A Survey on Geocoding: Algorithms and Datasets for Toponym Resolution

Anonymous ACL submission

Abstract

Geocoding, the task of converting unstructured text to structured spatial data, has recently seen progress thanks to a variety of new datasets, evaluation metrics, and machine-learning algorithms. We provide a survey to review, organize and analyze recent work on geocoding (also known as toponym resolution) where the text is matched to geospatial coordinates and/or ontologies. We summarize the findings of this research and suggest some promising directions for future work.

1 Introduction

Geocoding, also called toponym resolution or toponym disambiguation, is the subtask of geoparsing that disambiguates place names in text. The goal of geocoding is, given a textual mention of a location, to choose the corresponding geospatial coordinates, geospatial polygon, or entry in a geospatial database. Geocoders must handle place names (known as *toponyms*) that refer to more than one geographical location (e.g., *Paris* can refer to a town in the state of *Texas* in the *United States*, or the capital city of *France*), and geographical locations that may be referred to by more than one name (e.g., *Leeuwarden* and *Ljouwert* are two names for the same city in the Netherlands), as shown in fig. 1. Geocoding plays a critical role in tasks such as tracking the evolution and emergence of infectious diseases (Hay et al., 2013), analyzing and searching documents by geography (Bhargava et al., 2017), geospatial analysis of historical events (Tateosian et al., 2017), and disaster response mechanisms (Ashktorab et al., 2014; de Bruijn et al., 2018).

The field of geocoding, previously dominated by geographical information systems communities, has seen a recent surge in interest from the natural language processing community due to the interesting linguistic challenges this task presents. The four most recent geocoding datasets (see table 1) were all published at venues in the ACL Anthology.

Figure 1: An illustrative example of geocoding challenges. One toponym (*Paris*) can refer to more than one geographical location (a town in the state of *Texas* in the *United States* or the capital city of *France* in *Europe*), and a geographical location may be referred to by more than one toponym (*Leeuwarden* and *Ljouwert* are two names for the same city in the Netherlands).

And the recent ACL-SIGLEX sponsored SemEval 2019 Task 12: Toponym Resolution in Scientific Papers (Weissenbacher et al., 2019) resulted in several new natural language processing approaches to geocoding. The field has thus changed substantially since the most recent survey of geocoding (Gritta et al., 2017), including a doubling of the number of geocoding datasets, and the advent of modern neural network approaches to geocoding.

The field would thus benefit from a survey and critical evaluation of the currently available datasets, evaluation metrics, and geocoding algorithms. Our contributions are:

- the first survey on geocoding to include recent deep learning approaches
- coverage of new geocoding datasets (which increased by 100% since 2017) and geocoding systems (which increased by 50% since 2017)
- discussion of new directions, such as polygon-based prediction

In the remainder of this article, we first highlight some previous geocoding surveys (section 2) and explain the scope of the current survey (section 3). We then categorize the features of recent geocod-

ing datasets (section 4), compare different choices for geocoding evaluation metrics (section 5), and break down the different types of features and architectures used by geocoding systems (section 6). We conclude with a discussion of where the field should head next (section 7).

2 Background

An early formal survey of geocoding is Leidner (2007). This Ph.D. thesis distinguished the tasks of finding place names (known as *geotagging* or *toponym recognition*) from linking place names to databases (known as *geocoding* or *toponym resolution*). They found that most geocoding methods were based on combining natural language processing techniques, such as lexical string matching or word sense matching, with geographic heuristics, such as spatial-distance minimum and population maximum. Most geocoders studied in this thesis were rule-based.

Monteiro et al. (2016) surveyed work on predicting document-level geographic scope, which often includes mention-level geocoding as one of its steps. Most of this survey focused on the document-level task, but the geocoding section found techniques similar to those found by Leidner (2007).

Gritta et al. (2017) reviewed both geotagging and geocoding, and proposed a new dataset, WikToR. The survey portion of this article compared datasets for geoparsing, explored heuristics of rule-based and feature-based machine learning-based geocoders, summarized evaluation metrics, and classified common errors from several geocoders (misspellings, case sensitivity, processing fictional and historical text presents, etc.). Gritta et al. (2017) concluded that future geoparsers would need to utilize semantics and context, not just syntax and word forms as the geocoders of the time.

Geocoding research since these previous surveys has changed in several important ways, as will be described in the remainder of this article. Most notably, new datasets and evaluation metrics are enabling new polygon-based views of the problem, and deep learning methods are offering new algorithms and new approaches for geocoding.

3 Scope

We focus on the geocoding problem, where mentions of place names are resolved to database entries or polygons. We thus searched the Google Scholar and Semantic Scholar search engines for papers matching any of the keyword queries: *geocoding*, *geoparsing*, *geolocation*, *toponym resolution*, *toponym disambiguation*, or *spatial information extraxtion*. From the results, we excluded articles that described tasks other than mention-level geocoding, for example:

- matching a full document or full microblog post to a single location (Luo et al., 2020; Hoang and Mothe, 2018; Kumar and Singh, 2019; Lee et al., 2015)
- geographic document retrieval and classification (Gey et al., 2005; Adams and McKenzie, 2018)
- matching typonyms to each other within a geographical database (Santos et al., 2018)

We also excluded papers published before 2010 (e.g., Smith and Crane, 2001), as they have been covered thoroughly by prior surveys.

In total, we reviewed more than 60 papers and included more than 30 of them in this survey.

4 Geocoding Datasets

Many geocoding corpora have been proposed, drawn from different domains, linking to different geographic databases, with different forms of geocoding labels, and with varying sizes in terms of both articles/messages and toponyms. Table 1 cites and summarizes these datasets, and the following sections walk through some of the dimensions over which the datasets vary.

4.1 Domains

The news domain is the most common target for geocoding corpora, covering sources like broadcast conversation, broadcast news, and news magazines. Examples include the ACE 2005 English SpatialML Annotations (ACS), the Local Global Lexicon (LGL), CLUST, TR-NEWS, GeoVirus, and GeoWebNews. Though all these datasets include news text, they vary in what toponyms are included. For example, LGL is based on local and small U.S. news sources with most toponyms smaller than a U.S. state, while GeoVirus focuses on news about global disease outbreaks and epidemics with larger, often country-level, toponyms.

Web text is also a common target for geocoding corpora. Wikipedia Toponym Retrieval (WikToR) and GeoCoDe are both based on Wikipedia pages. ACS, mentioned above, also includes newsgroup and weblog data. And social media, specifically

Corpus	Domain	Geographic Database	Label Type	Articles / Messages	Toponyms
ACS, Mani et al. (2010)	News	GeoNames	Point	428	4783
LGL, Lieberman et al. (2010)	News	GeoNames	Point & GeoNamesID	588	4793
CLUST, Lieberman and Samet (2011)	News	GeoNames	Point & GeoNamesID	1082	11564
ZG, Zhang and Gelernter (2014)	Twitter	GeoNames	Point & GeoNamesID	956	1393
WOTR, DeLozier et al. (2016)	Historical	OpenStreetMap	Point & Polygon	9653	10380
WikTOR, Gritta et al. (2017)	Wikipedia	GeoNames	Point	5000	25000
Prussian , Ardanuy and Sporleder (2017)	Historical	GeoNames	Point	N/A	1529
Belgian, Ardanuy and Sporleder (2017)	Historical	GeoNames	Point	N/A	544
Antilles, Ardanuy and Sporleder (2017)	Historical	GeoNames	Point	N/A	301
EastIndies, Ardanuy and Sporleder (2017)	Historical	GeoNames	Point	N/A	210
DRegional, Ardanuy and Sporleder (2017)	Historical	GeoNames	Point	N/A	1037
TR-NEWS, Kamalloo and Rafiei (2018)	Historical	GeoNames	Point & GeoNamesID	118	1274
GeoCorpora, Wallgrün et al. (2018)	Twitter	GeoNames	Point & GeoNamesID	211	2966
GeoVirus, Gritta et al. (2018)	News	GeoNames	Point	229	2167
GeoWebNews, Gritta et al. (2019)	News	GeoNames	Point & GeoNamesID	200	5121
SemEval-2019-12, Weissenbacher et al. (2019)	Scientific	GeoNames	Point & GeoNamesID	150	8360
GeoCoDe, Laparra and Bethard (2020)	Wikipedia	OpenStreetMap	Polygon	360187	360187

Table 1: Summary of geocoding datasets covered by this survey, sorted by year of creation.

Twitter, is the target for ZG and GeoCorpora. These corpora vary as widely as the internet text upon which they are based. For example, GeoCoDe and WikToR include the first paragraphs of Wikipedia articles, while ZG and GeoCorpora contain Twitter messages with place names that were highly ambiguous and mostly unambiguous, respectively.

Other geocoding domains are less common, but have included areas such as historical documents and scientific journal articles. The Official Records of the War of the Rebellion (WOTR) corpus annotates historical toponyms of the U.S. Civil War. Ardanuy and Sporleder (2017) created 5 historical multi-lingual datesets based on national, regional, local, and colonial historical newspapers. The SemEval-2019 Task 12 dataset is based on scientific journal papers from PubMed Central[1].

4.2 Geographic Databases

All geocoding corpora rely on some database of geographic knowledge, sometimes also called a gazetteer or ontology. Such a database includes canonical names for places along with their geographic attributes such as latitude/longitude or geospatial polygon, and may include other information, such as population or type of place.

Most geocoding corpora have used GeoNames[2] as their geographic database, including ACS, LGL, CLUST, ZG, WikToR, TR-NEWS, GeoCorpora, GeoVirus, GeoWebNews, and SemEval-2019-12. GeoNames is a crowdsourced database of geospatial locations, with almost 7 million entries and a variety of information such as feature type (country, city, river, mountain, etc.), population, elevation, and positions within a political geographic hierarchy. The freely available version of GeoNames contains only a (latitude, longitude) point for each location, with the polygons only available with a premium data subscription, so most corpora based on GeoNames do not use geospatial polygons.

Geocoding corpora where recognizing geospatial polygons is important have typically turned to OpenStreetMap[3]. OpenStreetMap is another crowdsourced database of geospatial locations, which contains both (latitude, longitude) points and geospatial polygons for its locations. WOTR and GeoCoDe are based on OpenStreetMap.

4.3 Geospatial Label Types

Three different types of geospatial labels have been considered in geocoding corpora: database entries, (latitude, longitude) points, and polygons. All corpora except WTOR and GeoCoDe assign to each place name the (latitude, longitude) point that represents its geospatial center on the globe. Many of the GeoNames-based corpora (LGL, CLUST, TR-NEWS, GeoCorpora, GeoWebNews, and SemEval-2019-12) also assign to each place name its GeoNames database ID. The WTOR corpus assigns to each place name a point or a polygon, and GeoCoDe assigns to each place name only a polygon. Figure 2 shows an example of a polygon annotation from GeoCoDe.

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[2] https://www.geonames.org/

[3] https://www.openstreetmap.org/

Figure 2: The red-shaded area is the polygon label for *Biancavilla*, which is defined by the set of its boundary coordinates retrieved from OpenStreetMap.

4.4 Analysis: Geocoding Datasets

The most compelling improvements in geocoding datasets have been in the variety of domains, moving from exclusively news to include historical documents, scientific documents, Wikipedia, and social media. Less change has been seen in geographic databases, where GeoNames is still dominant over OpenStreetMap, and in geospatial label types, where points are still dominant over polygons. These latter two issues are intertwined: GeoNames polygons are only available for a fee, while OpenStreetMap polygons are freely available.

5 Geocoding Evaluation Metrics

Geocoding systems are evaluated on geocoding corpora using metrics that depend on the corpus's geospatial label type.

5.1 Database entry correctness metrics

When the target label type is a geospatial database entry ID, common evaluation metrics for multi-class classification tasks are applied. These metrics can also be used for corpora with (latitude, longitude) point labels by breaking the globe down into a discrete grid of geospatial tiles, and treating each geospatial tile like a database entry.

Accuracy is the number of place names where the system has predicted the correct database entry, divided by the number of place names. Accuracy is sometimes also called *Precision@1* or *P@1* when there is only one correct answer (as in the case for current geocoding datasets) and when the ranking-based system is turned into a classifier by taking the top-ranked result as its prediction (the current standard for geocoding evaluation).

$$Accuracy = \frac{|\hat{U}|}{|U|}$$

where U is the set of human-annotated place names, $\hat{U}$ is the set of place names where the system's single prediction or top-1 ranked result is correct.

5.2 Point distance metrics

When the target label type is a (latitude, longitude) point, common evaluation metrics attempt to measure the distance between the system-predicted point and the human-annotated point.

Mean error distance calculates the mean over all predictions of the distance between each system-predicted and human-annotated point:

$$MeanErrorDist = \frac{\sum_{u \in U} dis(l_s(u), l_h(u))}{|U|}$$

where U is the set of all human-annotated place names, $l_s(u)$ is the system-predicted (latitude, longitude) point for place name u, $l_h(u)$ is the human-annotated (latitude, longitude) point for place name u, and dis is the distance between the two points on the surface of the globe.

Median Error Distance is defined in a similar way to mean error distance, but takes the median of the error distances rather than the mean.

Accuracy@k km/miles measures the fraction of system-predicted (latitude, longitude) points that were less than k km/miles away from the human-annotated (latitude, longitude) points. Formally:

$$Acc@k = \frac{|\{u | u \in U \wedge dis(l_s(u), l_h(u)) <= k\}|}{|U|}$$

where U, l_s, l_h, and dis are defined as above, and k is a hyper-parameter. A common choice for k is 161 (Cheng et al., 2010).

Area Under the Curve (AUC) calculates the area under the curve of the distribution of geocoding error distances. A geocoding system is better if the area under the curve is smaller. Formally:

$$AUC = \ln \frac{ActualErrorDistance}{MaxPossibleErrors}$$

where $ActualErrorDistance$ is the area under the curve, and $MaxPossibleErrors$ is the farthest distance between two places on earth.

5.3 Polygon-based metrics

When the target label type is a polygon, evaluation metrics attempt to compare the overlap between the system-predicted polygon and the human-annotated polygon.

GeoCoder	Implementation	Prediction Type	Database Independent	Polygon based
Edinburgh Parser (Grover et al., 2010)	Rule-based	Ranking	No	No
TGBRW-2010 (Tobin et al., 2010)	Rule-based	Ranking	No	No
MAC-2010 (Martins et al., 2010)	Machine Learning	Ranking	No	No
IGeo (Lieberman et al., 2010)	Rule-based	Ranking	No	No
LS-2011 (Lieberman and Samet, 2011)	Rule-based	Ranking	No	No
MG (Freire et al., 2011)	Machine Learning	Ranking	No	No
CLAVIN (Berico Technologies, 2012)	Rule-based	Ranking	No	No
LS-2012 (Lieberman and Samet, 2012)	Machine Learning	Ranking	No	No
GeoTxt (Karimzadeh et al., 2013)	Rule-based	Ranking	No	No
SPIDER (Speriosu and Baldridge, 2013)	Machine Learning	Ranking	No	No
WISTR (Speriosu and Baldridge, 2013)	Machine Learning	Ranking	No	No
TRAWL (Speriosu and Baldridge, 2013)	Machine Learning	Ranking	No	No
CMU-Geolocator (Zhang and Gelernter, 2014)	Machine Learning	Ranking	No	No
SMFCM-2015 (Santos et al., 2015)	Machine Learning	Ranking	No	No
Topocluster (DeLozier et al., 2015)	Machine Learning	Classification	Yes	No
GeoSem (Ardanuy and Sporleder, 2017)	Machine Learning	Ranking	No	No
CBH, SHS Kamalloo and Rafiei (2018)	Machine Learning	Ranking	No	No
CamCoder (Gritta et al., 2018)	Deep Learning	Classification	No	No
DM_NLP (Wang et al., 2019)	Machine Learning	Ranking	No	No
CME-2019 (Cardoso et al., 2019)	Deep Learning	Classification & Regression	Yes	No
MLG (Kulkarni et al., 2020)	Deep Learning	Classification	Yes	No
LB-2020 (Laparra and Bethard, 2020)	Rule-based	Regression	Yes	Yes

Table 2: Summary of geocoding systems covered by this survey, sorted by year of creation.

Polygon-based precision and recall were proposed by Laparra and Bethard (2020) based on the intersection of system-predicted and human-annotated geometries. Formally:

$$Precision = \frac{1}{|S|} \sum_{i \in |S|} \frac{area(S_i \cap H_i)}{area(S_i)}$$

$$Recall = \frac{1}{|H|} \sum_{i \in |H|} \frac{area(S_i \cap H_i)}{area(H_i)}$$

where the S is the system-predicted set of polygons and H is the human-annotated set of polygons.

5.4 Analysis: Geocoding Evaluation Metrics

In point-based metrics, median error distance is generally preferred to mean error distance, as the latter is sensitive to outliers. For example, Gritta et al. (2017) found that the bulk of errors are triggered by roughly 20% of the places and the errors from the remaining places are relatively low. AUC is generally preferred to *Accuracy@k km/miles* because in AUC, the difference between two small errors (such as 10 and 20 km) is more significant than the same difference between two large errors (such as 110 and 120 km) (Jurgens et al., 2015).

Polygon-based metrics have so far only been applied to datasets with polygon labels, but future work should consider applying them to datasets with database entry labels. This could give credit when two database entries are equally applicable (e.g., a mention of *Dallas* that is ambiguous between city and county) and the polygons overlap (e.g., Dallas city, GeoNames ID 4684888, makes up most of Dallas county, GeoNames ID 4684904).

6 Geocoding Systems

Table 2 summarizes the approaches of geocoders over the last decade. These models have different approaches to the prediction problem, ranging from ranking to classification to regression. They implement their predictive models with technology ranging from hand-constructed rules and heuristics, to feature-based machine-learning models, to deep learning (i.e., neural network) models that learn their own features.

6.1 Prediction Types

Ranking is the most common approach to making geospatial predictions (Edinburgh Parser, TGBRW-2010, MAC-2010, IGeo, LS-2011, MG, CLAVIN, LS-2012, WISTR, GeoTxt, CMU-Geolocator, SMFCM-2015, GeoSem, CBH, SHS, DM_NLP). For example, most rule-based systems index their geospatial database with a search system like Lucene (`https://lucene.apache.org/`), and query that index to produce a ranked list of candidate database entries. This ranked list may be further re-ranked based on other features such as population or proximity. The type of scores using in re-ranking include binary classification score (MG, LS-2012, WISTR, CMU-Geolocator, CBH,

SHS, DM_NLP), regression distance MAC-2010, the precision at the first position of the ranked list SMFCM-2015, and heuristics based on information in the geospatial database (Edinburgh Parser, TGBRW-2010, IGeo, LS-2011, CLAVIN, GeoTxt).

Classification is commonly used in making geospatial predictions when the Earth's surface has been discretized into tiny areas (Topocluster, CamCoder, CME-2019, MLG). For example, *CamCoder* divides the Earth's surface into 7,823 tiles, and then changes the geospatial label of each toponym to the tile containing its coordinate. *CamCoder* then directly predicts one of 7823 classes for each toponym mention.

Regression is sometimes used for geospatial predictions when the label type is a (latitude, longitude) point or a polygon (CME-2019, LB-2020). For example, LB-2020 predict a set of coordinates (i.e., a polygon) by applying operations over reference geometries, where the operations take sets of coordinates as inputs and produce sets of coordinates as outputs. Regression approaches to geocoding are rare because directly predicting coordinates over the entire surface of the Earth is challenging.

6.2 Features and Heuristics

All geocoding systems combine string matching (exact string matching, Levenshtein distance, etc.) with other features and/or heuristics (population, words in nearby context, etc.). Details of such features are described in this section.

String match checks whether the place name matches any names in the geospatial database (Edinburgh Parser, TGBRW-2010, MAC-2010, IGeo, LS-2011, MG, CLAVIN, GeoTxt, CMU-Geolocator, SMFCM-2015, GeoSem, CBH, SHS, DM_NLP). String matching can be done exactly, or approximately with edit distance metrics like Levenshtein Distance. For example, GeoTxt calculates the Levenshtein Distance between the place name in the text and each candidate entry from the geospatial database, and selects the candidate with the lowest edit distance.

Population looks at the size of the population associated with candidate database entry, typically preferring more populous entries to less populous ones (Edinburgh Parser, TGBRW-2010, MAC-2010, IGeo, LS-2011, MG, LS-2012, CLAVIN, GeoTxt, CMU-Geolocator, SMFCM-2015, CBH, SHS, CamCoder, DM_NLP). For example, when the Edinburgh Parser geocodes the text *I love Paris*, it resolves *Paris* to PARIS, FRANCE instead of PARIS, TX, U.S. since the former has a greater population in the geospatial database.

Type of place looks at the geospatial feature type (country, city, river, populated place, facility, etc.) of a candidate database entry, typically preferring the more geographically prominent ones (Edinburgh Parser, TGBRW-2010, MAC-2010, IGeo, LS-2011, MG, CLAVIN, LS-2012, GeoTxt, TRAWL, CMU-Geolocator, SMFCM-2015, GeoSem, CBH, SHS, DM_NLP). For example, TGBRW-2010 prefers "populated places" to "facilities" such as farms and mines, when there are multiple candidate geospatial labels.

Words in the nearby context are used to disambiguate ambiguous place names (LS-2012, WISTR, CMU-Geolocator, SMFCM-2015, Topocluster, GeoSem, CBH, SHS, DM_NLP, CamCoder, CME-2019, MLG). Ways of using context words range from simple to complex. For example, WISTR uses a context window of 20 words on each side of the target place name, aiming to benefit from location-oriented words such as *uptown* and *beach*. In contrast, CMU-Geolocator searches for common country and state names in other nearby location expressions, using these mostly unambiguous place names to help resolve the target place name.

One sense per referent is a heuristic that assumes that all occurrences of a unique place name in the same document will refer to the same geographical database entry (Edinburgh Parser, TGBRW-2010, IGeo, LS-2011, GeoTxt, CBH, SHS, DM_NLP). For example, after each time that IGeo resolves a place name to a geospatial label, it propagates the same resolution to all identical place names in the remainder of the document.

Spatial minimality is a heuristic that assumes that place names in a text tend to refer to geospatial regions that are in close spatial proximity to each other (Edinburgh Parser, TGBRW-2010, IGeo, LS-2011, CLAVIN, SPIDER, Topocluster, GeoSem, CBH, SHS). For example, when IGeo geocodes the text *96 miles south of Phoenix, Arizona, just outside of Tucson*, it takes *Tucson* as an "anchor" toponym and resolves that first to get a target region. Then for *Phoenix*, it selects the geospatial label that is most geographically proximate to the target region.

6.3 Implementation Types

Rule-based systems use hand-crafted rules and heuristics to predict a geospatial label for a place name (Edinburgh Parser, TGBRW-2010, IGeo, LS-2011, CLAVIN, GeoTxt, LB-2020). The rule bases range in size from 2 to more than 200 rules, and rules may be formalized in rule grammars or defined more informally and provided as code. For example, IGeo uses a rule defined via code to identify place names in comma groups (e.g., "New York, Chicago and Los Angeles", all major cities in the U.S.), and then resolves all toponyms by applying a heuristic uniformly across the entire group. As another example, LB-2020 uses 219 synchronous grammar rules to parse a target polygon from reference polygons by constructing a tree of geometric operators (e.g., $\textsc{Between}(p_1, p_2)$ calculates the region between geolocation polygons p_1 and p_2).

Feature-based machine-learning systems use many of the same features and heuristics of rule-based systems, but provide these as input to a supervised classifier that makes the prediction of a geospatial label (MAC-2010, MG, LS-2012, WISTR, CMU-Geolocator, SMFCM-2015, Topocluster, GeoSem, CBH, SHS, DM_NLP). They typically operate in a two-step rank-then-rerank framework, where first an information retrieval system produces candidate geospatial labels, then a supervised machine-learning model produces a score for each candidate, and the candidates are reranked by these scores. Classification and ranking algorithms include logistic regression (WISTR), support vector machines (MAC-2010, CMU-Geolocator), random forests (MG, LS-2012), stacked LightGBMs (DM_NLP), and LambdaMART (SMFCM-2015). For example, MAC-2010 trains a support vector machine regression model using features such as the population and the number of alternative names for each candidate.

Deep learning systems often approach geocoding as a one-step classification problem by dividing the Earth's surface into an $N \times N$ grid, where the neural network attempts to map place names and their features to one of these $N \times N$ categories (CamCoder, CME-2019, MLG). Each system has a unique neural architecture for combining inputs to make predictions, typically based on either convolutional neural networks (CNNs) or recurrent neural networks (RNNs).

CamCoder was the first deep learning based-geocoder. Its lexical model uses CNNs to create vectors representing context words (a window of 200 words, location mentions excluded), location mentions (context words excluded) and the target place name. Its geospatial model produces a vector using a geospatial label's population (from the database) as its prior probability. CamCoder concatenates the lexical and geospatial vectors for the final classification.

MLG is also a CNN-based geocoder, but it does not use population or other geospatial database information. It captures lexical features in a similar manner to CamCoder, but takes advantage of the S2 geometry (`https://s2geometry.io/`) to represent its geospatial output space in hierarchical grid-cells from coarse to fine-grained. MLG can predict the geospatial label of a place name at multiple S2 levels by mutually maximizing both precision and generalization of predictions.

CME-2019 is an RNN-based geocoder that uses HEALPix geometry (Gorski et al., 2005) to discretize the Earth's surface. It uses long short-term memory network with pre-trained Elmo embeddings (Peters et al., 2018) to create vectors representing the place name, local context (50 words around the place name), and larger context (paragraph or 500 words around the place name). The three vectors are concatenated and used to predict both the class of the HEALPix region and the coordinates of the centroid of the HEALPix class. This joint learning approach allows the two tasks to be mutually promoted and restricted.

6.4 Analysis: Geocoding Systems

Despite much variability in choice of evaluation dataset, the LGL, WikTOR, GeoVirus, and WOTR datasets have been shared by multiple geocoders, so we summarize the reported results in table 3. Neural network models (CamCoder, CME-2019, and MLG) perform only slightly better than prior models on LGL and GeoVirus. (Though CME-2019 has larger gains LGL, it doesn't evaluate on GeoVirus). Neural network models achieve larger gains on WikTOR and WOTR, likely because these larger datasets (10,000+ toponyms) provide more training data to the data-hungry neural networks. WikTOR was also specifically designed to counteract the population heuristic popular in prior models, and WOTR's narrow domain (American Civil War military reports) likely has a similar effect.

While the advent of recent deep learning ap-

GeoCoder	Accuracy@161km (↑)				Mean error distance (↓)			
	LGL	GeoVirus	WikTOR	WOTR	LGL	GeoVirus	WikTOR	WOTR
Edinburgh Parser (Grover et al., 2010)	76	78	42	-	8	5	31	-
CLAVIN (Berico Technologies, 2012)	71	79	16	-	13	6	43	-
GeoTxt (Karimzadeh et al., 2013)	68	79	18	-	14	6	47	-
SPIDER (Speriosu and Baldridge, 2013)	68	-	-	67	12	-	-	4.8
SMFCM-2015 (Santos et al., 2015)	71	-	-	-	8	-	-	-
Topocluster (DeLozier et al., 2015)	63	-	26	-	12	-	38	-
GeoSem (Ardanuy and Sporleder, 2017)	-	-	-	68	-	-	-	4.5
CamCoder (Gritta et al., 2018)	76	82	65	-	7	3	11	-
CME-2019 (Cardoso et al., 2019)	86	-	-	82	2.4	-	-	1.6
MLG (Kulkarni et al., 2020)	73	85	85	-	6.2	2.8	3.5	-

Table 3: Reported results on LGL, WikToR, GeoVirus, and WOTR. For accuracy@161km, larger is better (↑). For mean error distance, smaller is better (↓).

proaches is an exciting step forward for geocoding research, most such models include only a few of the many features investigated by feature-based architectures. For example, no deep learning models yet incorporate document-level consistency features like *one sense per referent*, geospatial consistency features like *spatial minimality*, or database information beyond population.

7 Future Directions

A key direction of future research will be output representations. Many past geocoders focused on mapping place names to geospatial database entries (see column 4 of table 2). This was convenient, enabling fast resolution by applying standard information retrieval models to propose candidate entries from the database, but was limited by the simple types of matching that information retrieval systems could perform. Modern deep learning approaches to geocoding allow more complex matching of place names to geospatial locations, but typically rely on discretizing the Earth's surface into tiles to constrain the size of the network's output space. For the neural networks to achieve the fine-grained level of geocoding available in geocoding databases, they may need to consider hierarchical output spaces (e.g., Kulkarni et al., 2020) or compositional output spaces (e.g., Laparra and Bethard, 2020) that can express the necessary level of detail without exploding the output space.

Another key direction of future research will be the structure and evaluation of geocoding datasets. Most existing datasets and systems treat geocoding as a problem of identifying points rather than polygons (see column 4 of table 1 and column 5 of table 2). Yet the vast majority of real places in geospatial databases are complex polygons (as in fig. 2), not simple points. More polygon-based datasets are needed, especially ones like GeoCoDe (Laparra and Bethard, 2020) that include complex descriptions of locations (e.g., *between the towns of Adrano and S. Maria di Licodia*) and not just explicit place names (e.g., *Paris*). The current state-of-the-art for complex geographical description geocoding is rule-based, but more polygon-based datasets will drive algorithmic research that can improve upon these rule-based systems with some of the insights gained from deep neural network approaches to explicit place name geocoding.

Finally, geocoding evaluation is still an open research area. Future research will likely extend some of the new polygon-based evaluation metrics. For example, using polygon precision and recall would give credit to a geocoding system that predicted the GeoNames entry *Nakhon Sawan* even if the annotated data used the entry *Changwat Nakhon Sawan*, since the polygons of these two place names are nearly identical.

8 Conclusion

After surveying a decade of work on geocoding, we have identifed several trends. First, combining contextual features with geospatial database information makes geocoders more powerful. Second, like much of NLP, geocoders have moved from rule-based systems to feature-based machine-learning systems to deep-learning systems. Third, the older rank-then-rerank approaches, combining information retrieval and supervised classification, are being replaced by direct classification approaches, where the Earth's surface is discretized into many small tiles. Finally, the field of geocoding is just beginning to look beyond a point-based view of locations to a more realistic polygon-based view.

References

Benjamin Adams and Grant McKenzie. 2018. Crowdsourcing the character of a place: Character-level convolutional networks for multilingual geographic text classification. *Transactions in GIS*, 22(2):394–408.

Mariona Coll Ardanuy and Caroline Sporleder. 2017. Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 175–180.

Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. Tweedr: Mining twitter to inform disaster response. In *ISCRAM*, pages 269–272.

Berico Technologies. 2012. Cartographic location and vicinity indexer (clavin).

Preeti Bhargava, Nemanja Spasojevic, and Guoning Hu. 2017. Lithium NLP: A system for rich information extraction from noisy user generated text on social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 131–139, Copenhagen, Denmark. Association for Computational Linguistics.

Jens A de Bruijn, Hans de Moel, Brenden Jongman, Jurjen Wagemaker, and Jeroen CJH Aerts. 2018. Taggs: grouping tweets to improve global geoparsing for disaster response. *Journal of Geovisualization and Spatial Analysis*, 2(1):2.

Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2019. Using recurrent neural networks for toponym resolution in text. In *EPIA Conference on Artificial Intelligence*, pages 769–780. Springer.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768.

Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2382–2388. AAAI Press.

Grant DeLozier, Ben Wing, Jason Baldridge, and Scott Nesbit. 2016. Creating a novel geolocation corpus from historical texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 188–198, Berlin, Germany. Association for Computational Linguistics.

Nuno Freire, José Borbinha, Pável Calado, and Bruno Martins. 2011. A metadata geoparsing system for place name recognition and resolution in metadata records. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 339–348.

Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. 2005. Geoclef: the clef 2005 cross-language geographic information retrieval track overview. In *Workshop of the cross-language evaluation forum for european languages*, pages 908–919. Springer.

Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. 2005. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Which Melbourne? augmenting geocoding with maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1285–1296, Melbourne, Australia. Association for Computational Linguistics.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2019. A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, pages 1–30.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623.

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

Simon I Hay, Katherine E Battle, David M Pigott, David L Smith, Catherine L Moyes, Samir Bhatt, John S Brownstein, Nigel Collier, Monica F Myers, Dylan B George, et al. 2013. Global mapping of infectious disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120250.

Thi Bich Ngoc Hoang and Josiane Mothe. 2018. Location extraction from tweets. *Information Processing & Management*, 54(2):129–144.

David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Ninth international AAAI conference on web and social media*.

Ehsan Kamalloo and Davood Rafiei. 2018. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1287–1296.

Morteza Karimzadeh, Wenyi Huang, Siddhartha Banerjee, Jan Oliver Wallgrün, Frank Hardisty, Scott Pezanowski, Prasenjit Mitra, and Alan M MacEachren. 2013. Geotxt: a web api to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval*, pages 72–73.

Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, Eugene Ie, and Li Zhang. 2020. Spatial language representation with multi-level geocoding. *arXiv preprint arXiv:2008.09236*.

Abhinav Kumar and Jyoti Prakash Singh. 2019. Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction*, 33:365–375.

Egoitz Laparra and Steven Bethard. 2020. A dataset and evaluation framework for complex geographical description parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 936–948, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sunshin Lee, Mohamed Farag, Tarek Kanan, and Edward A Fox. 2015. Read between the lines: A machine learning approach for disambiguating the geo-location of tweets. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 273–274.

JL Leidner. 2007. *Toponym resolution: A comparison and taxonomy of heuristics and methods*. Ph.D. thesis, PhD Thesis, University of Edinburgh.

Michael D Lieberman and Hanan Samet. 2011. Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 843–852.

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

Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)*, pages 201–212. IEEE.

Xiangyang Luo, Yaqiong Qiao, Chenliang Li, Jiangtao Ma, and Yimin Liu. 2020. An overview of microblog user geolocation methods. *Information Processing & Management*, 57(6):102375.

Inderjeet Mani, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy. 2010. Spatialml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280.

Bruno Martins, Ivo Anastácio, and Pável Calado. 2010. A machine learning approach for resolving place references in text. In *Geospatial thinking*, pages 221–236. Springer.

Bruno R Monteiro, Clodoveu A Davis Jr, and Fred Fonseca. 2016. A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96:23–34.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

João Santos, Ivo Anastácio, and Bruno Martins. 2015. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80(3):375–392.

Rui Santos, Patricia Murrieta-Flores, Pável Calado, and Bruno Martins. 2018. Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, 32(2):324–348.

David A Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *International Conference on Theory and Practice of Digital Libraries*, pages 127–136. Springer.

Michael Speriosu and Jason Baldridge. 2013. Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1476.

Laura Tateosian, Rachael Guenter, Yi-Peng Yang, and Jean Ristaino. 2017. Tracking 19th century late blight from archival documents using text analytics and geoparsing. In *Free and open source software for geospatial (FOSS4G) conference proceedings*, volume 17, page 17.

Richard Tobin, Claire Grover, Kate Byrne, James Reid, and Jo Walsh. 2010. Evaluation of georeferencing. In *proceedings of the 6th workshop on geographic information retrieval*, pages 1–8.

Jan Oliver Wallgrün, Morteza Karimzadeh, Alan M MacEachren, and Scott Pezanowski. 2018. Geocorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32(1):1–29.

Xiaobin Wang, Chunping Ma, Huafei Zheng, Chu Liu, Pengjun Xie, Linlin Li, and Luo Si. 2019. Dm_nlp at semeval-2018 task 12: A pipeline system for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 917–923.

Davy Weissenbacher, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2019. SemEval-2019 task 12: Toponym resolution

in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 907–916, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Wei Zhang and Judith Gelernter. 2014. Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70.