Text-Driven Toponym Resolution using Indirect Supervision

Michael Speriosu **Jason Baldridge**
Department of Linguistics
University of Texas at Austin
Austin, TX 78712 USA
{speriosu,jbaldrid}@utexas.edu

Abstract

Toponym resolvers identify the specific locations referred to by ambiguous placenames in text. Most resolvers are based on heuristics using spatial relationships between multiple toponyms in a document, or metadata such as population. This paper shows that text-driven disambiguation for toponyms is far more effective. We exploit document-level geotags to indirectly generate training instances for text classifiers for toponym resolution, and show that textual cues can be straightforwardly integrated with other commonly used ones. Results are given for both 19th century texts pertaining to the American Civil War and 20th century newswire articles.

1 Introduction

It has been estimated that at least half of the world's stored knowledge, both printed and digital, has geographic relevance, and that geographic information pervades many more aspects of humanity than previously thought (Petras, 2004; Skupin and Esperbé, 2011). Thus, there is value in connecting linguistic references to places (e.g. placenames) to formal references to places (coordinates) (Hill, 2006). Allowing for the querying and exploration of knowledge in a geographically informed way requires more powerful tools than a keyword-based search can provide, in part due to the ambiguity of toponyms (placenames).

Toponym resolution is the task of disambiguating toponyms in natural language contexts to geographic locations (Leidner, 2008). It plays an essential role in automated geographic indexing and information retrieval. This is useful for historical research that combines age-old geographic issues like territoriality with modern computational tools (Guldi, 2009), studies of the effect of historically recorded travel costs on the shaping of empires (Scheidel et al., 2012), and systems that convey the geographic content in news articles (Teitler et al., 2008; Sankaranarayanan et al., 2009) and microblogs (Gelernter and Mushegian, 2011).

Entity disambiguation systems such as those of Kulkarni et al. (2009) and Hoffart et al. (2011) disambiguate references to people and organizations as well as locations, but these systems do not take into account any features or measures unique to geography such as physical distance. Here we demonstrate the utility of incorporating distance measurements in toponym resolution systems.

Most work on toponym resolution relies on heuristics and hand-built rules. Some use simple rules based on information from a gazetteer, such as population or administrative level (city, state, country, etc.), resolving every instance of the same toponym type to the same location regardless of context (Ladra et al., 2008). Others use relationships between multiple toponyms in a context (local or whole document) and look for containment relationships, e.g. *London* and *England* occurring in the same paragraph or as the bigram *London, England* (Li et al., 2003; Amitay et al., 2004; Zong et al., 2005; Clough, 2005; Li, 2007; Volz et al., 2007; Jones et al., 2008; Buscaldi and Rosso, 2008; Grover et al., 2010). Still others first identify unambiguous toponyms and then disambiguate other toponyms based on geopolitical relationships with or distances to the unambiguous ones (Ding et al., 2000). Many favor resolutions of toponyms within a local context or document that cover a smaller geographic area over those that are more dispersed (Rauch et al., 2003; Leidner, 2008; Grover et al., 2010; Loureiro et al., 2011; Zhang et al., 2012). Roberts et al. (2010) use relationships learned between people, organizations, and locations from Wikipedia to aid in toponym resolution when such named entities are present, but do not exploit any other textual context.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1466–1476,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

Most of these approaches suffer from a major weakness: they rely primarily on spatial relationships and metadata about locations (e.g., population). As such, they often require nearby toponyms (including unambiguous or containing toponyms) to resolve ambiguous ones. This reliance can result in poor coverage when the required information is missing in the context or when a document mentions locations that are neither nearby geographically nor in a geopolitical relationship. There is a clear opportunity that most ignore: use non-toponym textual context. Spatially relevant words like *downtown* that are not explicit toponyms can be strong cues for resolution (Hollenstein and Purves, 2012). Furthermore, the connection between non-spatial words and locations has been successfully exploited in data-driven approaches to document geolocation (Eisenstein et al., 2010, 2011; Wing and Baldridge, 2011; Roller et al., 2012) and other tasks (Hao et al., 2010; Pang et al., 2011; Intagorn and Lerman, 2012; Hecht et al., 2012; Louwerse and Benesh, 2012; Adams and McKenzie, 2013).

In this paper, we learn resolvers that use all words in local or document context. For example, the word *lobster* appearing near the toponym *Portland* indicates the location is Portland in Maine rather than Oregon or Michigan. Essentially, we learn a text classifier per toponym. There are no massive collections of toponyms labeled with locations, so we train models indirectly using geotagged Wikipedia articles. Our results show these text classifiers are far more accurate than algorithms based on spatial proximity or metadata. Furthermore, they are straightforward to combine with such algorithms and lead to error reductions for documents that match those algorithms' assumptions.

Our primary focus is toponym resolution, so we evaluate on toponyms identified by human annotators. However, it is important to consider the utility of an end-to-end toponym identification and resolution system, so we also demonstrate that performance is still strong when toponyms are detected with a standard named entity recognizer.

We have implemented all the models discussed in this paper in an open source software package called Fieldspring, which is available on GitHub: `http://github.com/utcompling/fieldspring` Explicit instructions are provided for preparing data and running code to reproduce our results.

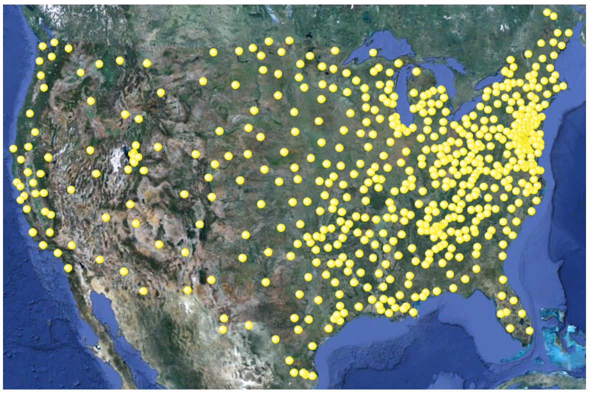

Figure 1: Points representing the United States.

2 Data

2.1 Gazetteer

Toponym resolvers need a gazetteer to obtain candidate locations for each toponym. Additionally, many gazetteers include other information such as population and geopolitical hierarchy information. We use GEONAMES, a freely available gazetteer containing over eight million entries worldwide.[1] Each location entry contains a name (sometimes more than one) and latitude/longitude coordinates. Entries also include the location's administrative level (e.g. city or state) and its position in the geopolitical hierarchy of countries, states, etc.

GEONAMES gives the locations of regional items like states, provinces, and countries as single points. This is clearly problematic when we seek connections between words and locations: e.g. we might learn that many words associated with the USA are connected to a point in Kansas. To get around this, we represent regional locations as a set of points derived from the gazetteer. Since regional locations are named in the entries for locations they contain, all locations contained in the region are extracted (in some cases over 100,000 of them) and then k-means is run to find a smaller set of spatial centroids. These act as a tractable proxy for the spatial extent of the entire region. k is set to the number of 1° by 1° grid cells covered by that region. Figure 1 shows the points computed for the United States.[2] A nice property of this representation is that it does not involve region shape files and the additional programming infrastructure they require.

[1] Downloaded April 16, 2013 from `www.geonames.org`.

[2] The representation also contains three points each in Hawaii and Alaska not shown in Figure 1.

Corpus	docs	toks	types	toks$_{top}$	types$_{top}$	amb$_{avg}$	amb$_{max}$
TRC-DEV	631	136k	17k	4356	613	15.0	857
TRC-DEV-NER	-	-	-	3165	391	18.2	857
TRC-TEST	315	68k	11k	1903	440	13.7	857
TRC-TEST-NER	-	-	-	1346	305	15.7	857
CWAR-DEV	228	33m	200k	157k	850	29.9	231
CWAR-TEST	113	25m	305k	85k	760	31.5	231

Table 1: Statistics of the corpora used for evaluation. Columns subscripted by *top* give figures for toponyms. The last two columns give the average number of candidate locations per toponym token and the number of candidate locations for the most ambiguous toponym.

A location for present purposes is thus a set of points on the earth's surface. The distance between two locations is computed as the great circle distance between the closest pair of representative points, one from each location.

2.2 Toponym Resolution Corpora

We need corpora with toponyms identified and resolved by human annotators for evaluation. The TR-CONLL corpus (Leidner, 2008) contains 946 REUTERS news articles published in August 1996. It has about 204,000 words and articles range in length from a few hundred words to several thousand words. Each toponym in the corpus was identified and resolved by hand.[3] We place every third article into a test portion (TRC-TEST) and the rest in a development portion. Since our methods do not learn from explicitly labeled toponyms, we do not need a training set.

The Perseus Civil War and 19th Century American Collection (CWAR) contains 341 books (58 million words) written primarily about and during the American Civil War (Crane, 2000). Toponyms were annotated by a semi-automated process: a named entity recognizer identified toponyms, and then coordinates were assigned using simple rules and corrected by hand. We divide CWAR into development (CWAR-DEV) and test (CWAR-TEST) sets in the same way as TR-CONLL.

Table 1 gives statistics for both corpora, including the number and ambiguity of gold standard toponyms for both as well as NER identified toponyms for TR-CONLL.[4] We use the pre-trained English NER from the OpenNLP project.[5]

2.3 Geolocated Wikipedia Corpus

The GEOWIKI dataset contains over one million English articles from the February 11, 2012 dump of Wikipedia. Each article has human-annotated latitude/longitude coordinates. We divide the corpus into training (80%), development (10%), and test (10%) at random and perform preprocessing to remove markup in the same manner as Wing and Baldridge (2011). The training portion is used here to learn models for text-driven resolvers.

3 Toponym Resolvers

Given a set of toponyms provided via annotations or identified using NER, a resolver must select a candidate location for each toponym (or, in some cases, a resolver may abstain). Here, we describe baseline resolvers, a heuristic resolver based on the usual cues used in most toponym resolvers, and several text-driven resolvers. We also discuss combining heuristic and text-driven resolvers.

3.1 Baseline Resolvers

RANDOM For each toponym, the RANDOM resolver randomly selects a location from those associated in the gazetteer with that toponym.

POPULATION The POPULATION resolver selects the location with the greatest population for each toponym. It is generally quite effective, but when a toponym has several locations with large populations, it is often wrong. Also, it can only be used when such information is available, and it is

[3] We found several systematic types of errors in the original TR-CONLL corpus, such as coordinates being swapped for some locations and some longitudes being zero or the negative of their correct values. We repaired many of these errors, though some more idiosyncratic mistakes remain. We, along with Jochen Leidner, will release this updated version shortly and will link to it from our Fieldspring GitHub page.

[4] States and countries are not annotated in CWAR, so we do not evaluate end-to-end using NER plus toponym resolution for it as there are many (falsely) false positives.

[5] `opennlp.apache.org`

less effective if the population statistics are from a time period different from that of the corpus.

3.2 SPIDER

Leidner (2008) describes two general and useful *minimality* properties of toponyms:

- *one sense per discourse*: multiple tokens of a toponym in the same text generally do not refer to different locations in the same text
- *spatial minimality*: different toponyms in a text tend refer to spatially near locations

Many toponym resolvers exploit these (Smith and Crane, 2001; Rauch et al., 2003; Leidner, 2008; Grover et al., 2010; Loureiro et al., 2011; Zhang et al., 2012). Here, we define SPIDER (Spatial Prominence via Iterative Distance Evaluation and Reweighting) as a strong representative of such textually unaware approaches. In addition to capturing both minimality properties, it also identifies the relative prominence of the locations for each toponym in a given corpus.

SPIDER resolves each toponym by finding the location for each that minimizes the sum distance to *all* locations for *all* other toponyms in the same document. On the first iteration, it tends to select locations that clump spatially: if *Paris* occurs with *Dallas*, it will choose Paris, Texas even though the topic may be a flight from Texas to France. Further iterations bring Paris, France into focus by capturing its prominence across the corpus. The key intuition is that most documents will discuss Paris, France and only a small portion of these mention places close to Paris, Texas; thus, Paris, France will be selected on the first iteration for many documents (though not for the *Dallas* document). SPIDER thus assigns each candidate location a weight (initialized to 1.0), which is re-estimated on each iteration. The adjusted distance between two locations is computed as the great circle distance divided by the product of the two locations' weights. At the end of an iteration, each candidate location's weight is updated to be the fraction of the times it was chosen times the number of candidates for that toponym. The weights are global, with one for each location in the gazetteer, so the same weight vector is used for each token of a given toponym on a given iteration.

For example, if after the first iteration Paris, France is chosen thrice, Paris, Texas once, and Paris, Arkansas never, the global weights of these locations are $(3/4)*3{=}2.25$, $(1/4)*3{=}.75$, and $(0/4)*3{=}0$, respectively (assume, for the example, there are no other locations named *Paris*). The sum of the weights remains equal to the number of candidate locations. The updated weights are used on the next iteration, so Paris, France will seem "closer" since any distance computed to it is divided by a number greater than one. Paris, Texas will seem somewhat further away, and Paris, Arkansas infinitely far away. The algorithm continues for a fixed number of iterations or until the weights do not change more than some threshold. Here, we run SPIDER for 10 iterations; the weights have generally converged by this point.

When only one toponym is present in a document, we simply select the candidate with the greatest weight. When there is no such weight information, such as when the toponym does not co-occur with other toponyms anywhere in the corpus, we select a candidate at random.

SPIDER captures prominence, but we stress it is not our main innovation: its purpose is to be a benchmark for text-driven resolvers to beat.

3.3 Text-Driven Resolvers

The text-driven resolvers presented in this section all use local context windows, document context, or both, to inform disambiguation.

TRIPDL We use a document geolocator trained on GeoWiki's document location labels. Others—such as Smith and Crane (2001)—have estimated a document-level location to inform toponym resolution, but ours is the first we are aware of to use training data from a different domain to build a document geolocator that uses all words (not only toponyms) to estimate a document's location. We use the document geolocation method of Wing and Baldridge (2011). It discretizes the earth's surface into 1° by 1° grid cells and assigns Kullback-Liebler divergences to each cell given a document, based on language models learned for each cell from geolocated Wikipedia articles. We obtain the probability of a cell c given a document d by the standard method of exponentiating the negative KL-divergence and normalizing these values over all cells:

$$P(c|d) = \frac{\exp(-KL(c,d))}{\sum_{c'} \exp(-KL(c',d))}$$

This distribution is used for all toponyms t in d to define distributions $P_{DL}(l|t,d)$ over candidate locations of t in document d to be the portion of $P(c|d)$ consistent with the t's candidate locations:

$$P_{DL}(l|t,d) = \frac{P(c_l|d)}{\sum_{l' \in G(t)} P(c_{l'}|d)}$$

where $G(t)$ is the set of the locations for t in the gazetteer, and c_l is the cell containing l. TRIPDL (Toponym Resolution Informed by Predicted Document Locations) chooses the location that maximizes P_{DL}.

WISTR While TRIPDL uses an off-the-shelf document geolocator to capture the geographic gist of a document, WISTR (Wikipedia Indirectly Supervised Toponym Resolver) instead directly targets each toponym. It learns text classifiers based on local context window features trained on instances automatically extracted from GEOWIKI.

To create the indirectly supervised training data for WISTR, the OpenNLP named entity recognizer detects toponyms in GEOWIKI, and candidate locations for each toponym are retrieved from GEONAMES. Each toponym with a location within 10km of the document location is considered a mention of that location. For example, the *Empire State Building* Wikipedia article has a human-provided location label of (40.75,-73.99). The toponym *New York* is mentioned several times in the article, and GEONAMES lists a *New York* at (40.71,-74.01). These points are 4.8km apart, so each mention of *New York* in the document is considered a reference to New York City.

Next, context windows w of twenty words to each side of each toponym are extracted as features. The label for a training instance is the candidate location closest to the document location. We extract 1,489,428 such instances for toponyms relevant to our evaluation corpora. These instances are used to train logistic regression classifiers $P(l|t,w)$ for location l and toponym t. To disambiguate a new toponym, WISTR chooses the location that maximizes this probability.

Few such probabilistic toponym resolvers exist in the literature. Li (2007) builds a probability distribution over locations for each toponym, but still relies on nearby toponyms that could refer to regions that contain that toponym and requires hand construction of distributions. Other learning approaches to toponym resolution (e.g. Smith and Mann (2003)) require explicit unambiguous mentions like *Portland, Maine* to construct training instances, while our data gathering methodology does not make such an assumption. Overell and Rüger (2008) and Overell (2009) only use nearby toponyms as features. Mani et al. (2010) and Qin et al. (2010) use other word types but only in a local context, and they require toponym-labeled training data. Our approach makes use of all words in local and document context and requires no explicitly labeled toponym tokens.

TRAWL We bring TRIPDL, WISTR, and standard toponym resolution cues about administrative levels together with TRAWL (Toponym Resolution via Administrative levels and Wikipedia Locations). The general form of a probabilistic resolver that utilizes such information to select a location $\hat{l}$ for a toponym t in document d may be defined as

$$\hat{l} = \arg\max_l P(l, a_l|t,d).$$

where a_l is the administrative level (country, state, city) for l in the gazetteer. This captures the fact that countries (like Sudan) tend to be referred to more often than small cities (like Sudan, Texas). The above term is simplified as follows:

$$\begin{aligned} P(l, a_l|t,d) &= P(a_l|t,d)P(l|a_l,t,d) \\ &\approx P(a_l|t)P(l|t,d) \end{aligned}$$

where we approximate the administrative level prediction as independent of the document, and the location as independent of administrative level. The latter term is then expressed as a linear combination of the local context (WISTR) and the document context (TRIPDL):

$$P(l|t,d) = \lambda_t P(l|t,c_t) + (1{-}\lambda_t)P_{DL}(l|t,d).$$

λ_t, the weight of the local context distribution, is set according to the confidence that a prediction based on local context is correct:

$$\lambda_t = \tfrac{f(t)}{f(t)+C},$$

where $f(t)$ is the fraction of training instances of toponym t of all instances extracted from GEOWIKI. C is set experimentally; C=.0001 was the optimal value for CWAR-DEV. Intuitively, the larger C is, the greater $f(t)$ must be for the local context to be trusted over the document context.

We define $P(a|t)$, the administrative level component, to be the fraction of representative points for a location $\hat{l}$ out of the number of representatives points for all candidate locations $l \in t$,

$$\tfrac{||R_{\hat{l}}||}{\sum_{l' \in t} ||R_{l'}||}$$

where $||R_l||$ is the number of representative points of l. This boosts states and countries since higher probability is assigned to locations with more points (and cities have just one point).

Taken together, the above definitions yield the TRAWL resolver, which selects the optimal candidate location $\hat{l}$ according to

$$\hat{l} = \arg\max_l P(a_l|t)(\lambda_t P(l|t, c_t) + (1-\lambda_t)P_{DL}(l|t, d)).$$

3.4 Combining Resolvers and Backoff

SPIDER begins with uniform weights for each candidate location of each toponym. WISTR and TRAWL both output distributions over these locations based on outside knowledge sources, and can be used as more informed initializations of SPIDER than the uniform ones. We call these combinations WISTR+SPIDER and TRAWL+SPIDER.[6]

WISTR fails to predict when encountering a toponym it has not seen in the training data, and TRIPDL fails when a toponym only has locations in cells with no probability mass. TRAWL fails when both of these are true. In these cases, we select the candidate location geographically closest to the most likely cell according to TRIPDL's $P(c|d)$ distribution.

3.5 Document Size

For SPIDER, runtime is quadratic in the size of documents, so breaking up documents vastly reduces runtime. It also restricts the minimality heuristic—appropriately—to smaller spans of text. For resolvers that take into account the surrounding document when determining how to resolve a toponym, such as TRIPDL and TRAWL, it can often be beneficial to divide documents into smaller subdocuments in order to get a better estimate of the overall geographic prominence of the text surrounding a toponym, but at a more coarse-grained level than the local context models provide. For these reasons, we simply divide each book in the CWAR corpus into small subdocuments of at most 20 sentences.

4 Evaluation

Many prior efforts use a simple accuracy metric: the fraction of toponyms whose predicted location is the same as the gold location. Such a metric can be problematic, however. The gazetteer used by a resolver may not contain, for a given toponym, a location whose latitude and longitude *exactly* match the gold label for the toponym (Leidner, 2008). Also, some errors are worse than others, e.g. predicting a toponym's location to be on the other side of the world versus predicting it to be a different city in the same country—accuracy does not reflect this difference.

We choose a metric that instead measures the distance between the correct and predicted location for each toponym and compute the mean and median of all such error distances. This is used in document geolocation work (Eisenstein et al., 2010, 2011; Wing and Baldridge, 2011; Roller et al., 2012) and is related to the root mean squared distance metric discussed by Leidner (2008).

It is important to understand performance on plain text (without gold toponyms), which is the typical use case for applications using toponym resolvers. Both the accuracy metric and the error-distance metric encounter problems when the set of predicted toponyms is not the same as the set of gold toponyms (regardless of locations), e.g. when a named entity recognizer is used to identify toponyms. In this case, we can use precision and recall, where a true positive is defined as the prediction of a correctly identified toponym's location to be as close as possible to its gold label, given the gazetteer used. False positives occur when the NER incorrectly predicts a toponym, and false negatives occur when it fails to predict a toponym identified by the annotator. When a correctly identified toponym receives an incorrect location prediction, this counts as both a false negative and a false positive. We primarily present results from experiments with gold toponyms but include an accuracy measure for comparability with results from experiments run on plain text with a named entity recognizer. This accuracy metric simply computes the fraction of toponyms that were resolved as close as possible to their gold label given the gazetteer.

5 Results

Table 2 gives the performance of the resolvers on the TR-CoNLL and CWAR test sets when gold toponyms are used. Values for RANDOM and SPIDER are averaged over three trials. The ORACLE row gives results when the candidate

[6] We scale each toponym's distribution as output by WISTR or TRAWL by the number of candidate locations for that toponym, since the total weight for each toponym in SPIDER is the number of candidate locations, not 1.

Resolver	TRC-TEST			CWAR-TEST		
	Mean	Med.	A	Mean	Med.	A
ORACLE	105	19.8	100.0	0.0	0.0	100.0
RANDOM	3915	1412	33.5	2389	1027	11.8
POPULATION	**216**	23.1	81.0	1749	**0.0**	59.7
SPIDER$_{10}$	2180	30.9	55.7	266	**0.0**	57.5
TRIPDL	1494	29.3	62.0	847	**0.0**	51.5
WISTR	279	**22.6**	**82.3**	855	**0.0**	69.1
WISTR+SPIDER$_{10}$	430	23.1	81.8	201	**0.0**	**85.9**
TRAWL	235	**22.6**	81.4	945	**0.0**	67.8
TRAWL+SPIDER$_{10}$	297	23.1	80.7	**148**	**0.0**	78.2

Table 2: Accuracy and error distance metrics on test sets with gold toponyms.

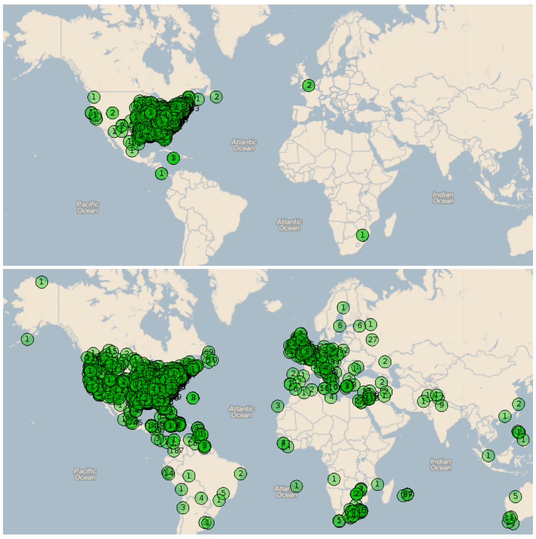

Figure 2: Visualization of how SPIDER clumps most predicted locations in the same region (above), on the CWAR-DEV corpus. TRAWL's output (below) is much more dispersed.

from GEONAMES closest to the annotated location is always selected. The ORACLE mean and median error values on TR-CONLL are nonzero due to errors in the annotations and inconsistencies stemming from the fact that coordinates from GEONAMES were not used in the annotation of TR-CONLL.

On both datasets, SPIDER achieves errors and accuracies much better than RANDOM, validating the intuition that authors tend to discuss places near each other more often than not, while some locations are more prominent in a given corpus despite violating the minimality heuristic. The text-driven resolvers vastly outperform SPIDER, showing the effectiveness of textual cues for toponym resolution.

The local context resolver WISTR is very effective: it has the highest accuracy for TR-CONLL, though two other text-based resolvers also beat the challenging POPULATION baseline's accuracy. TRAWL achieves a better mean distance metric for TR-CONLL, and when used to seed SPIDER, it obtains the lowest mean error on CWAR by a large margin. SPIDER seeded with WISTR achieves the highest accuracy on CWAR. The overall geographic scope of CWAR, a collection of documents about the American Civil War, is much smaller than that of TR-CONLL (articles about international events). This makes toponym resolution easier overall (especially error distances) for minimality resolvers like SPIDER, which primarily seek tightly clustered sets of locations. This behavior is quite clear in visualizations of predicted locations such as Figure 2.

On the CWAR dataset, POPULATION performs relatively poorly, demonstrating the fragility of population-based decisions for working with historical corpora. (Also, we note that POPULATION is not a resolver *per se* since it only ever predicts one location for a given toponym, regardless of context.)

Table 3 gives results on TRC-TEST when NER-identified toponyms are used. In this case, the ORACLE results are less than 100% due to the limitations of the NER, and represent the best possible results given the NER we used.

When resolvers are run on NER-identified toponyms, the text-driven resolvers that use local context again easily beat SPIDER. WISTR achieves the best performance. The named entity recognizer is likely better at detecting common toponyms than rare toponyms due to the na-

Resolver	P	R	F
ORACLE	82.6	59.9	69.4
RANDOM	25.1	18.2	21.1
POPULATION	71.6	51.9	60.2
SPIDER$_{10}$	40.5	29.4	34.1
TRIPDL	51.8	37.5	43.5
WISTR	**73.9**	**53.6**	**62.1**
WISTR+SPIDER$_{10}$	73.2	53.1	61.5
TRAWL	72.5	52.5	60.9
TRAWL+SPIDER$_{10}$	72.0	52.2	60.5

Table 3: Precision, recall, and F-score of resolvers on TRC-TEST with NER-identified toponyms.

ture of its training data, and many more local context training instances were extracted from common toponyms than from rare ones in Wikipedia. Thus, our model that uses *only* these local context models does best when running on NER-identified toponyms. We also measured the mean and median error distance for toponyms correctly identified by the named entity recognizer, and found that they tended to be 50-200km worse than for gold toponyms. This also makes sense given the named entity recognizer's tendency to detect common toponyms: common toponyms tend to be more ambiguous than others.

Results on TR-CoNLL indicate much higher performance than the resolvers presented by Leidner (2008), whose F-scores do not exceed 36.5% with either gold or NER toponyms.[7] TRC-TEST is a subset of the documents Leidner uses (he did not split development and test data), but the results still come from overlapping data. The most direct comparison is SPIDER's F-score of 39.7% compared to his LSW03 algorithm's 35.6% (both are minimality resolvers). However, our evaluation is more penalized since SPIDER loses precision for NER's false positives (Jack *London* as a location) while Leidner only evaluated on actual locations. It thus seems fair to conclude that the text-driven classifiers, with F-scores in the mid-50's, are much more accurate on the corpus than previous work.

6 Error Analysis

Table 4 shows the ten toponyms that caused the greatest total error distances from TRC-DEV with gold toponyms when resolved by TRAWL, the resolver that achieves the lowest mean error on that dataset among all our resolvers.

Washington, the toponym contributing the most total error, is a typical example of a toponym that is difficult to resolve, as there are two very prominent locations within the United States with the name. Choosing one when the other is correct results in an error of over 4000 kilometers. This occurs, for example, when TRAWL chooses Washington state in the phrase *Israel's ambassador to* **Washington**, where more knowledge about the status of Washington, D.C. as the political center of the United States (e.g. in the form of more or better contextual training instances) could overturn the administrative level component's preference for states.

An instance of *California* in a baseball-related news article is incorrectly predicted to be the town California, Pennsylvania. The context is: *...New York starter Jimmy Key left the game in the first inning after Seattle shortstop Alex Rodriguez lined a shot off his left elbow. The Yankees have lost 12 of their last 19 games and their lead in the AL East over Baltimore fell to five games. At* **California**, *Tim Wakefield pitched a six-hitter for his third complete game of the season and Mo Vaughn and Troy O'Leary hit solo home runs in the second inning as the surging Boston Red Sox won their third straight 4-1 over the California Angels. Boston has won seven of eight and is 20-6...* The presence of many east coast cues—both toponym and otherwise—make it unsurprising that the resolver would predict California, Pennsylvania despite the administrative level component's heavier weighting of the state.

The average errors for the toponyms *Australia* and *Russia* are fairly small and stem from differences in how countries are represented across different gazetteers, not true incorrect predictions.

Table 5 shows the toponyms with the greatest errors from CWAR-DEV with gold toponyms when resolved by WISTR+SPIDER. *Rome* is sometimes predicted as cities in Italy and other parts of Europe rather than Rome, Georgia, though it correctly selects the city in Georgia more often than not due to SPIDER's preference for tightly clumped sets of locations. *Mexico*, however, frequently gets incorrectly selected as a city in Maryland near many other locations in the corpus when TRAWL's administrative level component is not present. Many other of the toponyms contributing to the total error such as *Jackson* and *Lexington* are

[7] Leidner (2008) reports precision, recall, and F-score values even with gold toponyms, since his resolvers can abstain.

Toponym	N	Mean	Total
Washington	25	3229	80717
Gaza	12	5936	71234
California	8	5475	43797
Montana	3	11635	34905
WA	3	11221	33662
NZ	2	14068	28136
Australia	88	280	24600
Russia	72	260	18712
OR	2	9242	18484
Sydney	12	1422	17067

Table 4: Toponyms with the greatest total error distances in kilometers from TRC-DEV with gold toponyms resolved by TRAWL. N is the number of instances, and the mean error for each toponym type is also given.

Toponym	N	Mean	Total
Mexico	1398	2963	4142102
Jackson	2485	1210	3007541
Monterey	353	2392	844221
Haymarket	41	15663	642170
McMinnville	145	3307	479446
Alexandria	1434	314	450863
Eastport	184	2109	388000
Lexington	796	442	351684
Winton	21	15881	333499
Clinton	170	1401	238241

Table 5: Top errors from CWAR-DEV resolved by TRAWL+SPIDER.

simply the result of many American towns sharing the same names and a lack of clear disambiguating context.

7 Conclusion

Our text-driven resolvers prove highly effective for both modern day newswire texts and 19th century texts pertaining to the Civil War. They easily outperform standard minimality toponym resolvers, but can also be combined with them. This strategy works particularly well when predicting toponyms on a corpus with relatively restricted geographic extents. Performance remains good when resolving toponyms identified automatically, indicating that end-to-end systems based on our models may improve the experience of digital humanities scholars interested in finding and visualizing toponyms in large corpora.

Acknowledgements

We thank: the three anonymous reviewers, Grant DeLozier, and the UT Austin Natural Language Learning reading group, for their helpful feedback; Ben Wing, for his document geolocation software; Jochen Leidner, for providing the TR-CoNLL corpus as well as feedback on earlier versions of this paper; and Scott Nesbit, for providing the annotations for the CWAR corpus. This research was supported by a grant from the Morris Memorial Trust Fund of the New York Community Trust.

References

B. Adams and G. McKenzie. Inferring thematic places from spatially referenced natural language descriptions. *Crowdsourcing Geographic Knowledge*, pages 201–221, 2013.

E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, 2004.

D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301–313, 2008.

P. Clough. Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 25–30. ACM, 2005.

G. Crane. The Perseus Digital Library, 2000. URL `http://www.perseus.tufts.edu`.

J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 545–556, 2000.

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

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

J. Gelernter and N. Mushegian. Geo-parsing messages from microtext. *Transactions in GIS*, 15 (6):753–773, 2011.

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. Guldi. The spatial turn. *Spatial Humanities: a Project of the Institute for Enabling*, 2009.

Q. Hao, R. Cai, C. Wang, R. Xiao, J. Yang, Y. Pang, and L. Zhang. Equip tourists with knowledge mined from travelogues. In *Proceedings of the 19th international conference on World wide web*, pages 401–410, 2010.

B. Hecht, S. Carton, M. Quaderi, J. Schöning, M. Raubal, D. Gergle, and D. Downey. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 415–424. ACM, 2012.

L. Hill. *Georeferencing: The Geographic Associations of Information*. MIT Press, 2006.

J. Hoffart, M. 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*, pages 782–792. Association for Computational Linguistics, 2011.

L. Hollenstein and R. Purves. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, (1):21–48, 2012.

S. Intagorn and K. Lerman. A probabilistic approach to mining geospatial knowledge from social annotations. In *Conference on Information and Knowledge Management (CIKM)*, 2012.

C. Jones, R. Purves, P. Clough, and H. Joho. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 2008.

S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM, 2009.

S. Ladra, M. Luaces, O. Pedreira, and D. Seco. A toponym resolution service following the OGC WPS standard. In *Web and Wireless Geographical Information Systems*, volume 5373, pages 75–85. 2008.

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

H. Li, R. Srihari, C. Niu, and W. Li. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, pages 39–44, 2003.

Y. Li. Probabilistic toponym resolution and geographic indexing and querying. Master's thesis, The University of Melbourne, Melbourne, Australia, 2007.

V. Loureiro, I. Anastácio, and B. Martins. Learning to resolve geographical and temporal references in text. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 349–352, 2011.

M. Louwerse and N. Benesh. Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. *Cognitive science*, 36(8):1556–1569, 2012.

I. Mani, C. Doran, D. Harris, J. Hitzeman, R. Quimby, J. Richer, B. Wellner, S. Mardis, and S. Clancy. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280, 2010.

S. Overell. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. PhD thesis, Imperial College London, 2009.

S. Overell and S. Rüger. Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22:265–287, 2008.

Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang. Summarizing tourist destinations by mining user-generated travelogues and photos. *Computer Vision and Image Understanding*, 115(3):352 – 363, 2011.

V. Petras. Statistical analysis of geographic and language clues in the MARC record. Technical report, The University of California at Berkeley, 2004.

T. Qin, R. Xiao, L. Fang, X. Xie, and L. Zhang. An efficient location extraction algorithm by leveraging web contextual information. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 53–60. ACM, 2010.

E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, pages 50–54, 2003.

K. Roberts, C. Bejan, and S. Harabagiu. Toponym disambiguation using events. In *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference*, pages 271–276, 2010.

S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of EMNLP 2012*, 2012.

J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling. TwitterStand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51, 2009.

W. Scheidel, E. Meeks, and J. Weiland. ORBIS: The Stanford geospatial network model of the roman world. 2012.

A. Skupin and A. Esperbé. An alternative map of the United States based on an *n*-dimensional model of geographic space. *Journal of Visual Languages & Computing*, 22(4):290–304, 2011.

D. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 127–136, 2001.

D. Smith and G. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, pages 45–49, 2003.

B. Teitler, M. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: a new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 18. ACM, 2008.

R. Volz, J. Kleb, and W. Mueller. Towards ontology-based disambiguation of geographical identifiers. In *Proceedings of the 16th International Conference on World Wide Web*, 2007.

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

Q. Zhang, P. Jin, S. Lin, and L. Yue. Extracting focused locations for web pages. In *Web-Age Information Management*, volume 7142, pages 76–89. 2012.

W. Zong, D. Wu, A. Sun, E. Lim, and D. Goh. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 354–362, 2005.