



A survey on geocoding: algorithms and datasets for toponym resolution

Zeyu Zhang¹ · Steven Bethard¹

Accepted: 23 February 2024

© The Author(s), under exclusive licence to Springer Nature B.V. 2024

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. Geocoding plays a critical role in tasks such as tracking the evolution and emergence of infectious diseases, analyzing and searching documents by geography, geospatial analysis of historical events, and disaster response mechanisms. To assist those new to this area of research, we provide a survey that reviews, organizes and analyzes recent work on geocoding (also known as toponym resolution) where text is matched to geospatial coordinates and/or ontologies. We summarize the findings of this research, including the domains and databases covered by current geocoding corpora, point-based and polygon-based evaluation metrics, and features and architectures of geocoding systems.

Keywords Geocoding · Geographical entity normalization · Toponym resolution

1 Introduction

Geocoding, also called toponym resolution or toponym disambiguation, is the sub-task 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

✉ Zeyu Zhang
zeyuzhang@arizona.edu

Steven Bethard
bethard@arizona.edu

¹ School of Information, University of Arizona, 1103 E. 2nd St., Tucson, AZ 85721, USA

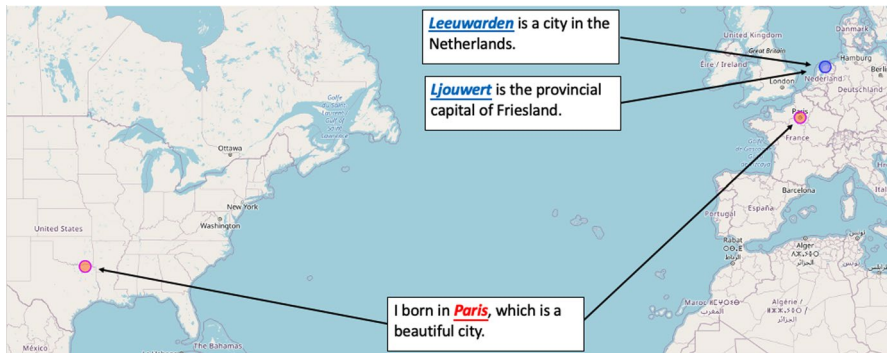


Fig. 1 An illustrative example of geocoding challenges. One toponym (*Paris*) can refer to more than one geographical location (a town in the state of *Texas* in the *United States* or the capital city of *France* in *Europe*), and a geographical location may be referred to by more than one toponym (*Leeuwarden*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

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. 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.

Those new to this area of research 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 (Sect. 2) and explain the scope of the current survey (Sect. 3). We then categorize the features of recent geocoding datasets (Sect. 5), compare different choices for geocoding evaluation metrics (Sect. 6), and break down the different types of features and architectures used by geocoding systems (Sect. 7). We conclude with a discussion of where the field should head next (Sect. 8).

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

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 & Samet, 2011)	News	GeoNames	Point & GeoNamesID	1082	11564
TUD-Loc-2013 (Katz & Schill, 2013)	Web	GeoNames	Point & GeoNamesID	152	3814
ZG (Zhang & Gelernter, 2014)	Twitter	GeoNames	Point & GeoNamesID	956	1393
WOTR (DeLozier et al., 2016)	Historical	OpenStreetMap	Point & Polygon	9653	10380
CLDW (Rayson et al., 2017)	Writing	Unlock	Point	80	–
WIKTOR (Gritta et al., 2017)	Wikipedia	GeoNames	Point	5000	25000
Prussian (Ardanuy & Sporleder, 2017)	Historical	GeoNames	Point	N/A	1529
Belgian (Ardanuy & Sporleder, 2017)	Historical	GeoNames	Point	N/A	544
Antilles (Ardanuy & Sporleder, 2017)	Historical	GeoNames	Point	N/A	301
EastIndies (Ardanuy & Sporleder, 2017)	Historical	GeoNames	Point	N/A	210
DRRegional (Ardanuy & Sporleder, 2017)	Historical	GeoNames	Point	N/A	1037
TR-NEWS (Kamalloo & 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., 2020)	News	GeoNames	Point & GeoNamesID	200	5121
SemEval-2019-12 (Weissenbacher et al., 2019)	Scientific	GeoNames	Point & GeoNamesID	150	8360
GeoCoDe (Laparra & Bethard, 2020)	Wikipedia	OpenStreetMap	Polygon	360187	360187
TopRes19th (Ardanuy et al., 2022)	News	Wikipedia	Point	343	3364

2 Related works

An early formal survey of geocoding is Leidner (2007), which distinguished 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.

Leidner (2021) reviewed many geospatial information processing tasks, but discussed only two geocoding systems in its section on geocoding.

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 Article inclusion criteria

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 extraction*. From the results, we excluded articles that described tasks other than mention-level geocoding, for example:

- Matching an entire document or microblog post to a single location (Luo et al., 2020; Hoang & Mothe, 2018; Kumar & Singh, 2019; Lee et al., 2015; Melo & Martins, 2017), as in geographic document retrieval and classification (Gey et al., 2005; Adams & McKenzie, 2018)
- Matching typonyms to each other within a geographical database (Santos et al., 2018)

- Location name recognition (geotagging) (Chen et al., 2022)

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 Overview of the survey

The survey is divided into three parts: geocoding datasets, geocoding evaluation metrics, and geocoding systems. In each part, we break down the relevant research to reveal the most common features shared across different research efforts and analyze the challenges and opportunities presented.

For geocoding datasets, we find that recent advances have led to an increased variety of domains, while the available geographic databases and geospatial label types have changed little. GeoNames remains the dominant geographic database, and point-based labels dominate over polygons. The availability of free polygon data on OpenStreetMap presents an opportunity to create new datasets that emphasize polygons over points.

For evaluation metrics, median error distance is preferred over mean error distance, and area under the curve of geocoding error distances (AUC) is favored over Accuracy@161 km. Yet these point-based metrics ignore the size and shape of geographic locations, while polygon-based metrics represent an opportunity to more carefully evaluate geocoding systems.

For geocoding systems, features like string matching and population are included in most systems regardless of whether they treat the problem as ranking or classification or whether they use deep neural networks or more traditional machine learning algorithms. Variability in selection of evaluation datasets makes direct comparison across systems difficult, but several systems have reported results on the LGL, WikTOR, GeoVirus, and WOTR datasets. These results generally show that deep neural network models outperform more traditional machine learning algorithms. The neural network models typically incorporate fewer features (e.g., having limited notion of spatial distance), thus there is an opportunity to design deep learning architectures that can incorporate such features.

The remainder of this survey elaborates on these findings in detail.

5 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.

5.1 Domains

The news domain is the most common target for geocoding corpora, covering sources like broadcast conversation, broadcast news, news magazines, and newspapers. Examples include the ACE 2005 English SpatialML Annotations (ACS), the Local Global Lexicon (LGL), CLUST, TR-NEWS, GeoVirus, GeoWebNews, and TopRes19th. 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 Twitter, is the target for ZG and GeoCorpora. TUD-Loc-2013 contains a variety of webpages including news articles and blogs. 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 datasets based on national, regional, local, and colonial historical newspapers. CLDW contains historical writings about the English Lake District in the early seventeenth and early twentieth centuries. The SemEval-2019 Task 12 dataset is based on scientific journal papers from PubMed Central.¹

5.2 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 GeoNames² 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

¹ <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>

² <https://www.geonames.org/>

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



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.

Wikipedia and Unlock⁴ have also been utilized although they are less common geographic databases. For example, in TopRes19th, the toponyms are annotated with the link to the corresponding Wikipedia entries, which can be used to obtain the geographic coordinates of the locations through their URLs.

5.3 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. Many of the GeoNames-based corpora (LGL, CLUST, TUD-Loc-2013, 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.

5.4 Challenges: geocoding datasets

While there have been significant improvements in geocoding datasets, the community has not successfully pivoted from point-based labels to the more precise representation of geographic areas as polygons. This is due primarily to the dominance of GeoNames as a geographic database. GeoNames provides polygons only for a fee,

³ <https://www.openstreetmap.org/>

⁴ <https://groups.inf.ed.ac.uk/geoparser/documentation/v1.1/epub/unlock.html>

creating a barrier for individuals and organizations that that would like to pursue polygon-based geocoding research.

An additional challenge is associative toponyms, such as *Canadian* or *Russian*. Associative toponyms are included in many geocoding datasets, such as LGL, GWN, and TR-News, but the geographic databases include only literal toponyms (e.g., *Canada* or *Russia*). Resolving such toponyms will thus be more difficult, especially when their demonymic forms diverge from their names (e.g., *Netherlands* vs. *Dutch*).

5.5 Opportunities: geocoding datasets

An opportunity for future research on geocoding datasets is to pivot to polygon based labels, which can more faithfully represent complex regions. OpenStreetMap, though used less widely in geocoding research to date, offers free polygon data, and thus provides an opportunity to design new polygon-based geocoding datasets that are not limited by GeoNames fees. Such datasets would allow the development of geocoding systems that better reflect the geography of the world.

Another opportunity in geocoding is to take advantage of the increased variety of domains now available, including historical documents, scientific documents, Wikipedia, and social media. Most work to date has focused on a single one of these domains, meaning there is a need to develop approaches to unify the various datasets, allowing more general and robust geocoding systems to be trained.

6 Geocoding evaluation metrics

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

6.1 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.

6.2 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:

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

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:

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

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 km \approx 100 miles (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:

$$AUC = \ln \frac{Actual\ Error\ Distanc}{Max\ Possible\ Errors}$$

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

6.3 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.

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.

6.4 Challenges: geocoding evaluation metrics

Some challenges exist with specific metrics. A challenge of using mean error distance is its sensitivity to outliers: a few locations with large errors can skew the results and obscure the accuracy of the majority of locations. For instance, Gritta et al. (2017) found that roughly 20% of the places caused most of the errors. A challenge of using *Accuracy@k km/miles* is that it weights small and large errors equally, which may not properly reflect the expectations of users of geocoding systems.

A challenge for all point-based evaluation metrics is that locations are not points on the globe, but regions, and thus the point-based evaluation metrics that are currently popular do a poor job of measuring the actual shapes predicted by geocoding systems.

6.5 Opportunities: geocoding evaluation metrics

For the metrics with specific challenges, alternative metrics have been defined and could be used more widely in future research. Median error distance is similar to mean error distance, but is more robust to outliers. AUC is similar to *Accuracy@k km/miles*, but it gives more weight to smaller errors, which are often more significant than larger errors in practical applications (Jurgens et al., 2015).

A larger opportunity in geocoding evaluation is the application of polygon-based metrics. While to date such metrics have been applied only to one polygon-based dataset, polygon-based metrics could also be applied to datasets with database entry labels. This would give credit to geocoding systems when two or more database entries are equally applicable, such as a mention of "Dallas" which is ambiguous between city and county, and where the polygons of both choices overlap. By considering the overlap of

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

GeoCoder	Method Type	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 & 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 & Samet, 2012)	Machine Learning	Ranking	No	No
GeoTxt (Karimzadeh et al., 2013)	Rule-based	Ranking	No	No
SPIDER (Speriosu & Baldrige, 2013)	Machine Learning	Ranking	No	No
WISTR (Speriosu & Baldrige, 2013)	Machine Learning	Ranking	No	No
TRAWL (Speriosu & Baldrige, 2013)	Machine Learning	Ranking	No	No
CMU-Geolocator (Zhang & 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 & Sportleder, 2017)	Machine Learning	Ranking	No	No
CBH, SHS (Kamalloo & Rafiei, 2018)	Machine Learning	Ranking	No	No
CamCoder (Gritta et al., 2018)	Deep Learning	Classification	No	No
HIS-2019 (Ardanuy et al., 2019)	Rule-based	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
RS-2020 (Aldana-Bobadilla et al., 2020)	Rule-based	Classification	No	No
MLG (Kulkarni et al., 2020)	Deep Learning	Classification	Yes	No
LB-2020 (Laparra & Bethard, 2020)	Rule-based	Regression	Yes	Yes

Table 2 (continued)

GeoCoder	Method Type	Prediction Type	Database Independent	Polygon based
DeezyMatch (Ardanuy et al., 2020)	Deep Learning	Classification	No	No
Bi-LSTM (Fize et al., 2021)	Deep Learning	Regression	Yes	No
LGGeoCoder (Yan et al., 2021)	Deep Learning	Classification	No	No
TR-2022 (Cardoso et al., 2022)	Deep Learning	Classification	Yes	No
Voting (Hu et al., 2023)	Deep Learning	Ensemble	No	No
GeoNorm (Zhang and Bethard, 2023)	Deep Learning	Ranking	No	No

polygons, polygon-based metrics could provide a more precise evaluation of geocoding performance in such cases.

7 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.

7.1 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, RS-2020, GeoNorm). 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, HIS-2019, CME-2019, MLG, DeezyMatch, TR-2022, LGGeoCoder). 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, Bi-LSTM). 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.

7.2 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-Geocator, SMFCM-2015, GeoSem, CBH, SHS, DM_NLP, HIS-2019, RS-2020, DeezyMatch, TR-2022, Bi-LSTM, GeoNorm). 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-Geocator, SMFCM-2015, CBH, SHS, CamCoder, DM_NLP, GeoNorm). 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-Geocator, SMFCM-2015, GeoSem, CBH, SHS, DM_NLP, TR-2022, GeoNorm). 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-Geocator, SMFCM-2015, Topocluster, GeoSem, CBH, SHS, DM_NLP, CamCoder, CME-2019, MLG, LGGeoCoder, TR-2022, GeoNorm). 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-Geocator 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, GeoNorm). 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, GeoNorm). 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.

7.3 Method 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, HIS-2019, RS-2020, 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., $BETWEEN(p_1, p_2)$ 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-Geocator, 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-Geocator), 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, DeezyMatch, Bi-LSTM, LGGeoCoder, TR-2022, GeoNorm). 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.

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

GeoCoder	Accuracy@161 km (\uparrow)				Mean error distance (\downarrow)			
	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 & Baldrige, 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 & 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	–
TR-2022 (Cardoso et al., 2022)	91	–	–	87	2.2	–	–	1.1

CME-2019 and TR-2022 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) or the embeddings generated by the pre-trained BERT Devlin 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.

GeoNorm is a geocoding architecture that improves toponym resolution by employing a two-stage generate-and-rerank method. Initially, it uses lexical-based information retrieval to suggest potential location entries from a geospatial ontology, GeoNames. These candidates are then prioritized using a transformer-based model that incorporates data such as population size. The first stage resolves clear entities like countries and states, while the second stage addresses more ambiguous locations, using results from the first as contextual support. This approach allows GeoNorm to achieve top-notch accuracy in identifying geographical references in text.

7.4 Challenges: geocoding systems

One of the challenges in geocoding research is the lack of consistency in evaluation datasets used by different geocoders. While the LGL, WikTOR, GeoVirus, and WOTR datasets have been shared by multiple geocoders, there is still much variability in the choice of evaluation datasets. This can make it difficult to compare the performance of different geocoders and to draw meaningful conclusions from the results. We nevertheless present the partial comparison that is possible in Table 3.

The table reveals a challenge for the neural network models: they are data hungry. The gains of neural network models over prior approaches are modest on smaller datasets, such as LGL and GeoVirus, and only become large on the larger datasets, such as WikTOR and WOTR. This need for large datasets may be due to the architectures themselves, or they may be a result of the simpler set of features input to neural network systems as compared to pre-neural-network systems.

7.5 Opportunities: geocoding systems

One opportunity for geocoding system research is to increase the size of the training datasets. This could be achieved by applying techniques like multi-task learning to train a single model using the variety of available geocoding datasets.

Another opportunity is to incorporate additional features into the deep learning models. For instance, document-level consistency features like *one sense per referent*, geospatial consistency features like *spatial minimality*, and additional database information beyond population were used by geocoding systems before deep learning models. Designing neural architectures that can incorporate such features could yield performance gains not possible with the current feature sets.

8 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 & 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.

9 Conclusion

After surveying a decade of work on geocoding, we have identified 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.

Author Contributions **ZZ**: Conceptualization, Methodology, Data Curation, Software, Formal analysis, Visualization, Writing - Original Draft, Writing - Review and Editing. **SB**: Funding Acquisition, Supervision, Conceptualization, Methodology, Writing - Original Draft, Writing - Review and Editing.

Funding Defense Advanced Research Projects Agency, W911NF-18-1-0014 and National Science Foundation, 1831551.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

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

- Aldana-Bobadilla, E., Molina-Villegas, A., Lopez-Arevalo, I., Reyes-Palacios, S., Muñoz-Sánchez, V., & Arreola-Trapala, J. (2020). Adaptive geoparsing method for toponym recognition and resolution in unstructured text. *Remote Sensing*, *12*(18), 3041. <https://doi.org/10.3390/rs12183041>
- Ardanuy, M. C., Beavan, D., Beelen, K., Hosseini, K., Lawrence, J., McDonough, K., Nanni, F., van Strien, D. & Wilson, D. C. (2022). A dataset for toponym resolution in nineteenth-century english newspapers. *Journal of Open Humanities Data* 8.
- Ardanuy, M. C., Hosseini, K., McDonough, K., Krause, A., van Strien, D. & Nanni, F. (2020). A deep learning approach to geographical candidate selection through toponym matching. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '20*, New York, NY, USA, pp. 385-388. Association for Computing Machinery.
- Ardanuy, M. C., McDonough, K., Krause, A., Wilson, D. C. S., Hosseini, K. & van Strien, D. (2019). Resolving places, past and present: Toponym resolution in historical british newspapers using multiple resources. In *Proceedings of the 13th Workshop on Geographic Information Retrieval, GIR '19*, New York, NY, USA. Association for Computing Machinery.
- Ardanuy, M. C. & Sporleder, C. (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*, pp. 175–180.
- Ashktorab, Z., Brown, C., Nandi, M. & Culotta, A. (2014). Tweed: Mining twitter to inform disaster response. In *ISCRAM*, pp. 269–272.
- Berico Technologies. (2012). Cartographic location and vicinity indexer (clavin).
- Bhargava, P., Spasojevic, N. & Hu, G. (2017, September). 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*, Copenhagen, Denmark, pp. 131–139. Association for Computational Linguistics.
- Cardoso, A. B., Martins, B. & Estima, J. (2019). Using recurrent neural networks for toponym resolution in text. In *EPIA Conference on Artificial Intelligence*, pp. 769–780. Springer.
- Cardoso, A. B., Martins, B. & Estima, J. (2022). A novel deep learning approach using contextual embeddings for toponym resolution. *ISPRS International Journal of Geo-Information* *11*(1). 10.3390/ijgi11010028 .
- Chen, P., Xu, H., Zhang, C., & Huang, R. (2022). Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In M. Carpuat, M.-C. de Marneffe & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3329–3339). Seattle, United States: Association for Computational Linguistics.
- Cheng, Z., Caverlee, J. & Lee, K. (2010). You are where you tweet: a content-based approach to geolocating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768.
- de Bruijn, J. A., de Moel, H., Jongman, B., Wagemaker, J., & Aerts, J. C. (2018). Taggs: Grouping tweets to improve global geoparsing for disaster response. *Journal of Geovisualization and Spatial Analysis*, *2*(1), 2.
- DeLozier, G., Baldrige, J. & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15*, pp. 2382–2388. AAAI Press.
- DeLozier, G., Wing, B., Baldrige, J. & Nesbit, S. (2016, August). 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)*, Berlin, Germany, pp. 188–198. Association for Computational Linguistics.
- Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) .
- Fize, J., Moncla, L., & Martins, B. (2021). Deep learning for toponym resolution: Geocoding based on pairs of toponyms. *ISPRS International Journal of Geo-Information*, *10*(12), 818.
- Freire, N., Borbinha, J., Calado, P. & Martins, B. (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*, pp. 339–348.
- Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P. & Petras, V. (2005). Geoclef: the clef 2005 cross-language geographic information retrieval track overview. In *Workshop of the cross-language evaluation forum for european languages*, pp. 908–919. Springer.

- Gorski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., & Bartelmann, M. (2005). Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2), 759.
- Gritta, M., Pilehvar, M. T. & Collier, N. (2018, July). Which Melbourne? augmenting geocoding with maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp. 1285–1296. Association for Computational Linguistics.
- Gritta, M., Pilehvar, M. T., & Collier, N. (2020). A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, 54(3), 683–712. <https://doi.org/10.1007/s10579-019-09475-3>
- Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2017). What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2), 603–623. <https://doi.org/10.1007/s10579-017-9385-8>
- Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & 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.
- Hay, S. I., Battle, K. E., Pigott, D. M., Smith, D. L., Moyes, C. L., Bhatt, S., Brownstein, J. S., Collier, N., Myers, M. F., George, D. B., et al. (2013). Global mapping of infectious disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614), 20120250.
- Hoang, T. B. N., & Mothe, J. (2018). Location extraction from tweets. *Information Processing & Management*, 54(2), 129–144.
- Hu, X., Sun, Y., Kersten, J., Zhou, Z., Klan, F., & Fan, H. (2023). How can voting mechanisms improve the robustness and generalizability of toponym disambiguation? *International Journal of Applied Earth Observation and Geoinformation*, 117, 103191.
- Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T. & Ruths, D. (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*.
- Kamaloo, E. & Rafiei, D. (2018). A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1287–1296.
- Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J.O., Hardisty, F., Pezanowski, S., Mitra, P. & MacEachren, A. M. (2013). Geotxt: a web api to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval*, pp. 72–73.
- Katz, P. & Schill, A. (2013). To learn or to rule: two approaches for extracting geographical information from unstructured text. *Data Mining and Analytics 2013 (AusDM'13)* 117 .
- Kulkarni, S., Jain, S., Hosseini, M. J., Baldridge, J., Ie, E. & Zhang, L. (2020). Spatial language representation with multi-level geocoding. *CoRR arXiv:2008.09236*.
- Kumar, A., & Singh, J. P. (2019). Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction*, 33, 365–375.
- Laparra, E. & Bethard, S. (2020, December). A dataset and evaluation framework for complex geographical description parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), pp. 936–948. International Committee on Computational Linguistics.
- Lee, S., Farag, M., Kanan, T. & Fox, E. A. (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*, pp. 273–274.
- Leidner, J. (2007). *Toponym resolution: A comparison and taxonomy of heuristics and methods*. Ph.D. thesis, PhD Thesis, University of Edinburgh.
- Leidner, J. L. (2021). *A survey of textual data & geospatial technology, Handbook of Big Geospatial Data* (pp. 429–457). Springer.
- Lieberman, M. D. & Samet, H. (2011). Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 843–852.
- Lieberman, M. D. & 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*, pp. 731–740.

- Lieberman, M. D., Samet, H. & Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)*, pp. 201–212. IEEE.
- Luo, X., Qiao, Y., Li, C., Ma, J., & Liu, Y. (2020). An overview of microblog user geolocation methods. *Information Processing & Management*, 57(6), 102375.
- Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S., & Clancy, S. (2010). Spatialml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3), 263–280.
- Martins, B., Anastácio, I., & Calado, P. (2010). *A machine learning approach for resolving place references in text, Geospatial thinking* (pp. 221–236). Springer.
- Melo, F., & Martins, B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21(1), 3–38.
- Monteiro, B. R., Davis, C. A., Jr., & Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96, 23–34.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018, June). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pp. 2227–2237. Association for Computational Linguistics.
- Rayson, P., Reinhold, A., Butler, J., Donaldson, C., Gregory, I. & Taylor, J. (2017). A deeply annotated testbed for geographical text analysis: The corpus of lake district writing. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, pp. 9–15.
- Santos, J., Anastácio, I., & Martins, B. (2015). Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80(3), 375–392.
- Santos, R., Murrieta-Flores, P., Calado, P., & Martins, B. (2018). Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, 32(2), 324–348.
- Smith, D. A. & Crane, G. (2001). Disambiguating geographic names in a historical digital library. In *International Conference on Theory and Practice of Digital Libraries*, pp. 127–136. Springer.
- Speriosu, M. & Baldrige, J. (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)*, pp. 1466–1476.
- Tateosian, L., Guenter, R., Yang, Y. P. & Ristaino, J. (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, pp. 17.
- Tobin, R., Grover, C., Byrne, K., Reid, J. & Walsh, J. (2010). Evaluation of georeferencing. In *proceedings of the 6th workshop on geographic information retrieval*, pp. 1–8.
- Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). Geocorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32(1), 1–29.
- Wang, X., Ma, C., Zheng, H., Liu, C., Xie, P., Li, L. & Si, L. (2019). Dm_nlp at semeval-2018 task 12: A pipeline system for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 917–923.
- Weissenbacher, D., Magge, A., O'Connor, K., Scotch, M. & Gonzalez-Hernandez, G. (2019, June). SemEval-2019 task 12: Toponym resolution in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA, pp. 907–916. Association for Computational Linguistics.
- Yan, Z., Yang, C., Hu, L., Zhao, J., Jiang, L., & Gong, J. (2021). The integration of linguistic and geospatial features using global context embedding for automated text geocoding. *ISPRS International Journal of Geo-Information*, 10(9), 572.
- Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9), 37–70.
- Zhang, Z., & Bethard, S. (2023). Improving toponym resolution with better candidate generation, transformer-based reranking, and two-stage resolution. In A. Palmer & J. Camacho-collados (Eds.), *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)* (pp. 48–60). Toronto, Canada: Association for Computational Linguistics.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.