Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles

Grant DeLozier and **Jason Baldridge** and **Loretta London**
University of Texas at Austin
Austin TX, 78712
{grantdelozier, jbaldrid}@utexas.edu
loretta.r.london@gmail.com

Abstract

Toponym resolution, or grounding names of places to their actual locations, is an important problem in analysis of both historical corpora and present-day news and web content. Recent approaches have shifted from rule-based spatial minimization methods to machine learned classifiers that use features of the text surrounding a toponym. Such methods have been shown to be highly effective, but they crucially rely on gazetteers and are unable to handle unknown place names or locations. We address this limitation by modeling the geographic distributions of words over the earth's surface: we calculate the geographic profile of each word based on local spatial statistics over a set of geo-referenced language models. These geo-profiles can be further refined by combining in-domain data with background statistics from Wikipedia. Our resolver computes the overlap of all geo-profiles in a given text span; without using a gazetteer, it performs on par with existing classifiers. When combined with a gazetteer, it achieves state-of-the-art performance for two standard toponym resolution corpora (TR-CoNLL and Civil War). Furthermore, it dramatically improves recall when toponyms are identified by named entity recognizers, which often (correctly) find non-standard variants of toponyms.

Introduction

Toponym Resolution (TR) refers to the task of automatically assigning geographic reference to place names in text. Place names are highly ambiguous: some have hundreds of possible geographic referents (e.g. *Washington* is associated with at least 64 unique geographic referents). In addition to being highly ambiguous, most place references occur only rarely. Nonetheless, toponym resolution has far reaching applications, with uses in question answering and information retrieval tasks (Leidner 2008; Daoud and Huang 2013), crisis response, and social and historical research (Smith and Crane 2001; Grover et al. 2010; Nesbit 2013).

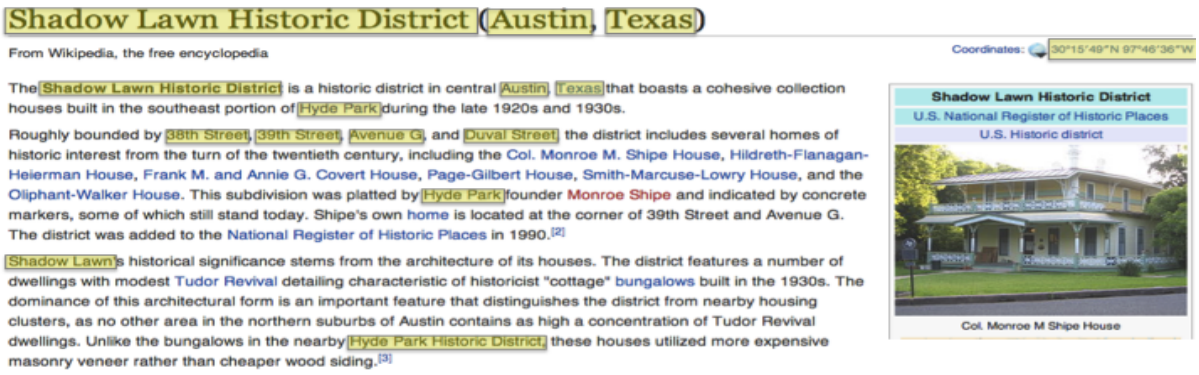
The majority of work in TR until recent years employed heuristic techniques for disambiguating place references. Common approaches included selecting place references with the highest population, selecting more administratively prominent places, and selecting places contained in a minimal geographic area given a set of possible referents. Some of these heuristics have been shown to perform relatively well in some text domains, despite their simplicity (Leidner 2008; Speriosu and Baldridge 2013).

There is surprisingly little research that uses machine learning for toponym resolution. This is in large part due to limited availability of annotated corpora. Most corpora only contain training data for a small proportion of possible referents; e.g., Leidner's TR-CoNLL contains 800 unique referents, while common gazetteers have 9 million unique place referents. In dealing with this paucity of training data, one strategy uses indirect supervision to create large amounts of training data from links and annotations in Wikipedia, such as those shown in Figure 1 (Speriosu and Baldridge 2013; Santos, Anastácio, and Martins 2014). Another strategy incorporates more generalized features sets, such as certain document metadata and population features (Lieberman, Samet, and Sankaranarayanan 2010; Lieberman and Samet 2012; Santos, Anastácio, and Martins 2014). These systems substantially improved upon the performance of heuristic techniques, and they have also improved replicable comparisons with standard datasets and resources.

This paper 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, Caverlee, and Lee 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 regionally specific language models. These regional language models capture differences not just over explicitly geographic context words like *Philadelphia* and *Midwest*, but also over latently geographic words such as *y'all* and *hockey*. We use spatial statistics over these models to flip them to a word-centric perspective that forms the basis of our toponym resolver.

Only limited attempts have been made to use local geo-

Figure 1: Wikipedia page for Shadow Lawn Historic District in Austin, Texas. Place Names and coordinates are highlighted.



graphic clustering techniques in the context of text-based geographic disambiguation. Cheng, Caverlee, and Lee (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 2 gives an example of how the toponym *Washington* is characterized via multiple prominent geographic clusters.

In this paper, we 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.

Data

Corpora

Our toponym resolution system requires documents with latitude-longitude locations to learn per-word spatial distributions. For this, we use **GeoWiki**, the subset of Wikipedia pages that contain latitude-longitude pairs in their info box. We pulled 700,000 such pages in January 2013. Documents were created from these pages by extracting all titles and text from the page and associating these documents with the latitude-longitude in the info box.

We use two corpora used previously by (Speriosu and Baldridge 2013): **TR-CoNLL** (Leidner 2008) and **CWar** (Speriosu 2013). TR-CoNLL 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 resolved locations range from coarse geographic scales (e.g. countries) to fine geographic scales (e.g. parks and neighborhoods). TR-CoNLL was split by Speriosu and Baldridge (2013) into a dev (4,356 Toponyms) and a held-out test set (1,903 Toponyms). Problems have been noted with some annotations in the dataset , including single latitude-longitude gold references for large discontinuous geographies and simple annotation errors (Speriosu and Baldridge 2013).

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. 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 substantially 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 appropriate for evaluating known gold toponyms, but not those identified by a named entity recognizer.

The Local-Global Lexicon corpus (**LGL**) was developed by (Lieberman, Samet, and Sankaranarayanan 2010) to evaluate TR systems on geographically localized text domains. It 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). The dataset contains 5,088 toponyms among which 41% are small populated places.

LGL has critically 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 innermost entities is unconventional in the context of other Toponym Resolution work. The other evaluation corpora used in this study as well as our NER procedure opt for outermost named entity annotations, so we do not directly compare our results to (Lieberman and Samet 2012).

Gazetteers

Previous work has relied on the **Geonames** gazetteer, which contains a wide range of place references, from countries to local parks and bridges. The Geonames gazetteer is global in scope and very large, containing almost 9 million unique places. All types of places are referenced with a single latitude-longitude pair. We additionally use **Natural Earth**, which has country, region, and state shapefiles for 500 unique referents.[1] The shapefiles contain multi-polygon geometries as geographic references, which are important for appropriately representing large, discontinuous geographies.

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*.

The above motivation requires identifying geographic clusters for every word. We derive these by applying local spatial statistics over large numbers of geo-referenced 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.

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

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, Grenager, and Manning 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.

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

This weights the importance of particular documents at locations j to their near grid points i. When used in the $G_i*$of equation 1, the kernel has the effect of smoothing the contributions from each document according to their nearness to i, the current cell under consideration.

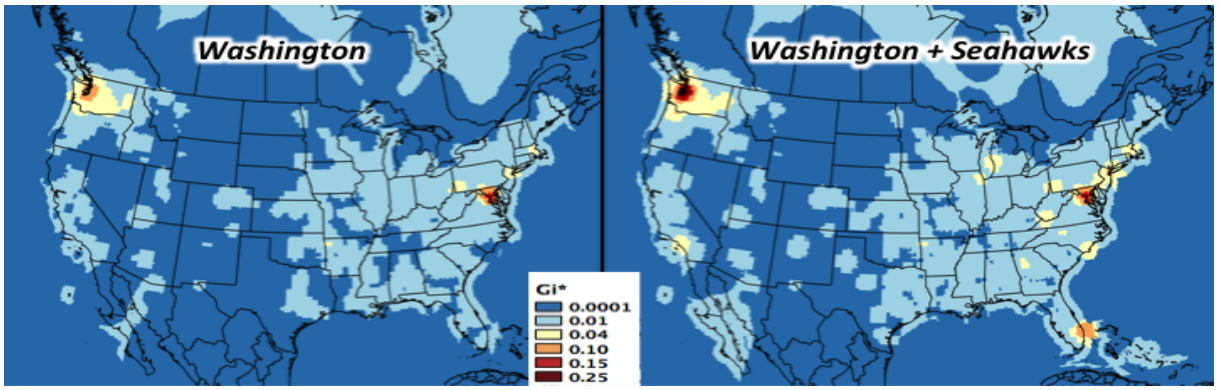
The output of the local statistic calculations is a matrix of statistics with grid cells as columns and each word as a row vector $\vec{g^*}(x)$. The G_i* statistic serves primarily to create a geographically aggregated and smoothed likelihood of seeing each word at certain points in geographic space.

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 2. *Washington* has strong clusters around Washington state and Washington DC, with a slight preference

[1] http://www.naturalearthdata.com/

Figure 2: Left: Local G_i* values for *Washington*. Right: interpolated G_i* values for *Washington* + *Seahawks*.



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 [2].

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.

$$\vec{g^*} = \lambda\, \vec{g^*}_{InDomain} + (1-\lambda)\, \vec{g^*}_{GeoWiki} \quad (3)$$

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 (15 words on each side, filtering out function words). We then compute a weighted sum of all the $\vec{g^*}$ values of the toponyms t and words x in c.

$$g^*(z,c) = \theta_1 \vec{g^*}(z) + \theta_2 \sum_{t \in c} \vec{g^*}(t) + \theta_3 \sum_{x \in c} \vec{g^*}(x) \quad (4)$$

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 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.

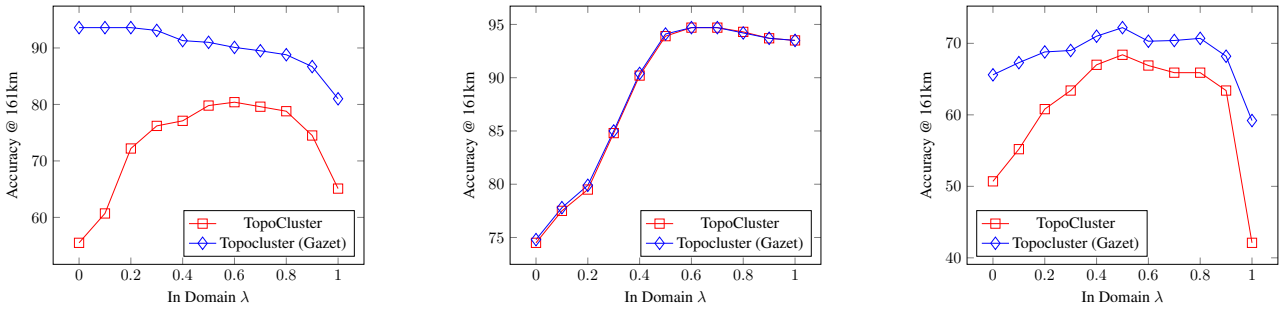
Experimental Setup

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 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

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

Figure 3: Domain adaptation: optimizing λ for each corpus. Left: TR-CoNLL, Middle: CWar, Right: LGL.



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, Anastácio, and Martins 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, Samet, and Sankaranarayanan 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.

Evaluation Metrics

Simple accuracy gives the number of correctly resolved toponyms out of all gold-standard toponyms. This is problematic because not all systems use the same gazetteer. Because of this weakness, other metrics emphasize the distance from predicted to true locations rather than identity of gazetteer entries, and are thus gazetteer-independent. These are standard in document geolocation, where three primary metrics are used: mean error distance, median error distance and accuracy within 161 kilometers (A@161) (Leidner 2008; Eisenstein et al. 2010; Wing and Baldridge 2011; Speriosu and Baldridge 2013; Santos, Anastácio, and Martins 2014).

It is also important to measure the coverage of TR systems. Recall has typically been a problem for TR systems because of incompleteness of gazetteers and spelling variants of toponyms, which means that many candidate toponyms are not resolved. One of our goals with TopoCluster is to handle such candidates well and push up recall, which is rarely measured in other toponym resolution work.

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 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 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.

Table 1: Toponym Weights Resulting from Gridsearch.

Dataset	**Resolver**	θ_1	θ_2	θ_3
TR-CoNLL	TopoCluster	40	1	0.5
TR-CoNLL	TopoClusterGaz	40	1	0.5
LGL	TopoCluster	20	5	1
LGL	TopoClusterGaz	10	5	1
CWar	TopoCluster	40	1	1
CWar	TopoClusterGaz	40	1	1

Domain Adaptation: We also determine values for the $\lambda's$ of Equation 3 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. 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.

Toponym resolution results

Table 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 competitive with

Table 2: Toponym resolution performance of all models using gold-standard toponyms.

	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**	1029	28	69.0
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**	1228	**0**	71.4

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 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 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.

Error Analysis: TopoCluster's most common errors on

Table 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**

LGL were for *Russian, American, Sudanese, Athens,Sudan* and *Iran*; for TR-ConLL they were *Sakai, Malaysia, Kashmir, Iran, Nigeria* and *Sweden*. Overall, the most problematic items for TopoClusterGaz were demonyms. Such expressions–which are usually not included in toponym corpora–are hard for a few reasons. First, the most significant G_i* clusters associated with demonyms tend to be either near capitals or highly populated places, while corpus annotations in LGL geo-reference these entities as a centroid for the large, discontinuous geography associated with its residents. Gazetteer matching fails in these cases because demonyms are not listed as alternate names for countries in Geonames. Also, the demonym string is at times identical to a toponym associated with a different referent, e.g. a city in Venezuela named *Russian* is the only toponym reference for *Russian* in Geonames. Highly ambiguous toponyms like *Athens* are of course inherently challenging.

Gazetteer independent models make their largest errors on large geographic entities (e.g. countries). TopoCluster $_{\lambda=.6}$ for example will typically geo-reference *Sweden* as point near Stockholm, while the oracle TR-CoNLL annotation indicates a point much further north at the geometric center of the country. Similarly, *Sudan* often resolves to a point near Khartoum and not the geographic center of the country. Such resolutions are not incorrect so much as different geographic representations of the same entity.

Conclusions

Our toponym resolvers perform 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. Further improvements could be made on TopoCluster by combining with a spatial minimization algorithm like SPIDER, and by learning (e.g. via logistic regression) fine-grained per-word/per-toponym weights rather than using a grid search over three coarse parameters.

The results of the gazetteer-independent models call into question whether gazetteer 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. We suspect that our results and those of others would change significantly if more complex, multi-polygon annotations would have been given for gold toponyms in annotated corpora.

Acknowledgements

We thank three anonymous reviewers for their helpful feedback; Ben Wing and Michael Speriosu for their help getting Fieldspring's toponym resolution software running on LGL. We additionally thank Michael Lieberman for providing us with the LGL corpus. This research was supported by a grant from the Morris Memorial Trust Fund of the New York Community Trust.

References

Backstrom, L.; Kleinberg, J.; Kumar, R.; and Novak, J. 2008. Spatial variation in search engine queries. In *Proc. of the 17th International Conference on World Wide Web*, WWW '08, 357–366. New York, NY, USA: ACM.

Cheng, Z.; Caverlee, J.; and Lee, K. 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*, 759–768. ACM.

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

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287. Association for Computational Linguistics.

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

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

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

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

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

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

Nesbit, S. 2013. In Zander, J., and Mosterman, P. J., eds., *Computation for Humanity: Information Technology to Advance Society*. New York: Taylor & Francis. chapter Visualizing Emancipation: Mapping the End of Slavery in the American Civil War, 427–435.

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

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

Roller, S.; Speriosu, M.; Rallapalli, S.; Wing, B.; and Baldridge, J. 2012. 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*, 1500–1510.

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

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

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

Speriosu, M. 2013. *Methods and Applications of Text-Driven Toponym Resolution with Indirect Supervision*. Ph.D. Dissertation, University of Texas at Austin.

Wing, B. P., and Baldridge, J. 2011. 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*, 955–964.

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