoothed likelihood of seeing each word at certain points in geographic space.

In practice the G_i* statistic can be run directly from the points in the observed documents, or it can be calculated from points in a regularized grid. We use the the latter to reduce the computational cost of calculating G_i* for all words. A grid of $.5^o$ geographic degree spaced points was created, beginning with a point at latitude of 70^o N proceeding down to -70^o S. The grid was clipped to only include points within $.25^o$ of land mass. In total, the grid used for this study represents 60,326 unique points on the earth.

Because more prominent senses of a place name are represented in more documents, clustering based on regional language models derived from a source

like Wikipedia is likely to show preferences for prominent senses of a place name without being overly tied to a specific aspect of a place (e.g. administrative level or population). This is seen in the interpolated heat map of the local G_i* clusters for *Washington* in Figure 3.1. *Washington* has strong clusters around Washington state and Washington DC, with a slight preference toward the latter in an empty context. However, this preference changes in contexts favorable towards other senses (e.g. *Seahawks* in the context shifts towards the state referent). TopoCluster code and precomputed local statistic calculations are available online [1].

Domain adaptation: Because the local G_i* statistic is bounded between 0 and 1, it is straightforward to adapt it to new domains and new data with a simple linear interpolation of values derived from different corpora.

$$\vec{g^*} = \lambda \, \vec{g^*}_{InDomain} + (1-\lambda) \, \vec{g^*}_{GeoWiki} \tag{3.3}$$

We run several experiments to test the importance of domain adapting G_i* values. For each corpus (TR-CoNLL, CWar, and LGL), we construct pseudo-documents from its development set by converting each toponym and the 15 words to each side of it into a document. Each pseudo-document is labeled with the latitude-longitude pair of the corpus annotation for the toponym, which allows us to train domain-specific regional unigram language models.

Resolution: To disambiguate a toponym z, we separate the toponyms t from non-toponym words x in that z's context window c (size of window is optimized differently for each domain, and function words are filtered). We then compute a weighted sum of all the $\vec{g^*}$ values of the toponyms t and words x in c. For metrics deriving Accuracy @ 161km, mean, and median distance toponyms are sourced from oracle recognition, or directly from the annotation source. Precision, Recall and F-1 Score are given for appropriate datasets when off-the-shelf Named Entity Recognition systems are used as a prerequisite to the Toponym Resolution process.

$$g^*(z,c) = \theta_1 \vec{g^*}(z) + \theta_2 \sum_{t \in c} \vec{g^*}(t) + \theta_3 \sum_{x \in c} \vec{g^*}(x) \tag{3.4}$$

The parameters θ_1, θ_2, and θ_3 weight the contribution of the main toponym, other toponyms and the generic words, respectively. The chosen location is then the grid

[1] `https://github.com/grantdelozier/TopoCluster`

cell i with the largest value in $g^*(z, c)$, which represents the most strongly overlapped point in the grid given all words in the context. Weights are decided on a per domain basis, based on a training procedure described in section on toponym weighting.

TopoClusterGaz: The output of the above disambiguation process is gazetteer-free: a single point i representing a cell in the grid is produced. However, we can restrict the prediction to a gazetteer by forcing place names to match title case and primary names, alternate names, and 2-3 letter abbreviations contained in our Geonames-Natural Earth hybrid gazetteer. Reference for the toponym is snapped to the gazetteer entry that matches the term in one of these fields and has geometry closest to the most overlapped point i.

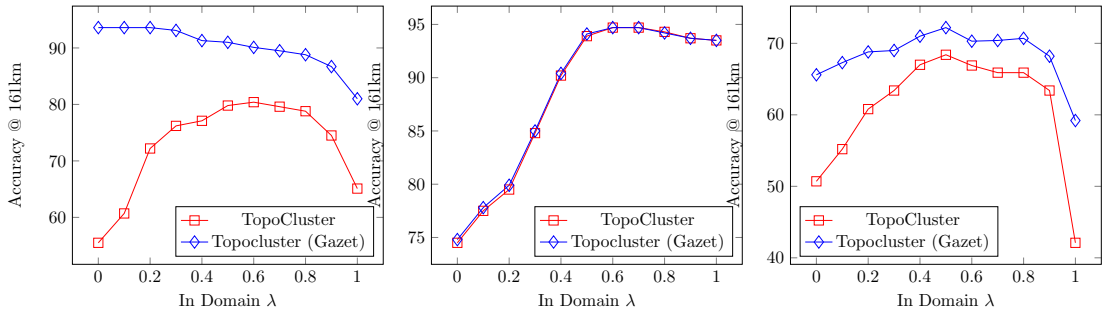
Self-training using metadata: Documents often contain metadata that is useful for toponym resolution Lieberman and Samet (2012). One such feature is domain locality, wherein certain geographies are weighted according to their correspondence with the spatial extent of a document's intended audience. This typically requires explicit correspondence with such metadata at test time (e.g. 'publisher' or 'domain') and also requires additional training annotations corresponding to an oracle geolocation for a publication's place of focus. As such, they do not easily generalize to all use cases; nonetheless, their usefulness is naturally of interest, particularly in very localized datasets such as LGL.

We explore use of a very limited domain locality feature through a self-training procedure which only uses the name of the publication at train and test time. For every publisher (e.g. theparisnews.com, dallasnews.com), TopoClusterGaz is run on all documents. The predictions are then filtered to only include references to countries, regions, states, and counties. This filtered set of toponyms is then associated with the publication domain. Later, when toponyms in the respective local domains are disambiguated, our system injects the domain's associated country, region, state, and county toponyms, applying the θ_2 weight used with other toponyms in the text context. In this way, we use very little manually specified knowledge to bootstrap and exploit a characterization of the domain.

3.2 Experimental Setup

We consider both **TopoCluster** and **TopoClusterGaz** (which uses a gazetteer), and we compare using domain adaptation (λ>0) or not (λ=0). These are compared

Figure 3.2: Domain adaptation: optimizing λ for each corpus. Left: TR-CoNLL, Middle: CWar, Right: LGL.



to two gazetteer-based baselines: **Random**, which randomly selects an entry from the possible referents for each toponym, and **Population**, which selects the entry with the greatest population (according to the gazetteer). We also compare to six of the systems of Speriosu and Baldridge (2013):

- **SPIDER**: a weighted *spatial minimization* algorithm that selects referents for all toponyms in a given text span.
- **TRIPDL**: a document geolocation model that predicts the probability of a referent being in a grid cell given a document, restricted to valid cells according to a gazetteer.
- **WISTR**: a discriminative classifier per toponym, trained using distant supervision from GeoWiki.
- **TRAWL**: a hybrid of WISTR and TRIPDL with preferences for administratively prominent locations.
- **WISTR+SPIDER** and **TRAWL+SPIDER**: two combinations of spatial minimization with classification.

Other systems of interest include Santos et al. (2014) and Lieberman and Samet (2012). However, direct comparison with those is challenging because they exploit metadata features, such as a hand-annotated indicator of a newspaper's geographic focus Lieberman et al. (2010), that are not available in the version of LGL we have. We compare where possible, but our primary focus is resolution using

only the text, in large part because we are interested in resolution on historical corpora such as CWar, which do not have such metadata available.

3.2.1 Evaluation Metrics

One of the most commonly used metrics in geolocation is Accuracy at 161 kilometers (100 miles). This measure is preferred to simple accuracy because it allows systems to stray from evaluation ontologies (i.e. you can predict Austin, TX to be a point somewhere near the particular gazetteer reference used for annotation). This measure originated in document geolocation, along with two other primary metrics three primary metrics: mean error distance, median error distance Leidner (2008); Eisenstein et al. (2010); Wing and Baldridge (2011); Speriosu and Baldridge (2013); Santos et al. (2014). Unfortunately however there have been different interpretations as to what Accuracy @ 161km means in practice. Some authors seem to calculate this as (the number of predictions within 161km of their referent) / (total number of predictions made) (as with (Speriosu, 2013) and *likely* (Lieberman and Samet, 2012)). Others include all gold toponyms, even ones not predicted on in the calculation (the number of predictions within 161km of their referent) / (total number of annotated toponyms) ((DeLozier et al., 2015) with respect to the TopoCluster systems). For purposes of clarity, one can say that A @161 breaks into two interpretations: precision @ 161km and recall @ 161km).

It is also important to measure the coverage of TR systems with NER errors, as this is the natural use case of TR systems in the wild. Recall in this scenario is affected not only by gazetteer incompleteness, but also failure of NER systems to recognize actual place names. This is one of the places where gazetteer independent techniques should shine, as they remove one of the sources of error in recall calculation.

3.2.2 Parameter tuning

Toponym Weighting: A grid search was run on the dev portions of the datasets to derive values of three parameters θ_1, θ_2, and θ_3 corresponding to weights on the $\vec{g^*}$ of the main toponym, context toponyms, and other context words, respectively. The search was performed by running the disambiguation procedure on 80/20 splits of the dev set using a closed set of parameter values ranging from .5 to 40. Performance

Table 3.1: Toponym Weights Resulting from Gridsearch.

Dataset	**Resolver**	θ_1	θ_2	θ_3	window
TR-CoNLL	TopoCluster	40	1	0.0	10
TR-CoNLL	TopoClusterGaz	40	1	0.0	10
LGL	TopoCluster	20	5	0.0	100
LGL	TopoClusterGaz	10	5	0.0	100
CWar	TopoCluster	40	1	0.0	10
CWar	TopoClusterGaz	40	1	0.0	10
Wotr-Topo	TopoClusterGaz	3	1	0.0	200

of the theta combinations was then averaged over the splits. The combination that produced the lowest average kilometer error scores for the respective models were then selected for future runs on the corpus. Table 3.1 shows the values obtained for the respective Model-Domain combinations. The weights for CWar and TR-CoNLL are very similar, with very strong preferences being shown for spatial statistics of the main toponym. The weights obtained for LGL show more balanced preferences for the clusters associated with both the main and context toponyms.

Domain Adaptation: We also determine values for the $\lambda's$ of Equation 3.2 by varying them from 0 to 1 and measuring A@161, again on 80/20 splits of the dev portions. An average was taken of the accuracy over the 5 splits and is depicted in Figure 3.2. In five of six cases, TopoCluster benefits from domain adaptation; the exception is when using gazetteer matching on TR-CoNLL. This is unsurprising since TR-CoNLL is the corpus most similar to the background GeoWiki corpus and it contains many large, discontinuous geographic entities (e.g. states, countries) that are poorly represented as single points. Predictions for such geographic entities constitute a large portion of changes as the in-domain λ increases. Both CWar and LGL constitute substantially different domains; for these, λ values that equally balance the in-domain and GeoWiki models are best.

3.3 Toponym resolution results

Table 3.2 shows test set performance for all models when resolving gold-standard toponyms. The base TopoCluster model (trained only on GeoWiki and not using a gazetteer) performs relatively poorly, even on TR-CoNLL. However, when combined with in-domain data, it ties for best performance on CWar (A@161 of 93.1) and is

Table 3.2: Toponym resolution performance of all models using oracle NER. *A @ 161 does not mean the same thing with respect to all systems. TopoCluster and TopoClusterGaz actually measure Recall @ 161km while Random, Population, SPIDER, TRIPDL, WISTR, WISTR+SPIDER, TRAWL, and TRAWL+SPIDER all measure Precision @ 161km

	TR-CoNLL			**CWar**			**LGL**		
Resolver	**Mean**	**Median**	**A@161**	**Mean**	**Median**	**A@161***	**Mean**	**Median**	**A@161**
Random	3891	1523.9	38.4	2393	1029	13.4	2852	1078	26.1
Population	219	30.3	90.5	1749	**0.0**	62.0	1529	38	62.7
SPIDER	2175	40.1	65.3	266	**0.0**	67.0	1233	16	68.4
TRIPDL	1488	37.0	72.9	848	**0.0**	60.2	1311	46	60.9
WISTR	281	30.5	89.1	855	**0.0**	73.3	1264	27	64.0
WISTR+SPIDER$_{10}$	432	30.7	87.4	201	**0.0**	87.1	**830**	3	**77.7**
TRAWL	237	30.5	89.7	944	0.0	70.8	2072	324	46.9
TRAWL+SPIDER$_{10}$	300	30.5	89.1	148	**0.0**	88.9	873	6	74.4
TopoCluster $_{\lambda=0}$	560	122	53.2	1226	27	68.4	1735	274	45.5
TopoCluster $_{\lambda=.6\ (.5,LGL)}$	597	20	85.2	141	22	**93.1**	1193	41	67.0
TopoClusterGaz $_{\lambda=0}$	**209**	**0.0**	**93.2**	1199	**0.0**	68.7	1540	1.5	61.4
TopoClusterGaz $_{\lambda=.6\ (.5,LGL)}$	351	**0.0**	91.6	**120**	**0.0**	**93.1**	1183	**0**	74.8

competitive with others for TR-CoNLL (85.2) and LGL (69.0). Furthermore, this strategy is more effective than TopoClusterGaz without domain adaptation on both CWar and LGL, though vanilla TopoClusterGaz does obtain the best performance on TR-CoNLL. This is mostly likely due to two factors: GeoWiki is a good match for the international news domain of TR-CoNLL and the GeoNames gazetteer was one of the main resources used to create TR-CoNLL Leidner (2008).

TopoClusterGaz with domain adaptation is the best overall performer across all datasets. It beats the best models of Speriosu and Baldridge for both TR-CoNLL and CWar by large margins. LGL proves to be a more challenging dataset: TopoClusterGaz is second (by a large margin of 6 absolute percentage points), to WISTR+SPIDER. This indicates an opportunity for further gains by combining TopoCluster and SPIDER. We also performed the self-training technique described previously to see whether bootstrapping information on metadata can help. It does: TopoClusterGaz with domain adaptation and self-training obtains A@161 of 75.8 on LGL, near the 77.7 of WISTR+SPIDER. It also easily beats the 77.5 A@**250** obtained by Santos et al. (2014).

Table 3.3 shows final performance scores for versions of TopoCluster run using an off-the-shelf NER on simple tokenized versions of the TR-CoNLL corpora. In this combined system evaluation, large differentiation is seen between the models

Table 3.3: TR-CoNLL performance with predicted toponyms.

Resolver	**P**	**R**	**F**
Random	26.4	19.2	22.2
Population	71.7	52.0	60.2
SPIDER	49.1	35.6	41.3
TRIPDL	51.8	37.5	43.5
WISTR	73.9	53.6	62.1
WISTR+SPIDER$_{10}$	73.2	53.0	61.5
TRAWL	72.6	52.6	61.0
TRAWL+SPIDER$_{10}$	72.4	52.5	60.8
TopoCluster $_{\lambda=.6}$	75.1	84.0	79.3
TopoCluster $_{\lambda=0}$	46.7	52.2	49.2
TopoCluster-Gaz $_{\lambda=0}$	**81.9**	**91.6**	**86.5**

of Speriosu and Baldridge (2013) and our own, with the largest differences being seen in the Recall metric–though some of the difference likely comes from Speriosu and Balridge's use of OpenNLP NER as opposed to TopoCluster's use of Stanford NER. This large difference in Recall is in part due to our model's non-reliance on gazetteer matching. This makes it possible for TopoCluster to make correct resolutions even in cases when the NER output uses a non-standard place name variant (e.g. *Big Apple* for *NYC*) or when slight errors are made in tokenization or NER (e.g. *NYC.* is output as opposed to *NYC*). TopoCluster succeeds in these cases because language models typically include these variants, and their distributions pattern in ways that are similar to the more commonly occurring dominant form. The advantage of our models in the combined NER and TR evaluation matters because almost all real-world use cases of TR apply to toponyms identified by a named entity recognizer.

Table 3.4 shows the resolution results of many state-of-the-art toponym resolution systems on the WoTR corpus. As can be seen, TopoClusterGazoutperforms all resolvers on all metrics when oracle NER is used, and significantly outperforms others on Recall and F-1 Score. The exact reason for this success is explored more deeply in the next chapter, but key to the system's success is the ability to predict on non-gazetteer matched entities. The ability to use toponyms who lack gazetteer referents aids in direct and indirect ways. First, the system is able to predict in cases where other systems cannot, directly boosting Recall and F-1 Score. Second, the ability to predict predict on non-gazetteer matched entities indirectly bene-

Table 3.4: WoTR Toponym Resolution Results

System	**A@161km***	**Mean**	**P**	**R**	**F-1**
Random	27.0	2259	15.2	6.2	8.8
Population	63.7	1507	39.4	16.1	22.8
SPIDER	68.6	593	37.8	15.4	21.9
WISTR	62.9	965	**53.3**	15.9	24.5
WISTR+SPIDER	68.7	610	38.1	15.5	22.1
TopoCluster	60.0	662	36.2	26.9	30.8
TopoClusterGaz	**71.28**	**564**	40.6	**30.3**	**34.7**

fits gazetteer matched entities. For example, a correspondence includes toponyms "Northern Mississippi" and "Corinth". *Corinth* is a highly ambiguous toponym and *Northern Mississippi* does not appear in gazetteers. Systems such as SPIDER cannot utilize the information from Northern Mississippi as it does not match a toponym; in contrast, TopoClusterGazutilizes the densely clustered Gi* vector of *Northern Mississippi* to disambiguate the ambiguous *Corinth*. NER inclusive scores are generally much lower for WoTR-Topo than other datasets because NER systems utilized (Stanford-NER and openNLP-NER) are trained are very different domains. Nonetheless, TopoClusterGaz outperforms all systems in NER-inclusive scores except for precision, which WISTR prevails.

Chapter 4

Error Analsyis and Conclusions

In this chapter I analyze in detail the errors that TopoCluster, SPIDER, and WISTR make on LGL, TR-CoNLL, and WoTR-Topo.

4.1 Error Analysis

Error in Toponym Resolution systems can be sourced to a relatively small subset of underlying causes. The first group of errors I class as system design or setup design errors. Most commonly this appears as systems being unable to predict on a correct referent due to a dependency in the system on gazetteer matching and also as a prediction that is correct in some sense, but differs from that of the evaluation ontology. The second group of errors are endemic to all TR systems, irrespective of system design. The principal error type in this group is disambiguation error, where a system incorrectly predicts the wrong referent for a toponym.

In the genre of spatial minimization systems, which SPIDER is an example of, disambiguation errors frequently occur in situations where a single article jumps in its geographic focus (e.g. articles with lists of country names). SPIDER also suffers in situations where texts are short and lack a large number toponyms in context. As can be seen in Tables Table 3.2 and Table 3.4 this type of system does much better in domains that have relatively restricted spatial focus (Local news articles of LGL and spatially organized reports/letters of WoTR), but does much poorer on the international news of TR-ConLL.

In the genre of language-context based systems, such as TopoCluster and

WISTR, disambiguation errors frequently occur when the natural language context lacks geographically specifying information. These systems can typically do much better in international domains such as TR-CoNLL, where toponyms are less likely to be geographically close to one another and where frequency of 'prominent' sense toponyms if very high (i.e. 'prominent' toponyms are much more likely to have training contexts which characterize them well).

4.1.1 LGL Errors

The first major class of errors on LGL involve demonyms. Two of these that appear in LGL-dev of are *American* and *Georgian* (Table 4.1). For TopoClusterGaz, the source of errors is listed as 'Gazet Match Error'. This type of error corresponds to situations wherein the system correctly disambiguates to a Gi* cluster in the correct location, but a gazetteer 'snap' step causes an error. Plain TopoCluster correctly disambiguates these demonyms to locations very near their intended referents (LGL annotation dictates they should be at the centroids of the related country). Geonames lacks demonyms as alternate names for countries, but in fact *American* and *Georgian* refer to actual cities completely unrelated to the country. SPIDER and WISTR also fail to predict on the demonyms, but are not penalized in the A @ 161km measure because they are not penalized for non-predictions. While SPIDER and WISTR utilize the same Geonames resource as TopoClusterGaz, TopoClusterGaz allows toponyms to match on the geonames 'alternate names' field, while SPIDER and WISTR do not.

SPIDER and WISTR also fail to make predictions are certain acronyms (Ga.) while TopoClusterGaz does well on these type of references. This, once again, goes back to the difference in whether the systems utilize the altnames field of Geonames.

Certain toponyms TopoClusterGaz is better at disambiguating, while SPIDER and WISTR succeed at others. TopoClusterGaz does well on Boone County, Cookville, and Dublin while SPIDER and WISTR do not. SPIDER fails on these toponyms because the spatial minimization procedure is mislead by clusters of toponyms in areas unrelated to the true referent. One example of this is *Boone County*, which is a county that appears in Missouri, Kentucky, Indiana, and Illinois. In LGL the Boone County (KY) co-occurs with a set of toponyms with possible referents in Indiana. This misleads SPIDER into selecting Boone County (IN). Both TopoClus-

Figure 4.1: TR-CoNLL instances of corpus errors

1. Jiri Novak (Czech Republic) beat Ben Ellwood (Austrialia <lat=27, long=0>) 6-2 6- 4 6-3

2. Sao Paulo <lat=-7.233, long=-62.9> state power firm Electropaulo said it has named Eduardo Bernini as new president, replacing Emmanuel Sobral...

terGaz and WISTR succeed at disambiguating *Israel*, while SPIDER fails. This is due to *Israel* frequently appearing in articles talking about US, Israeli relations. In these situations, *Israel* becomes disambiguated to a place within the US, due to the variety of US toponyms that appear in the articles.

SPIDER on the other hand is better at disambiguating articles with highly concentrated, ambiguous toponyms. It does well where TopoClusterGaz and WISTR fail on toponyms such as *Georgia* and *Iran*. In these cases of LGL-dev, the country names appear with a variety of toponyms which uniquely predict to points in the Caucuses and Iran (the country). TopoClusterGaz can sometimes fail in these scenarios because the GI* densities of country names are typically widely dispersed (i.e. the terms are associated and used over wide areas), and the clusters which do appear don't necessarily overlap with all regions and cities contained in the country. TopoClusterGaz ends up disambiguating *Iran* to a city in western Syria due to the frequency of *Palestine* and *Syria* in its context.

4.1.2 TRCoNLL Errors

Errors observed with TR-CoNLL are a little different than with LGL and WoTR. Suprisingly, the biggest class of errors observed turned out not to actually be errors but instead problems with the annotated corpora. These problems in the corpora exist despite the use of TR-CoNLLf ('f' for fixed) developed by (Speriosu, 2013) to fix problems with the base TR-CoNLL. Because the primary errors are corpora related, I only list the results from TopoClusterGaz.

Table 4.2 displays the primary errors on TR-CoNLL. As an example of 'corpus errors', sentences and associated annotations for *Australia* and *Sao Paulo* are displayed in Figure 4.1. As can be seen, the annotated geo-references for the places are incorrect.

Table 4.1: Principal Errors LGL-Dev

LGL-Dev									
	TopoCluster-Gaz			SPIDER			WISTR		
Toponym	**Freq**	**Avg Er**	**Explanation**	**Freq**	**Avg Err**	**Explanation**	**Freq**	**Avg Err**	**Explanation**
American	6	2000km	Gazet Match Error	6	N/A	No Prediction	6	N/A	No Prediction
Athens	5	9000km	Disambiguation	4	300km	Disambiguation	5	8000km	Disambiguation
Boone County	0	0km	No Errors	3	191km	Disambiguation	3	350km	Disambiguation
Cookville	0	0km	No Errors	3	1000km	Disambiguation	6	2200km	Disambiguation
Dublin	0	0km	No Errors	0	0km	No Errors	8	5700km	Disambiguation
Ga.	0	0km	No Errors	2	N/A	No Prediction	2	N/A	No Prediction
Gainesville	7	800km	Disambiguation	4	800km	Disambiguation	7	800km	Disambiguation
Georgia	2	6000km	Disambiguation	0	50km	No Error	3	10000km	Disambiguation
Georgian	5	8900km	Gazet Match Error	5	N/A	No Prediction	5	N/A	No Prediction
Grapeland	0	0km	No Errors	0	1000km	Disambiguation	2	2000km	Disambiguation
Harrisburg	0	0k	No Errors	3	981km	Disambiguation	0	0k	No Errors
Iran	2	1500km	Disambiguation	0	100km	No Error	0	100km	No Error
Israel	0	0km	No Errors	9	9600km	Disambiguation	0	0km	No Errors
Norfolk	0	0km	No Errors	3	3000km	Disambiguation	3	1900km	Disambiguation
Paris	14	7200km	Disambiguation	14	2200km	Disambiguation	18	7100km	Disambiguation
Springfield	1	428km	Disambiguaton	3	1000km	Disambiguaton	4	1100km	Disambiguation

A second source of error in TR-CoNLL are Gazet Match Errors. These are seen in TR-CoNLL dev with the toponym *Guerrero*. Gi* clusters for *Guerrero* and its context correctly disambiguate to a point within the state of Guerrero Mexico within 161 km of the true referent. Nevertheless, the gazeteer matching procedure causes TopoCluster to snap to a city in a neighboring state rather than the correct referent, due to there being no match for that particular spelling of the state within GeoNames.

Disambiguation errors occur as well with TR-CoNLL, but are generally much less of a problem. *Georgia* for appears 3 times in the dev set (two times as the country, one time as the state). TopoClusterGaz tends to prefer the state sense of the toponym in empty contexts due to fact that there are many more Wikipedia articles for places within the US state of Georgia than the country in the Caucuses.

Figure 4.2: Sao Paulo Error in sentence (2)

Table 4.2: TopoCluster Principal Errors TrCoNLL-Dev

TrCoNLL-Dev			
Toponym	**Frequency**	**Avg Error**	**Error Explanation**
Argentina	14	7600km	Corpus Error
Australia	14	13000km	Corpus Error
Austria	14	1400km	Corpus Error
Berlin	7	337km	Corpus Error
China	3	6300km	Corpus Error
Georgia	2	9800km	Disambiguation Error
Guerrero	6	268km	Gazet Match Error
Johannesburg	5	373km	Corpus Error
Malaysia	7	500km	Corpus Error
Paris	4	440km	Corpus Error
Sao Paulo	7	2500km	Corpus Error
Sudan	8	329km	Corpus Error
WA	3	12000km	Disambiguation Error
Zaire	9	1468km	Disambiguation Error

4.1.3 Wotr-Topo Errors

The errors seen with the systems on WoTR-Topo are mostly disambiguation errors, though gazetteer matching and ontology mismatch errors are also represented.

Toponyms that do well in SPIDER but not other systems are *Alexandria*, *Clarksville*, *Houston*, and *Pocahontas*.

Alexandria solely refers to Alexandria, Lousiana in WoTR-Topo Dev. TopoClusterGaz tends to disambiguate the toponym to the city in Egypt due to the strength of the Gi* cluster there. WISTR tends to disambiguate to Alexandria, Virginia due to the amount of material written on the place, especially concerning the civil war. *Clarksville* refers to Clarksville, Arkansas in WoTR-Topo Dev. TopoClusterGaz tends to disambiguate the toponym to Clarksville, Tennessee due to the strength of the Gi* cluster there and the lack of specifying context. SPIDER succeeds because the documents containing Clarksville are typically headed by a single contextual toponym, Arkansas. This is enough to set SPIDER on the right track, but unfortunately for TopoCluster the Gi* clusters for Arkansas and Clarksville, Arkansas overlap only partially. In documents with a single instance of *Clarksville* and a single instance of *Arkansas*, TopoCluster makes the right decision; however, when a document has multiple instances of the same toponym, the top clustered

sense merely stacks in its preference. In future work, it may be advantageous to TopoCluster to avoid this kind of stacking of redundant context toponyms. *Houston* solely refers to Houston, Missouri in the dev set analyzed here. Because the sense is so minor in Wikipedia and the context not especially helpful, TopoClusterGaz and WISTR struggle to disambiguate correctly. *Pocahontas*, Arkansas is a struggle for TopoClusterGaz and WISTR primarily because instances of this name are strongly associated with Virginia. SPIDER is able to make many fewer errors due to using instances of Arkansas that co-occur in the document. Unfortunately for TopoCluster, the Arkansas clusters only slightly overlap with Pocahontas and the single instance of Arkansas is not enough to outweigh the multiple instances of Pocahontas.

Toponyms that do well in TopoClusterGaz but not other systems are *Ark.*, *Camp Lapwai*, *Colo. Terr.*, *Richmond*, *San Francisco*, and *West Tennessee*.

Ark. does well in TopoCluster and TopoClusterGaz because even without a gazetteer, use of Ark. is centered in Arkansas. WISTR and SPIDER fail to make predictions for this acronym because they do not utilize GeoNames altnames field. *Camp Lapwai*, *West Tennessee*, and *Colo. Terr.* lack entries in GeoNames, but Gi* Vectors of TopoCluster center the strings in the correct locations near western Montana, western Tennesee, and central Colorado. WISTR and SPIDER fail to make predictions because the toponyms are not represented in GeoNames. *Richmond* and *San Francisco* are handled much better by TopoClusterGaz primarily because they only refer to their dominant senses in the dev portion of the corpus. SPIDER struggles with *Richmond* due to a sense in which it refers to a place in southern Pennsylvania; in documents that preample with *Washington*, SPIDER prefers the technically closer *Richmond* within Pennsylvania.

Toponyms which are difficult for all systems are *Bonnet Carre*, *Cal.*, *Jacksonport*, *Mississippi River*, *Red River*, and *Washington*.

Bonnet Carre lacks a GeoNames entry at the annotated referent, but does have an entry in the alternate name field that links to a place in France. For this reason, TopoClusterGaz is listed as a 'Gazet Match Error'–TopoCluster produces the correct reference but the act of gazetteer matching snaps to a place on the completely wrong side of the globe. WISTR and SPIDER do not utilize alternate names in GeoNames, so they simply produce no prediction. *Cal.* lacks an entry in GeoNames, including an altnames field entry. TopoCluster tends to disambiguate

Table 4.3: Principal Errors Wotr-Topo

Wotr-Topo Dev										
	TopoCluster-Gaz				Spider			WISTR		
Toponym	**N**	**Freq**	**Avg Er**	**Explanation**	**Freq**	**Avg Err**	**Explanation**	**Freq**	**Avg Err**	**Explanation**
Alexandria	8	6	10000km	Disambiguation	0	3km	Disambiguation	8	1600km	Disambiguation
Ark.	28	0	0km	No Errors	28	N/A	No Prediction	28	N/A	No Prediction
Arkansas	17	5	1000km	Disambiguation	10	181km	Disambiguation	10	181km	Disambiguation
Black River	10	7	947km	Disambiguation	10	1225km	Disambiguation	10	1600km	Disambiguation
Bonnet Carre	2	2	16479km	Gazet Match Error	2	N/A	No Prediction	2	N/A	No Prediction
Burton's Ford	4	4	400km	Disambiguation	4	N/A	No Prediction	4	N/A	No Prediction
Cal.	20	19	600km	Ontology Mismatch	20	N/A	No Prediction	20	N/A	No Prediction
Camp Curtis	5	5	691km	Gazet Match Error	5	N/A	No Prediction	5	N/A	No Prediction
Camp Lapwai	8	0	15km	No Errors	8	N/A	No Prediction	8	N/A	No Prediction
Clarksville	8	6	564km	Disambiguation	0	2km	No Errors	8	1800km	Disambiguation
Colo. Ter.	6	0	74km	No Errors	6	N/A	No Prediction	6	N/A	No Prediction
Fort Bascom	2	2	400km	Disambiguation	2	N/A	No Prediction	2	N/A	No Prediction
Houston	5	5	896km	Disambiguation	0	0km	No Errors	5	900km	Disambiguation
Jacksonport	11	11	1094km	Disambiguation	10	1094km	Disambiguation	11	1094km	Disambiguation
Mississippi River	6	6	1000km	Ontology Mismatch	6	N/A	No Prediction	6	N/A	No Prediction
N. Mex.	12	9	500km	Ontology Mismatch	12	N/A	No Prediction	12	N/A	No Prediction
North Texas	1	1	247km	Ontology Mismatch	1	N/A	No Prediction	1	N/A	No Prediction
Pocahontas	8	8	1120km	Disambiguation	1	181km	Disambiguation	8	400km	Disambiguation
Red River	30	14	1000km	Disambiguation	30	1200km	Disambiguation	30	1200km	Disambiguation
Richmond	17	3	600km	Disambiguation	17	280km	Disambiguation	6	4000km	Disambiguation
Rio Grande	10	6	2000km	Disambiguation	10	280km	Disambiguation	9	8000km	Disambiguation
San Francisco	28	0	4km	No Errors	8	2000km	Disambiguation	1	1500km	Disambiguation
Springfield	13	10	225km	Disambiguation	8	225km	Disambiguation	11	1852km	Disambiguation
Texas	15	10	800km	Disambiguation	15	190km	Ontology	15	190km	Ontology
Wash. Ter.	16	16	290km	Ontology Mismatch	16	N/A	No Prediction	16	N/A	No Prediction
Washington	25	15	2000km	Disambiguation	18	1200km	Disambiguation	9	3100km	Disambiguation
West Tennessee	1	0	100km	No Errors	1	N/A	No Prediction	1	N/A	No Prediction

this to a point near San Francisco, which is about 300km away from the California state centroid. SPIDER and WISTR cannot disambiguate the toponym due to lacking a gazetteer entry for this particular spelling. *Jacksonport* is always disambiguated to the city in Wisconsin by all systems, even though the correct referent is in Arkansas. The reason for this error is complex; the vasty majority of written material in Wikipedia on Jacksonport concerns the much more populous city in Wisconsin. This causes the Gi* clusters and classifiers of WISTR to more easily select that sense. Adding to this, contexts where *Jacksonport* frequently mention the highly ambiguous physical feature the *Black River*, which has referents in both Arkansas and Wisconsin. Because of this geographically shared referent ambiguity, SPIDER selects *Jacksonport*, Wisconsin over the Arkansas sense. *Mississippi River* runs into ontology mismatch issues; annotators tended to select the Wikipedia entry referent (at the mouth of the river), while GeoNames links to the head waters of the same river. The evaluation metric captures this behavior of TopoClusterGaz as an error, but it is more a matter of parallel ontologies pointing to the same referent. *Washington* is one of the most ambiguous toponyms in GeoNames, and within WoTR-Topo both the Washington, DC sense and Washington state sense are prevalent. Most of the Washington errors occur in scenarios where the dominant focus of the document is geographically localized, but contains a phrase such as "orders from Washington", often in the document header. WISTR does best on *Washington*, getting about 60% of its instances right.

4.2 Conclusions

The gazetteer independent toponym resolver Topocluster-Gaz performs well on international news and historical corpora, beating other state-of-the-art resolvers by large margins. Gazetteer-independent versions of our models perform competitively with many high performing resolvers, and TopoCluster works especially well on predicted toponyms—which is arguably the key use case for toponym resolution in the wild.

One criticism that could me made of toponym resolution systems more generally is that they ought to be subsumed into the more general task of named entity linking, where important scholarship has explored the use of joint modeling (Durrett and Klein, 2014), as well as a variety of machine-learned and probabilistic classifiers

(Shen et al., 2015). The work contained in this thesis however should undermine at least some of this criticism. Toponym grounding is unique among word sense disambiguation in that the words can naturally be projected into an objective 2 dimensional (or 2d + elevation) space. This constrained prediction space allows the development of systems like TopoCluster which are less dependent on human-curated reference ontologies, and as I have demonstrated these systems can perform very well in a variety of domains, side-stepping some issues endemic with entity linking, such as poor recall resulting from ontology mismatches and query expansion failures.

The results of the gazetteer-independent models call into question whether gazetteer matching or ontology matching is truly an essential component of a toponym resolution process. Theoretically, gazetteer matching could do significant work correcting a completely wrong output of a non-gazetteer utilizing model like TopoCluster, particularly in cases where a toponym string is unambiguous in a gazetteer. Empirically however we found these cases to be extremely rare—the primary benefit of the gazetteer in our models was ontology correction of large geographic entities and not disambiguation per se.

4.3 Future Work

Two areas within text geolocation that I think are promising, but I didn't have enough time to explore are (1) automated geographic feature type inference, and (2) application of sequence modeling and related techniques to toponym resolution.

I observed two undesirable behaviors of TopoCluster that were related to issues of geographic feature typing and scale. The first was that more administratively prominent toponyms tend to have G_i*clusters that are more dispersed and thus have much lower peak values. The second, and somewhat related, is that toponyms in the same administrative hierarchy (e.g. Spokane, and Washington) do not necessary overlap in their G_i*distributions. This leads to certain G_i*clustered toponyms, such as cities being a much more dominating force than desired. One way that I could have alleviated this would be to automatically infer the feature type from patterns in the G_i*distribution, and re-weight certain points in the G_i*vector according to the inferred feature typing.

It is somewhat surprising that sequence modeling approaches, which are

widely found in the problem of Named Entity Recognition, have not been applied to resolving sequences of toponyms. Besides being an elegant computational framework for optimally resolving sequences, it gives the toponym resolution community access to a body of scholarship that ought to help with ancillary problems (i.e. handling the tendency of named entity strings to have one referent per discourse (Finkel and Manning, 2009)). The main roadblock I had in applying sequence modeling approaches in TR is that they seem to require a relatively small number of states to work well. This requires partitioning geograhic space into a relatively small subset of areas, which fails to produce predictions at the scale necessary for most toponym resolution tasks. One possible work-around in this framework would be to apply hierarchical modeling techniques like those used in Wing and Baldridge (2014) to alter the scale of the prediction to that for which there is training data.

Bibliography

L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *Proc. of the 17th International Conference on World Wide Web*, WWW '08, pages 357–366, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367546. URL `http://doi.acm.org/10.1145/1367497.1367546`.

Z. Cheng, J. Caverlee, and K. Lee. 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. ACM, 2010.

M. Daoud and J. X. Huang. Mining query-driven contexts for geographic and temporal search. *International Journal of Geographical Information Science*, 27(8): 1530–1549, 2013. URL `http://dblp.uni-trier.de/db/journals/gis/gis27.html\#DaoudH13`.

G. DeLozier, J. Baldridge, and L. London. Gazetteer-independent toponym resolution using geographic word profiles. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. URL `http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9823`.

G. Durrett and D. Klein. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2: 477–490, 2014.

J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empir-*

ical Methods in Natural Language Processing, pages 1277–1287. Association for Computational Linguistics, 2010.

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

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

M. Graham and S. De Sabbata. Mapping information wealth and poverty: The geography of gazetteers. *Environment and Planning A.*, 2016.

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

J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4. URL `http://dl.acm.org/citation.cfm?id=2145432.2145521`.

H. Kamp and U. Reyle. *From discourse to logic: introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory. Part 1.* From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Kluwer Academic, 1993. ISBN 9780792310273. URL `https://books.google.com/books?id=np0hxQVrJxMC`.

S. A. Kripke. *Naming and Necessity.* Harvard University Press, 1980.

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

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

M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 201–212. IEEE, 2010.

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

D. R. Montello, M. F. Goodchild, J. Gottsegen, and P. Fohl. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-3):185–204, 2003.

S. Nesbit. Computation for humanity: Information technology to advance society. chapter Visualizing Emancipation: Mapping the End of Slavery in the American Civil War, pages 427–435. New York: Taylor & Francis, 2013.

N. O'Hare and V. Murdock. Modeling locations with social media. *Information Retrieval*, 16(1):30–62, 2013. ISSN 1386-4564. doi: 10.1007/s10791-012-9195-y. URL `http://dx.doi.org/10.1007/s10791-012-9195-y`.

J. K. Ord and A. Getis. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, 27(4):286–306, 1995.

D. O'Sullivan and D. J. Unwin. *Geographic Information Analysis.* John Wiley & Sons, Hoboken, New Jersey, 2010.

S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proce. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510, 2012.

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

W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460, 2015.

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

M. Speriosu. *Methods and Applications of Text-Driven Toponym Resolution with Indirect Supervision*. PhD thesis, University of Texas at Austin, August 2013.

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

O. van Laere, S. Schockaert, and B. Dhoedt. Georeferencing flickr resources based on textual meta-data. *Information Sciences*, 238:52–74, 2013. URL `http://dblp.uni-trier.de/db/journals/isci/isci238.html\#LaereSD13`.

B. Wing and J. Baldridge. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 336–348, 2014.

B. P. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *Proc. of the 49th Annual Meeting of the Assoc. for Computational Linguistics: Human Language Technologies-Volume 1*, pages 955–964, 2011.

Vita

Permanent Address: 506 West 37th Street Apt 206, Austin TX, 78705

This thesis was typeset with LaTeX 2$_\varepsilon$[1] by the author.

[1] LaTeX 2$_\varepsilon$ is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this thesis were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, Ayman El-Khashab, and Nicholas Gaylord.