How can voting mechanisms improve the robustness and generalizability of toponym disambiguation?

Xuke Hu[a], Yeran Sun[b], Jens Kersten[a], Zhiyong Zhou[c], Friederike Klan[a], Hongchao Fan[d]

[a]Institute of Data Science, German Aerospace Center, Germany
[b]Department of Geography, University of Lincoln, UK
[c]Department of Geography, University of Zurich, Switzerland
[d]Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Norway

ARTICLE HISTORY
Compiled October 12, 2022

Abstract
A vast amount of geospatial information exists in natural language texts, such as tweets and news. Extracting geospatial information from texts is called Geoparsing, which includes two subtasks: toponym recognition and toponym disambiguation, i.e., to identify the geospatial representations of toponyms. This paper focuses on toponym disambiguation, which is approached by toponym resolution and entity linking. Recently, many novel approaches have been proposed, especially deep learning-based, such as CamCoder, GENRE, and BLINK. In this paper, a spatial clustering-based voting approach combining several individual approaches is proposed to improve SOTA performance regarding robustness and generalizability. Experiments are conducted to compare a voting ensemble with 20 latest and commonly-used approaches based on 12 public datasets, including several highly challenging datasets (e.g., WikToR). They are in six types: tweets, historical documents, news, web pages, scientific articles, and Wikipedia articles, containing 98,300 places across the world. Experimental results show that the voting ensemble performs the best on all the datasets, achieving an average *Accuracy@161km* of 0.86, proving its generalizability and robustness. Besides, it drastically improves the performance of resolving fine-grained places, i.e., POIs, natural features, and traffic ways.

KEYWORDS
Toponym disambiguation; Toponym resolution; Geocoding; Geoparsing; Entity linking; Entity disambiguation; Voting.

1. Introduction

Huge and ever-increasing amounts of semi- and unstructured text data, like news articles, historical archives, and social media posts are available online and offline. These documents often refer to geographic regions or specific places on earth, and therefore contain valuable but hidden geospatial information in the form of toponyms or location references. The information is useful not only for scientific studies, such as spatial humanities (Gregory et al. 2015), but can also contribute to many practical applications, such as geographical information retrieval (Purves et al. 2018), disaster management (Shook and Turner 2016), disease surveillance (Scott et al. 2019), and traffic management (Milusheva et al. 2021). Extracting geospatial information from texts is called geoparsing, which includes two subtasks: toponym recognition, i.e., to recognize toponyms from texts, and toponym disambiguation, i.e., to handle the situation in which one toponym can refer to more than one geographical location, as shown in Figure 1. Toponym recognition has been extensively studied (Hu et al. 2022b; Wang, Hu, and Joseph 2020; Hu et al. 2021) and the recognition performance is already very high due to the advancement of deep learning techniques, seeing Hu et al. (2022a) for an overview. Therefore, this paper will focus on toponym disambiguation, which is still challenging.

Toponym disambiguation can be approached by entity linking and toponym resolution. Entity linking aims to link an entity (e.g., person, location) mentioned in texts to an entry of Knowledge Bases (KBs), such as Wikipedia (Wikipedia 2004) and DBpedia (Auer et al. 2007). Recently, many deep learning-based entity linkers (ELs) emerged,

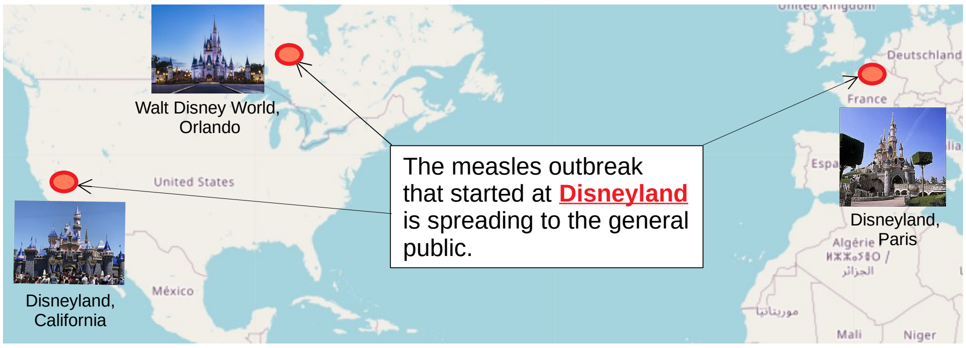

Figure 1. Example of toponym disambiguation. *'Disneyland'* can refer to multiple different locations, such as the park in Paris (France), California (US), Orlando (Canada), and other places named with *'Disneyland'*.

such as GENRE (De Cao et al. 2021) and BLINK (Wu et al. 2020), pushing the state-of-the-art performance (Sevgili et al. 2022). Toponym resolution aims to determine the coordinates of toponyms, focusing only on location entities. Quite a few toponym resolution approaches have also been proposed, such as Edinburgh Geoparser (Grover et al. 2010), CamCoder (Gritta, Pilehvar, and Collier 2018), and CHF (Kamalloo and Rafiei 2018). Despite the impressive advancement of ELs, many toponyms mentioned in texts cannot be linked to KBs since the current KBs contain only a small proportion of location entries, lacking many small, unpopular, or fine-grained places (e.g., roads and shops). For instance, the largest KB, Wikipedia contains about one million places, while over 23 million and 12 million places have been recorded in OpenStreetMap [1] and GeoNames [2], respectively (Hu et al. 2022a). Toponym resolution approaches normally link toponyms to gazetteers, such as GeoNames. However, according to the comparison study (Wang and Hu 2019), the performance of toponym resolution approaches varies by datasets and no one can always perform the best.

In this paper, we propose a spatial clustering-based voting approach that combines seven individual approaches to overcome the shortcomings of existing toponym disambiguation approaches and further push state-of-the-art performance. The principle

[1]https://www.openstreetmap.org/
[2]http://www.geonames.org/

of the proposed voting mechanism is that the minority is subordinate to the majority. We then compare the voting ensemble with 20 latest and commonly-used toponym disambiguation approaches based on 12 public datasets.

Our contributions are twofold. First, we propose a more general and robust voting-based approach at the cost of moderate increased computational costs. Second, it is the first time that many competing approaches (especially deep learning-based ELs) are compared based on numerous datasets.

2. Related works

There are two main ways to disambiguate toponyms: entity linking and toponym resolution. Recently, surveys (Sevgili et al. 2022; Möller, Lehmann, and Usbeck 2022) on entity linking have been conducted. Therefore, we will review only toponym resolution approaches by dividing them into three groups: (1) Rules, (2) Learning and ranking, and (3) Learning and classification.

2.1. *Rules*

Given a toponym, rule-based approaches first search gazetteers to find all the candidates that match or partially match the toponym, and then rank or score the candidates by manually defined IF-THEN rules, using heuristics like string similarity, the candidate's population and admin levels, spatial proximity, and one-sense-per-referent. Representative rule-based approaches include Edinburgh Geoparser (Grover et al. 2010), CLAVIN [3], GeoTxt (Karimzadeh et al. 2019), TAGGS (de Bruijn et al. 2018), and CHF, CBH, and SHS (Kamalloo and Rafiei 2018). Qi et al. (2019) rules that if a toponym appears in training examples, the candidate with the highest frequency in the training examples is selected. Otherwise, the candidate with the highest population is selected. Karimzadeh et al. (2019) accumulates the score of a candidate based on nine optional heuristics, such as population, the number of alternate names, GeoNames feature codes, hierarchical relationship, and proximity relationship between two toponyms in the same tweet.

Rule-based approaches are easy to implement and computational-efficiency. However, manually defined rules are often fragile and ineffective, considering the variability of describing or mentioning toponyms in unstructured texts.

2.2. *Learning and ranking*

The workflow of learning and ranking-based approaches (Lieberman and Samet 2012; Santos, Anastácio, and Martins 2015; Ardanuy and Sporleder 2017; Wang et al. 2019) is similar to the rule-based approaches. The only difference lies in the rules, which are not explicitly defined but learned from annotated examples. For example, Lieberman and Samet (2012) train a random forest model using context-free features (e.g., population and the distance of a candidate to a news's local location) and adaptive context features, such as sibling and proximate relationships between the candidates of the toponyms in a certain context window. The input of the classifiers are the features related to the pair of (*toponym*, *candidate*). The output is 1 or 0, indicating if the toponym refers to the candidate or not. Classification confidence is regarded as the

[3] `https://github.com/Novetta/CLAVIN`

score of the candidate. Wang et al. (2019) train a LightGBM (Ke et al. 2017) model using features like name string similarity, candidate attributes, neighboring toponyms, and context features. Context features refer to the contextual similarity between the toponym and the candidate, while the context of the candidate is obtained from its Wikipage. Apart from fully supervised approaches, weakly-supervised and unsupervised approaches (Speriosu and Baldridge 2013; Ardanuy and Sporleder 2017; Fize, Moncla, and Martins 2021) have also been proposed to reduce the amount of annotated data required. For example, Ardanuy and Sporleder (2017) define a model to score and rank candidates, using features like the context similarity between a toponym and a candidate, geographic closeness to the base location of a collection, and geographic closeness of toponyms. The parameters of the model are learned from a small training set.

Learning and ranking-based approaches can automatically disambiguate toponyms without requiring as much expert knowledge as rule-based approaches do. However, the trained models are often not general enough due to the paucity of sufficient and accurate training data although unsupervised or weakly supervised techniques have been adopted.

2.3. *Learning and classification*

Learning and classification-based approaches Gritta, Pilehvar, and Collier (2018); Kulkarni et al. (2020); Yan et al. (2021); Cardoso, Martins, and Estima (2021); DeLozier, Baldridge, and London (2015) divide the earth's surface into multiple cells and then locate a toponym to a certain cell (class). For example, Gritta, Pilehvar, and Collier (2018) proposed a CNN-based model, named CamCoder, using features like the target toponym, the other toponyms in the text, and the context removing the toponyms, and the prior probability of the candidate of the target toponym based on the its population.

1.4M training examples are generated from over 1M geographically annotated Wikipages. Different from CamCoder that uses only local context features (e.g., co-occurrence of words and toponyms in texts), Yan et al. (2021) uses also global context features, including topic and location embedding. Some studies also leverage language models, i.e., the spatial distribution of the words in texts(Speriosu and Baldridge 2013; Wing and Baldridge 2011; DeLozier, Baldridge, and London 2015), based on the assumption that apart from toponyms, common language words, such as *'howdy'* and *'phillies'* can often be geographically indicative. For example, DeLozier, Baldridge, and London (2015) proposed TopoCluster, a gazetteer-independent approach using geographic word profiles.The per-word spatial distribution is first learned based on 700,000 geographically annotated Wikipages. Disambiguation is then performed by merging the shared geographic preferences (cells) of a toponym and all words in the context of the toponym.

Learning and classification-based approaches are normally trained on geographically annotated Wikipages, which contain around 1 million places. However, there are still many places, which are not presented on Wikipedia.

3. Proposed approach

In this section, we introduce the voting approach, summarize 20 individual approaches used to form or to be compare with a voting ensemble, and illustrate the voting approach with four examples.

3.1. *Voting approach*

The idea of this study is inspired by Won, Murrieta-Flores, and Martins (2018); Hoang and Mothe (2018), which combine multiple existing toponym recognition approaches as a voting ensemble, achieving promising recognition performance. Each approach has its own limitations while combining multiple approaches can overcome these shortcomings. Different approaches normally return (or vote for) different locations (candidates) for a toponym in texts. We count the votes for them and choose the one with the most votes. Since some approaches outperform the other approaches, the superior approaches' votes should have a higher weight. That is, we set a higher weight to superiors approaches by copying the coordinate estimation of the approaches multiple times. To realize the voting approach, we adopt DBSCAN (Khan et al. 2014), which groups together points that are close to each other based on a distance measurement (denoted by eps) and a minimum number of points required to form a group (denoted by $minPts$). The workflow of the voting approach is as follows:

(1) Group the coordinate estimation of the individual approaches of a voting ensemble with DBSCAN.
(2) If clusters are formed, select the largest cluster or randomly select one when multiple clusters of same size exist. Treat the centroid of the coordinate estimations in the selected cluster as the voting result.
(3) If no clusters are formed, traverse the individual approaches of the ensemble and treat the first valid estimation as the voting result.

Invalid estimation refers to the situation where an approach fails to estimate the coordinates of a toponym, such as the one not appearing in gazetteers. The maximum possible error distance (half of the earth's circumference) is assigned to an invalid estimation, which equals 20,039 km (Gritta, Pilehvar, and Collier 2020).

3.2. *Individual approaches*

Table 1 lists representative approaches, covering all types as discussed in Section 2. By default, we would modify their implementation to input gold toponyms to their entity (toponym) disambiguation step. We obtain the coordinates of DBpedia and Wikipedia entities if they are geographically annotated. Otherwise, the coordinates of (0,0) are returned, denoting an invalid estimation. Details of the 20 approaches are as follows:

- **DBpedia Spotlight** is a popular EL. We use the provided HTTPS API [5] to annotate and link entities in texts.
- **Entity-Fishing** is an EL based on Random Forest and Gradient Tree Boosting. We use its spaCy wrapper [6].
- **MulRel-NEL** is a neural entity-linker. We use the provided API [7] of Radboud Entity Linker (REL) (van Hulst et al. 2020), which uses **mulrel-nel** for entity disambiguation.

[5] `https://www.dbpedia-spotlight.org/api`
[6] `https://github.com/Lucaterre/spacyfishing`
[7] `https://github.com/informagi/REL`

Table 1. 20 representative approaches for toponym disambiguation. ML and DL denote traditional machine learning algorithms based on feature engineering and deep learning algorithms, respectively.

Name	Method Type
DBpedia Spotlight (Mendes et al. 2011)	EL
Entity-Fishing (ent 2016–2022)	EL
MulRel-NEL (Le and Titov 2018)	EL
DCA (Yang et al. 2019)	EL
BLINK (Wu et al. 2020)	EL
Bootleg (Orr et al. 2020)	EL
GENRE (De Cao et al. 2021)	EL
ExtEnD (Barba, Procopio, and Navigli 2022)	EL
LUKE (Yamada et al. 2022)	EL
Nominatim [4]	Geocoder
Adaptive learning (Lieberman and Samet 2012)	ML (Ranking)
Edinburgh Geoparser (Grover et al. 2010)	Rule
Population-Heuristics (Speriosu and Baldridge 2013)	Rule
CLAVIN	Rule
TopoCluster (DeLozier, Baldridge, and London 2015)	ML (Classification)
Mordecai (Halterman 2017)	Rule
CBH, SHS, CHF (Kamalloo and Rafiei 2018)	Rule
CamCoder (Gritta, Pilehvar, and Collier 2018)	DL (Classification)

- **DCA** is a neural entity-linker. We retrain the model [8] based on the public and widely used AIDA CoNLL-YAGo dataset (Hoffart et al. 2011).
- **Bootleg** adopts a transformer architecture. We use the provided model [9] which is trained on weakly-labeled training data.
- **BLINK** is an EL based on fine-tuned BERT (Devlin et al. 2018). We use the provided model [10] that was pre-trained on nearly 9M unique triples document-mention-entity from Wikipedia.
- **GENRE** uses a transformer-based architecture. We use the provided model [11] that was first pre-trained on nearly 9M unique triples document-mention-entity from Wikipedia and then fine-tuned with the AIDA dataset.
- **LUKE** is a entity disambiguation model based on BERT, using both local (word-based) and global (entity-based) contextual information. We use the provided model [12] that was trained on a large entity-annotated corpus generated from Wikipedia.
- **ExtEnD** adopts Transformer-based architectures, which was first pretrained on the same Wikipedia dataset as BLINK and then fine-tuned on the AIDA dataset. We use the fine-tuned model [13] directly.
- **Nominatim** is a geocoder, built on OpenStreetMap , which is used as a baseline system. We input a toponym to it and keep the first result it returns.
- **Population-Heuristics** uses the heuristic of the largest population, which is used as a baseline system. We implement the approach based on GeoNames.
- **Edinburgh Geoparser** [14] is a geoparsing tool developed by the Language Tech-

[8] `https://github.com/YoungXiyuan/DCA`

[9] `https://github.com/HazyResearch/bootleg`

[10] `https://github.com/facebookresearch/BLINK`

[11] `https://github.com/facebookresearch/GENRE`

[12] `https://github.com/studio-ousia/luke/tree/master/examples/entity_disambiguation`

[13] `https://github.com/SapienzaNLP/extend`

[14] `https://www.ltg.ed.ac.uk/software/geoparser/`

nology Group (LTG) at the Edinburgh University. We use its online API to annotate and disambiguate toponyms.

- **CLAVIN** applies several heuristics and fuzzy search for toponym resolution. We use its implementation directly.
- **Adaptive Learning** is a random forest-based toponym resolution approach. We use its implementation [15] to retrain a model based on one dataset, i.e., LGL (Lieberman, Samet, and Sankaranarayanan 2010).
- **Mordecai** is a geoparsing tool, using word2vec for inferring the correct country for a set of toponyms in texts. We use its implementation [16] directly.
- **CBH,SHS,CHF** are three rule-based approaches proposed by Kamalloo and Rafiei (2018). We use their implementation [17] directly.
- **TopoCluster** is a language model-based geoparsing tool. We use its implementation [18] directly.
- **CamCoder** is a CNN-based geoparsing tool.We use the trained model [19] directly.

3.3. *Examples*

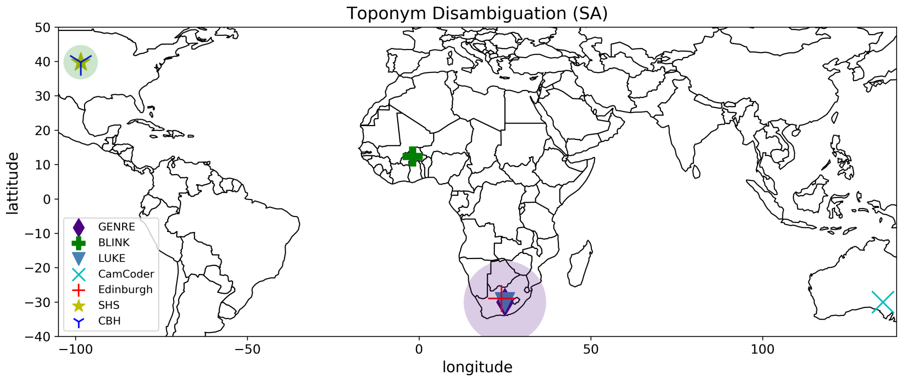

Figure 2. An example to show how the voting approach works. The target toponym is *'SA'*, whose true location is in the largest cluster (purple circle). The context of the toponym is: *'Kgosi (chief) Nyalala Pilane of the Bakgatla-ba-Kgafela community — perhaps even more than any other chief in* ***SA*** *— has been the subject of a litany of maladministration and corruption allegations'*.

We use four examples to illustrate the principle of the voting approach. We assume that the voting ensemble combines seven individual approaches with each having one vote: **GENRE**, **BLINK**, **LUKE**, **CamCoder**, **Edinburgh Geoparser**, **SHS**, and **CBH**. Figures 2, 3, 4, and 5 show the estimated location of *'SA'*, *'False River'*, *'Victoria Park'*, and *'Mount Sheridan'* by the seven individual approaches and formed clusters, respectively. In GeoNames we can find 58 records of *'SA'*, 23 records of *'False River'*, 589 records of *'Victoria Park'*, and 25 records of *'Mount Sheridan'*, which actually refer to the country of South Africa, a county in Louisiana, US, a park in London,

[15] https://github.com/ehsk/CHF-TopoResolver
[16] https://github.com/openeventdata/mordecai
[17] https://github.com/ehsk/CHF-TopoResolver
[18] https://github.com/grantdelozier/TopoCluster
[19] https://github.com/milangritta/Geocoding-with-Map-Vector

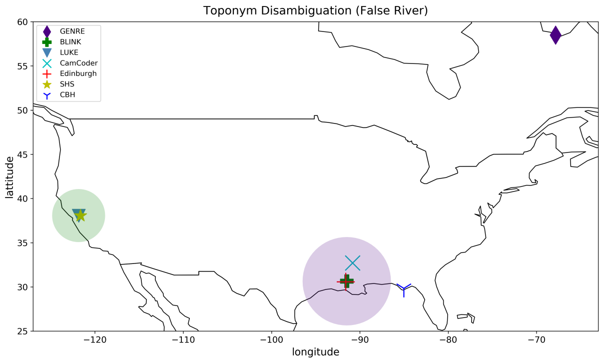

Figure 3. An example to show how the voting approach works. The target toponym is *'False River'*, whose true location is in the largest cluster (purple circle). The context of the toponym is: '*The enemy have now left Waterloo, and that is of no importance, but the Rosedale country is of to visit, with the cavalry, and so also is the **False River** country. The cavalry must go to Rosedale and return by False River*'.

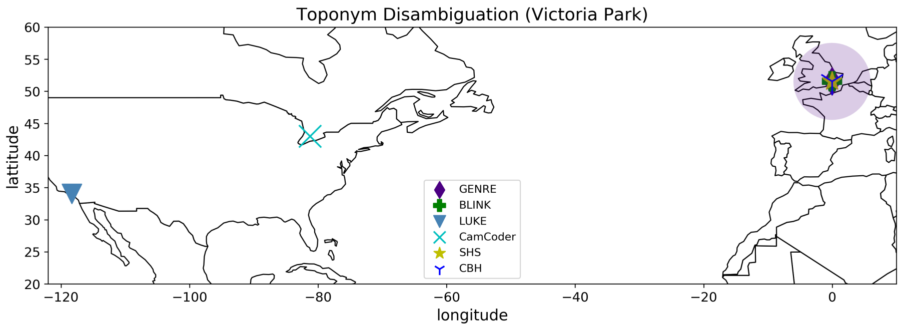

Figure 4. An example to show how the voting approach works. The target toponym is *'Victoria Park'*, whose true location is in the largest cluster (purple circle). The context of the toponym is: '*The Clash - White Riot (Live 1978 **Victoria Park**, London): via @YouTube Let's start our shift!*'.

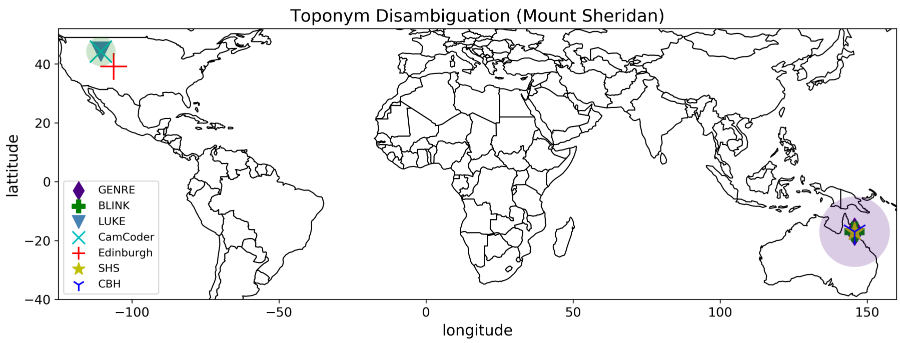

Figure 5. An example to show how the voting approach works. The target toponym is *'Mount Sheridan'*, whose true location is in the largest cluster (purple circle). The context of the toponym is: '*Nine cases of the mosquito-borne illness have been confirmed in the Cairns suburbs of Edmonton, **Mount Sheridan**, Bentley Park, and Trinity Beach.*'.

UK, and a suburb of Cairns in the Cairns Region, Queensland, Australia, respectively. Their true location are all in the largest cluster, denoted by the purple circle. Note that, Edinburgh Geoparser cannot recognize *'Victoria Park'*, which thus cannot vote. From the four examples, we can observe that no one individual approach can correctly resolve the toponym in all the examples except the voting ensemble.

4. Experiments

In this section, we first introduce the used test datasets and evaluation metrics. We then propose a voting ensemble and compare the voting ensemble with the 20 approaches regarding correctness and computational efficiency. Finally, we conduct a sensitivity analysis of the voting approach.

4.1. *Test datasets*

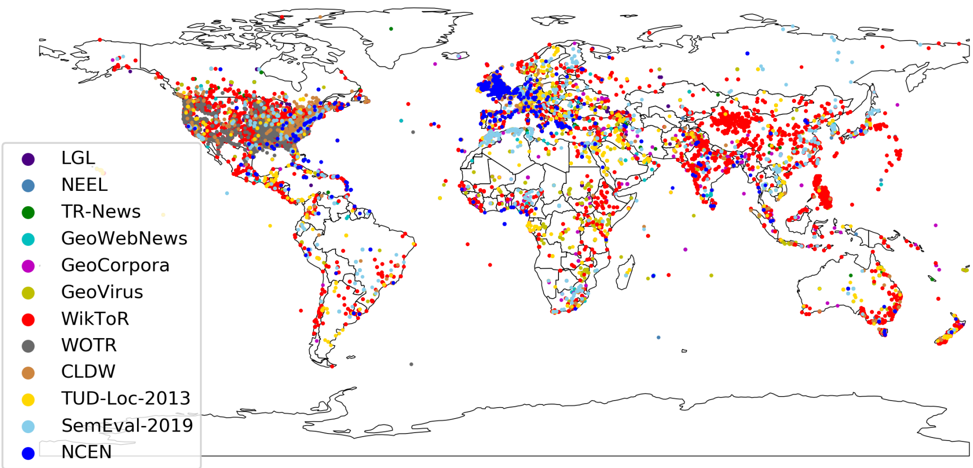

Figure 6. Spatial distribution of the 98,300 toponyms in the 12 datasets.

To thoroughly evaluate the voting ensemble and compare it with the 20 approaches, we used 12 public datasets, as shown in Table 2 and Figure 6. Note that, ELs normally use Wikipedia as the target KB, while the toponyms in most of the datasets are linked to GeoNames except WikToR, NEEL, GeoVirus, and NCEN. However, the coordinates of some coarse-grained places (e.g., country) in Wikipedia and GeoNames are inconsistent. For instance, *'United States'* is geocoded to (40,-100) and (39.76, -98.5) and *'China'* is geocoded to (35, 103) and (35, 105) in Wikipedia and GeoNames, respectively. Such places appear frequently in the datasets, which can cause incorrect evaluation. From the datasets, we found 3,147 records of 29 frequent and misaligned places, including [*'China', 'Chinese', 'Russia', 'Russian', 'Russians', 'Australia', 'Canada','Canadians','Canadian', 'United States', 'American','USA', 'America', 'U.S.', 'U.S', 'United States of America',' Americans', 'North America', 'South America', 'India', 'Algeria', 'Europe','European', 'Western Europe', 'Asia', 'Africa',* *'West Africa', 'North Africa', 'Middle East'*], and will be ignored during the evaluation.

Table 2. Summary of test datasets.

Dataset	**Text Count**	**Toponym Count**	**Type**	**KB/Gazetteer**
LGL	588	5,088	News	GeoNames
NEEL	2,135	481	Tweet	DBpedia
TR-News	118	1,319	News	GeoNames
GeoWebNews	200	5,121	News	GeoNames
GeoCorpora	6,648	3,100	Tweet	GeoNames
GeoVirus	230	2,170	News	Wikipedia
WikToR	5,000	31,500	Wikipedia article	Wikipedia
WOTR	1,643	11,795	History	GeoNames
CLDW	62	3,814	History	GeoNames
TUD-Loc-2013	152	3,850	Web page	GeoNames
SemEval-2019	90	8,360	Scientific article	GeoNames
NCEN	455	3,364	History	Wikipedia

Details of the 12 test datasets are as follows:

- **LGL** [20] (Local-Global Lexicon) corpus was created by Lieberman, Samet, and Sankaranarayanan (2010), containing 588 human-annotated news articles published by 78 local newspapers.
- **NEEL** [21] is the gold dataset of 2016 Named Entity rEcognition and Linking challenge, including 2,135 tweets covering multiple noteworthy events from 2011 to 2013, such as the Westgate Shopping Mall shootout.
- **TR-News** [22] was created by Kamalloo and Rafiei (2018) through annotating and linking Toponyms to entries in GeoNames from news articles of various news sources.
- **GeoWebNews** [23] was shared by Gritta, Pilehvar, and Collier (2018), comprising human-annotated news articles from 200 globally distributed news sites collected from April 1st to 8th in 2018.
- **GeoCorpora** [24] was created by Wallgrün et al. (2018), containing 6,648 tweets related to multiple noteworthy events (e.g., ebola, flood, and rebel) that happened across the world in 2014 and 2015.
- **GeoVirus** [25] was created by Gritta, Pilehvar, and Collier (2018), containing 230 news articles related to disease outbreaks and epidemics, such as Ebola and Swine Flu.
- **WikToR** [26] was created by Gritta et al. (2018) in an automatic manner, containing 5,000 Wikipedia articles with many ambiguous toponyms, such as (*Santa Maria, California*), (*Santa Maria, Bulacan*), (*Santa Maria, Ilocos Sur*), and (*Santa Maria, Romblon*).

[20] `https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/lgl.xml`
[21] `http://microposts2016.seas.upenn.edu/challenge.html`
[22] `https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/TR-News.xml`
[23] `https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/tree/master/data`
[24] `https://github.com/geovista/GeoCorpora`
[25] `https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/GeoVirus.xml`
[26] `https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/WikToR.xml`

- **WOTR** [27] was created by DeLozier et al. (2016) based on a set of American Civil War archives, known as *Offical Records of the War of the Rebellion*.
- **CLDW** [28] (The Corpus of Lake District Writing) was created by Rayson et al. (2017) based on 80 texts which are writing samples about the English Lake District between the early seventeenth and the beginning of the twentieth century.
- **TUD-Loc-2013** [29] was first utilized in (Katz and Schill 2013), containing 152 texts from web pages.
- **SemEval-2019-12** [30] is the gold dataset of the Task 12 (Toponym Resolution in Scientific Papers) of the 13th International Workshop on Semantic Evaluation (SemEval) (Weissenbacher et al. 2019).
- **NCEN** [31] (The Nineteenth-Century English Newspapers) was created by Ardanuy et al. (2022), containing 343 newspaper articles published between 1780 and 1870 in four different locations of England (i.e., Manchester, Ashton-under-Lyne, Poole, and Dorchester).

4.2. *Evaluation metrics*

To fairly evaluate toponym disambiguation approaches, we assume that all the toponyms in the datasets can be correctly recognized at the toponym recognition step. However, DBpedia Spotlight and Edinburgh Geoparser provide only an online API and deploy the toponym recognition module on servers. Therefore, when evaluating the correctness of the two approaches, we will compare them with the other approaches on the correctly recognized toponyms (a subset of gold toponyms) by them, respectively.

From the standard metrics defined in (Gritta, Pilehvar, and Collier 2020), we adopt the three most important metrics. They are: (1) *Accuracy@161km*, which is the percentage of geocoding errors that are smaller than 100 miles (161 km); (2) *Mean Error (ME)*, which is the mean distance error of toponyms; (3) *Area Under the Curve (AUC)*, which is the total area under the curve of the normalized log error distance. AUC is calculated using Equation 1, where x_i denotes the distance error of the i-th toponym, N denotes the count of toponyms, and 20039 is the maximum possible error in km on earth.

$$AUC = \int_{i=1}^{N} \frac{ln(x_i + 1)}{ln(20039)} dx \qquad (1)$$

4.3. *Voting ensemble*

We propose a voting ensemble that combines seven individual approaches and manually assign a weight (count of votes) to each approach, denoted by the number in the brackets. We decide the combination manner and weights according to numerous experimental results. For the voting algorithm, we set the DBSCAN parameters $minPts$ and eps to 2 and 10 (km), respectively.

- **Voting**: GENRE (3), BLINK (2), LUKE (2), CamCoder (1), SHS (1), CBH (1),

[27] `https://github.com/barbarainacioc/toponym-resolution/tree/master/corpora/WOTR`
[28] `https://github.com/UCREL/LakeDistrictCorpus`
[29] `https://bitbucket.org/palladian_pk/tud-loc-2013/src/master/`
[30] `https://github.com/TharinduDR/SemEval-2019-Task-12-Toponym-Resolution-in-Scientific-Papers`
[31] `https://bl.iro.bl.uk/concern/datasets/f3686eb9-4227-45cb-9acb-0453d35e6a03`

Edinburgh Geoparser (1).

4.4. *Results*

We average each metric of the approaches on the 12 datasets. Figure 7 shows the result of the 18 approaches (excluding Edinburgh Geoparser and DBpedia Spotlight) and the voting ensemble on the gold toponyms of the test datasets. The voting ensemble achieves an *Accuracy@161km* at 0.86, improving the best individual approach, GENRE, by 5%. Similarly, the ensemble achieves the best *ME* and *AUC*, improving the best individual approach, GENRE, by 57% and 13%, respectively. Figures 8 and 9

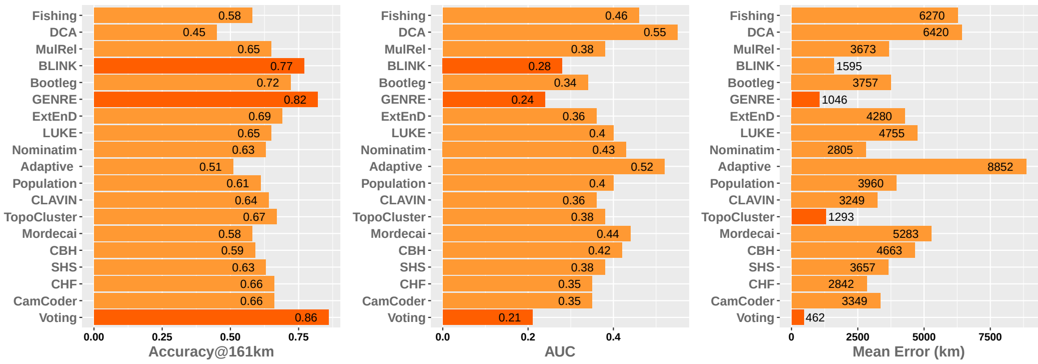

Figure 7. Average *Accuracy@161km* (↑), *AUC* (↓), and *ME* (↓) of approaches on gold toponyms.

shows the average *Accuracy@161km*, *Mean Error*, and *AUC* of the approaches on the subset of gold toponyms, which are correctly recognized by DBpedia Spotlight and by Edinburgh Geoparser, respectively, and the voting ensemble still performs the best. Among individual approaches, the state-of-the-art ELs, GENRE and BLINK, achieve promising results, outperforming the other ELs and toponym resolution approaches.

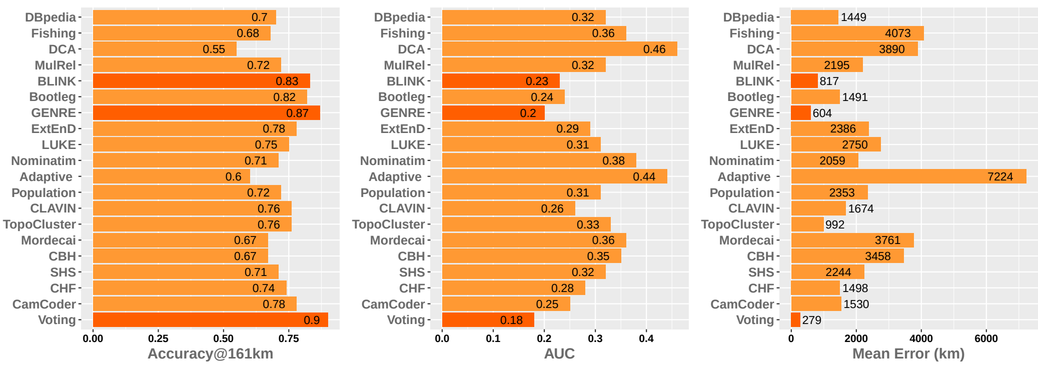

Figure 8. Average *Accuracy@161km* (↑), *AUC* (↓), and *ME* (↓) on the subset of gold toponyms, which are correctly recognized by DBpedia Spotlight.

We provide the raw result of each dataset in the supplement materials, from which we can see that on the gold toponyms, the voting ensemble achieves the best result on 35/36 (3 metrics evaluated for 12 datasets) indicators. Besides, GENRE and BLINK are especially effective on highly ambiguous datasets,including WikToR, WOTR, and

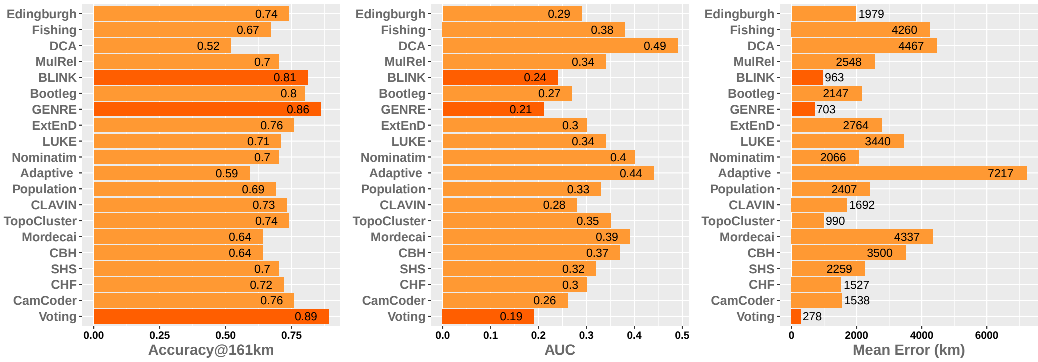

Figure 9. Average *Accuracy@161km* (↑), *AUC* (↓), and *ME* (↓) on the subset of gold toponyms, which are correctly recognized by Edinburgh Geoparser.

LDC, performing much better than the other individual approaches. These datasets contain many less-common or low-frequency places, such as *'Paris, Missouri'* and *'Lima, Oklahoma'*, on which the two baseline systems, Nominatim and Population-Heuristics adopting simple heuristics (i.e., popularity) thus perform poorly. Regardless of whether highly ambiguous or general datasets, the voting ensemble that combines several individual approaches can always achieve state-of-the-art performance, proving its generalizability and robustness.

4.5. *Place category*

We investigate the disambiguation performance of the approaches on four different types of places: admin units (e.g., country, state, and county), POIs (e.g., park, church, and hospital), traffic ways (e.g., street, highway, and bridge), natural features (e.g., river, beach, and hill). In the datasets of GeoCorpora, LGL, TR-News, GeoWebNews, CLDW, and Semeval-2019, the GeoNames ID of places have been provided, through which we determine in total 13,878 admin units (e.g., *'EU'*, *'Berlin'*, and *'Boone County'*), 820 POIs (e.g., *'Lambert-St. Louis International Airport'*, *'Sam Houston High School'*, and *'westboro baptist church'*), 1,605 natural features (e.g., *'Pine Island Bayou'*, *'Skiddaw Mountain'*, and *'Little Pine Creek'*), and 336 traffic ways (e.g., *'High Street'*, *'Lynchburg Railroad bridge'*, and *'Highway 49'*).

We then calculate *Accuracy@161km* of the approaches on each place type, ruling out Edinburgh Geoparser and DBpedia Spotlight since their toponym recognition modules can only correctly recognize a small proportion of toponyms. For example, they can only recognize 21/336 and 13/336 traffic ways, respectively. Figure 10 shows that most of the approaches perform well in resolving coarse-grained places (i.e., admin units), with fifteen can correctly resolving over 70% of the admin units. However, they are incapable of resolving fine-grained places, with only four, three, and one of them correctly resolving over 60% of the POIs, natural features, and traffic ways, respectively. The voting ensemble performs the best, on average improving the best individual approach, GENRE, by 11% in resolving fine-grained places. However, there is still space for improving the performance of resolving fine-grained places.

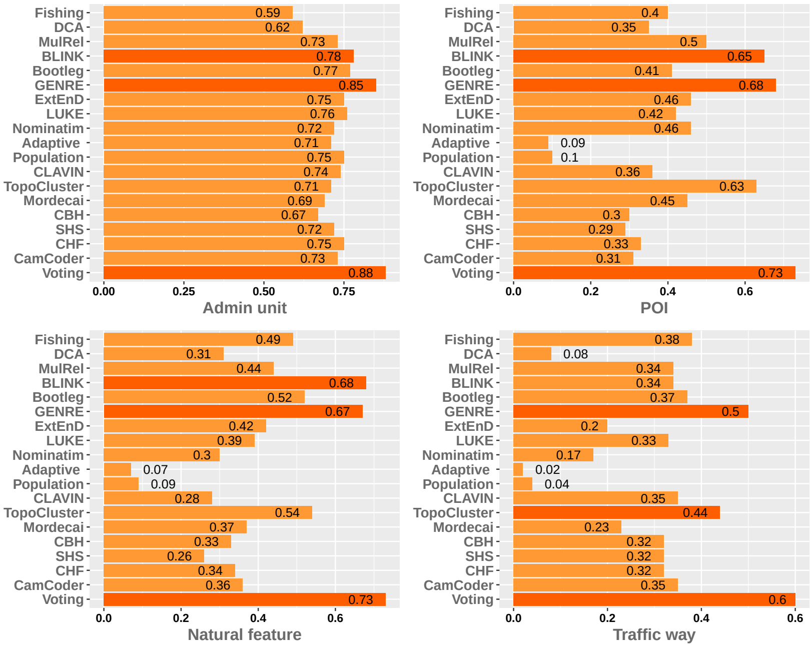

Figure 10. *Accuracy@161km* of the approaches on four categories with 13,878 admin units, 820 POIs, 1605 natural features, and 336 traffic ways.

4.6. *Computational efficiency*

We further investigate the computational efficiency of different approaches. We run each approach on the total datasets and record the consumed time without counting the training phase, as shown in Figure 11. Note that, we do not include Edinburgh Geoparser, Nominatim, DBpedia Spotlight, and Entity-Fishing in the comparison since they are online services and it is impossible to count the amount of time of processing done on the server. We run the toponym resolution approaches on a Dell laptop with an Intel Core i7-8650U CPU (1.90 GHz 8-Core) and a RAM of 16 GB, while we run the ELs on an NVIDIA Tesla V100 GPU of a cluster node since they require a GPU execution environment.

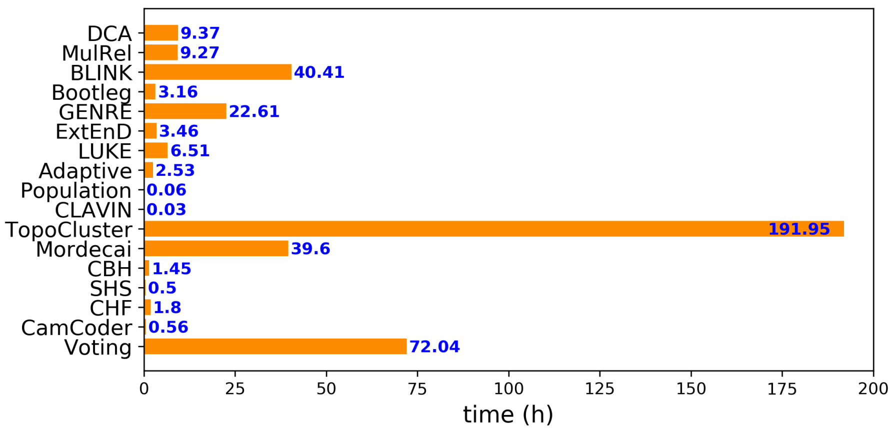

Figure 11. Time consumption of the approaches running on the total test datasets.

Generally, ELs take more time (from 3 hours to 40 hours) than toponym resolution approaches (from 2 minutes to 2.5 hours) except TopoCluster and Mordecai since the former was normally built on large language model, such as BERT and deals with more complex issues (disambiguating not only toponyms but also other types of entities) than the latter. TopoCluster is the slowest, taking nearly 191 hours while CLAVIN is the fastest, taking only 2 minutes. The time consumption for a voting ensemble equals the sum of the time of every individual approach that it combines. Therefore, the voting ensemble takes 72 hours for resolving 98,300 toponyms, which means on average resolving a toponym takes 2.6 seconds. There is a trade-off between correctness and speed.

4.7. ***Sensitivity analysis***

4.7.1. Configuration

We first investigate how the removal of an individual approach would affect the performance of voting ensembles. A basic voting ensemble is first proposed, including all the 20 individual approaches with each approach having one vote. A degraded ensemble is then constructed by removing one approach from the basic ensemble. We then subtract the average *ME, Accuracy@161km*, and *AUC* achieved by the degraded ensemble from that of the basic ensemble. The result is shown in Figure 12. Regarding *Accuracy@161km*, GENRE, BLINK, and SHS make the largest positive contribution, while CLAVIN and DBpedia Spotlight make the largest negative contribution. Regarding *ME*, GENRE and BLINK make the largest positive contribution, while CLAVIN and Population-Heuristics make the largest negative contribution. Regarding *AUC*, CamCoder, SHS, and Adaptive Learning make the largest positive contribution, while DCA and MulRel make the largest negative contribution. Generally, the disambiguation ability of each single approach determines their contribution to the voting ensemble, such as GENRE (with high disambiguation ability) and Population-Heuristics (with low disambiguation ability), which contribute positively and negatively, respectively.

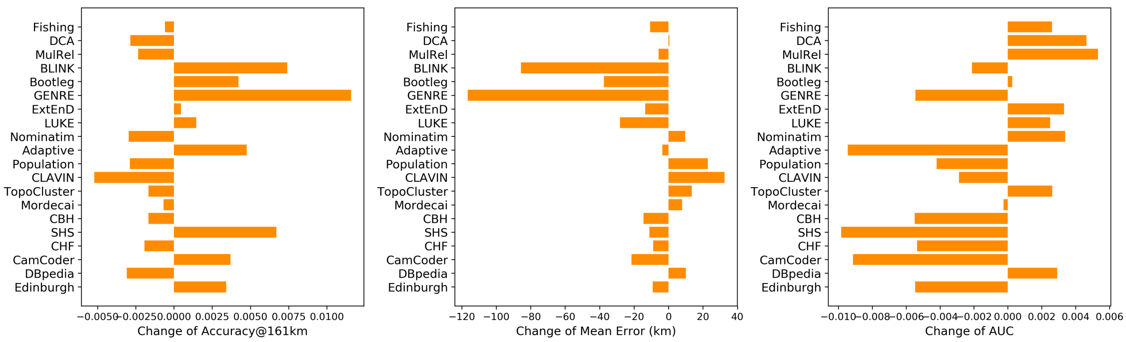

Figure 12. Change of *Accuracy@161km*, *AUC*, and *ME* when adding an approach to a voting ensemble.

4.7.2. Parameters

In the first experiment, the DBSCAN parameter *eps* was defined as $eps \in 1, ..., 800$ and a step size of 30. Figure 13 shows the performance of the voting ensemble as the change of *eps*. The red line denotes the performance of the best individual approach,

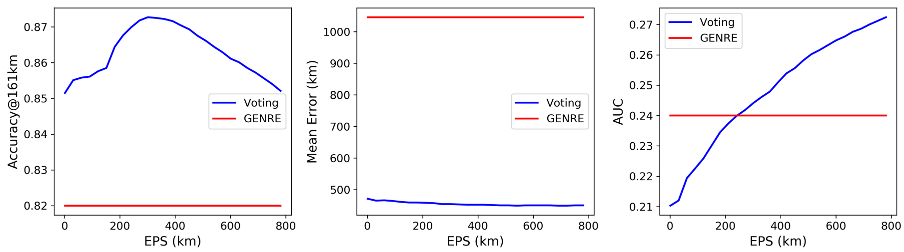

Figure 13. Impact of *eps* on the performance of the voting ensemble.

GENRE. We can see that *eps* has distinct impact on *ME, Accuracy@161km*, and AUC. The best *Accuracy@161km* is achieved when *eps* is set to 350 km, while as the increase of *eps*, *ME* decreases slightly from 470 km to 450 km, and AUC increases rapidly from 0.21 to 0.27.

In the second experiment, the DBSCAN parameter $minPts$ was defined as $minPts \in 1, ..., 11$ and a step size of 1. Figure 14 shows the result of the voting ensemble as the change of $minPts$. We can see $minPts$ has a large impact on the performance of the voting ensemble. When $minPts$ is set to 1 and 2, the best performance is reached. The higher the $minPts$, the worse the voting ensemble performs.

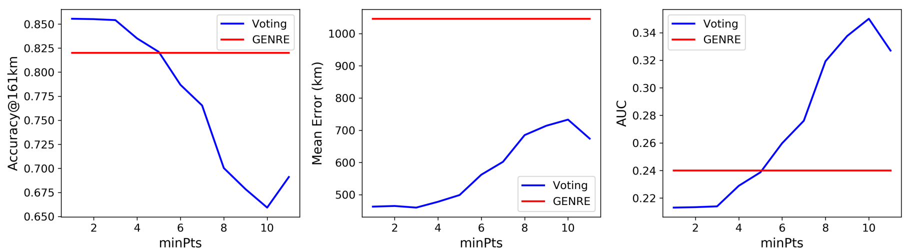

Figure 14. Impact of $minPts$ on the performance of the voting ensemble.

5. Conclusion

In this paper, we investigate how voting ensembles that combine several individual approaches can push state-of-the-art performance of toponym disambiguation. Experimental results on 12 public datasets of six types prove the generalizability and robustness of the voting approach. The deep learning-based ELs (i.e., GENRE and BLINK) that are pretrained on nearly 10 million Wikipedia entities show impressive disambiguation performance, performing much better than toponym resolutions approaches. However, there is a trade-off between correctness and speed since the voting approach and the two ELs take much more time than most of the others. Moreover, there is still space for improving the performance of resolving fine-grained places, such as POIs, natural features, and traffic ways. This will be one of our future research tasks. Furthermore, the idea of voting approaches can be extended to the more general issue, such as entity linking since combining several ELs can improve the disambiguation performance for toponyms, and might also for other entities, such as person, and organization.

6. Data and code availability

The code and data that support the findings of this study is available in GitHub with the link https://github.com/uhuohuy/toponym-disambiguation-voting.

7. Competing interests

The authors declare no competing interests.

References

2016–2022. "entity-fishing." `https://github.com/kermitt2/entity-fishing`.

Ardanuy, Mariona Coll, David Beavan, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Katherine McDonough, Federico Nanni, Daniel van Strien, and Daniel CS Wilson. 2022. "A Dataset for Toponym Resolution in Nineteenth-Century English Newspapers." *Journal of Open Humanities Data* 8.

Ardanuy, Mariona Coll, 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*, 175–180.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. "Dbpedia: A nucleus for a web of open data." In *The semantic web*, 722–735. Springer.

Barba, Edoardo, Luigi Procopio, and Roberto Navigli. 2022. "ExtEnD: Extractive Entity Disambiguation." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Online and Dublin, Ireland, May. Association for Computational Linguistics.

Cardoso, Ana Bárbara, Bruno Martins, and Jacinto Estima. 2021. "A Novel Deep Learning Approach Using Contextual Embeddings for Toponym Resolution." *ISPRS International Journal of Geo-Information* 11 (1): 28.

de Bruijn, Jens A, 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.

De Cao, Nicola, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. "Autoregressive Entity Retrieval." In *International Conference on Learning Representations*, `https://openreview.net/forum?id=5k8F6UU39V`.

DeLozier, Grant, Jason Baldridge, and Loretta London. 2015. "Gazetteer-independent toponym resolution using geographic word profiles." In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, .

DeLozier, Grant, Benjamin 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)*, 188–198.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* .

Fize, Jacques, Ludovic Moncla, and Bruno Martins. 2021. "Deep Learning for Toponym Resolution: Geocoding Based on Pairs of Toponyms." *ISPRS International Journal of Geo-Information* 10 (12): 818.

Gregory, Ian, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. 2015. "Geoparsing, GIS, and textual analysis: current developments in spatial humanities research." *International Journal of Humanities and Arts Computing* 9 (1): 1–14.

Gritta, Milan, Mohammad Pilehvar, and Nigel Collier. 2018. "Which melbourne? augmenting geocoding with maps." .

Gritta, Milan, Mohammad Taher Pilehvar, and Nigel Collier. 2020. "A pragmatic guide to geoparsing evaluation." *Language resources and evaluation* 54 (3): 683–712.

Gritta, Milan, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018. "What's missing in geographical parsing?" *Language Resources and Evaluation* 52 (2): 603–623.

Grover, Claire, 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.

Halterman, Andrew. 2017. "Mordecai: Full text geoparsing and event geocoding." *Journal of Open Source Software* 2 (9): 91.

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

Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. "Robust disambiguation of named entities in text." In *Proceedings of the 2011 conference on empirical methods in natural language processing*, 782–792.

Hu, Xuke, Hussein Al-Olimat, Jens Kersten, Matti Wiegmann, Friederike Klan, Yeran Sun, and Hongchao Fan. 2021. "GazPNE: Annotation-free Deep Learning for Place Name Extraction from Microblogs Leveraging Gazetteer and Synthetic Data by Rules." *International Journal of Geographical Information Science* 1–28.

Hu, Xuke, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2022a. "Location reference recognition from texts: A survey and comparison." *arXiv preprint arXiv:2207.01683* .

Hu, Xuke, Zhiyong Zhou, Yeran Sun, Jens Kersten, Friederike Klan, Hongchao Fan, and Matti Wiegmann. 2022b. "GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models." *IEEE Internet of Things Journal* 1–1.

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

Karimzadeh, Morteza, Scott Pezanowski, Alan M MacEachren, and Jan O Wallgrün. 2019. "GeoTxt: A scalable geoparsing system for unstructured text geolocation." *Transactions in GIS* 23 (1): 118–136.

Katz, Philipp, and Alexander Schill. 2013. "To learn or to rule: two approaches for extracting geographical information from unstructured text." *Data Mining and Analytics 2013 (AusDM'13)* 117.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems* 30.

Khan, Kamran, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. 2014. "DBSCAN: Past, present and future." In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, 232–238. IEEE.

Kulkarni, Sayali, 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* .

Le, Phong, and Ivan Titov. 2018. "Improving entity linking by modeling latent relations between mentions." *arXiv preprint arXiv:1804.10637* .

Lieberman, Michael D, 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*, 731–740.

Lieberman, Michael D, 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)*, 201–212. IEEE.

Mendes, Pablo N, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. "DBpedia spotlight: shedding light on the web of documents." In *Proceedings of the 7th international conference on semantic systems*, 1–8.

Milusheva, Sveta, Robert Marty, Guadalupe Bedoya, Sarah Williams, Elizabeth Resor, and Arianna Legovini. 2021. "Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning." *PloS one* 16 (2): e0244317.

Möller, Cedric, Jens Lehmann, and Ricardo Usbeck. 2022. "Survey on English Entity Linking on Wikidata: Datasets and approaches." *Semantic Web* (Preprint): 1–42.

Orr, Laurel, Megan Leszczynski, Simran Arora, Sen Wu, Neel Guha, Xiao Ling, and Christopher Re. 2020. "Bootleg: Chasing the tail with self-supervised named entity disambiguation." *arXiv preprint arXiv:2010.10363* .

Purves, Ross S, Paul Clough, Christopher B Jones, Mark H Hall, and Vanessa Murdock. 2018. "Geographic information retrieval: Progress and challenges in spatial search of text." *Foundations and Trends in Information Retrieval* 12 (2-3): 164–318.

Qi, Tao, Suyu Ge, Chuhan Wu, Yubo Chen, and Yongfeng Huang. 2019. "THU_NGN at SemEval-2019 Task 12: Toponym Detection and Disambiguation on Scientific Papers." In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 1302–1307.

Rayson, Paul, Alex Reinhold, James Butler, Chris Donaldson, Ian Gregory, and Joanna Taylor. 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*, 9–15.

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

Scott, Peter, Martin K-F Bader, Treena Burgess, Giles Hardy, and Nari Williams. 2019. "Global biogeography and invasion risk of the plant pathogen genus Phytophthora." *Environmental Science & Policy* 101: 175–182.

Sevgili, Özge, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. "Neural entity linking: A survey of models based on deep learning." *Semantic Web* (Preprint): 1–44.

Shook, Eric, and Victoria K Turner. 2016. "The socio-environmental data explorer (SEDE): a social media–enhanced decision support system to explore risk perception to hazard events." *Cartography and Geographic Information Science* 43 (5): 427–441.

Speriosu, Michael, 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)*, 1466–1476.

van Hulst, Johannes M., Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. "REL: An Entity Linker Standing on the Shoulders of Giants." In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20. ACM.

Wallgrün, Jan Oliver, 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.

Wang, Jimin, and Yingjie Hu. 2019. "Are we there yet? Evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform." In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*, 1–6.

Wang, Jimin, Yingjie Hu, and Kenneth Joseph. 2020. "NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages." *Transactions in GIS* .

Wang, Xiaobin, 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*, 917–923.

Weissenbacher, Davy, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez. 2019. "Semeval-2019 task 12: Toponym resolution in scientific papers." In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 907–916.

Wikipedia. 2004. *Wikipedia.* PediaPress.

Wing, Benjamin, and Jason Baldridge. 2011. "Simple supervised document geolocation with geodesic grids." In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 955–964.

Won, Miguel, Patricia Murrieta-Flores, and Bruno Martins. 2018. "ensemble named entity recognition (ner): evaluating ner Tools in the identification of Place names in historical corpora." *Frontiers in Digital Humanities* 5: 2.

Wu, Ledell, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. "Zero-shot Entity Linking with Dense Entity Retrieval." In *EMNLP*, .

Yamada, Ikuya, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. "Global Entity Disambiguation with BERT." In *NAACL*, Association for Computational Linguistics.

Yan, Zheren, Can Yang, Lei Hu, Jing Zhao, Liangcun Jiang, and Jianya Gong. 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.

Yang, Xiyuan, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, and Xiang Ren. 2019. "Learning dynamic context augmentation for global entity linking." *arXiv preprint arXiv:1909.02117* .