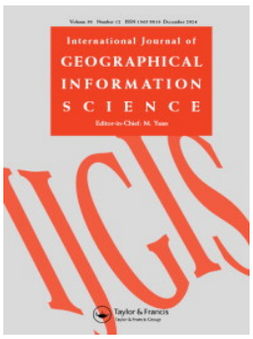


International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgis20

Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge

Xuke Hu, Jens Kersten, Friederike Klan & Sheikh Mastura Farzana

To cite this article: Xuke Hu, Jens Kersten, Friederike Klan & Sheikh Mastura Farzana (24 Sep 2024): Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2024.2405182

To link to this article: https://doi.org/10.1080/13658816.2024.2405182

Published online: 24 Sep 2024.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2024.2405182

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge

Xuke Hu[a], Jens Kersten[a], Friederike Klan[a] and Sheikh Mastura Farzana[b]

[a]Institute of Data Science, German Aerospace Center (DLR), Jena, Germany; [b]Institute of Software Technology, German Aerospace Center (DLR), Cologne, Germany

ABSTRACT

Toponym resolution is crucial for extracting geographic information from natural language texts, such as social media posts and news articles. Despite the advancements in current methods, including state-of-the-art deep learning solutions like GENRE and a sophisticated voting system that integrates seven individual methods, further enhancing their accuracy is essential. To achieve this goal, we propose a novel method that combines lightweight and open-source large language models and geo-knowledge. Specifically, we first fine-tune Mistral (7B), Baichuan2 (7B), Llama2 (7B & 13B), and Falcon (7B) to estimate toponyms' unambiguous reference (e.g., city, state, country) given their contexts. Subsequently, we correct inaccuracies in generated references and determine their geo-coordinates via sequentially querying GeoNames, Nominatim, and ArcGIS geocoders until a successful geocoding result is achieved. Our methods demonstrate enhanced performance compared to 20 existing methods, as evidenced across seven challenging datasets including 83,365 toponyms worldwide, with the Mistral-based method leading, followed by Baichuan2, Llama2, and Falcon-based methods. Specifically, the Mistral-based method achieves an *Accuracy@161km* of 0.91, surpassing GENRE, the best individual method, by 17% and the seven-methods composite voting system by 7%. Moreover, our methods are computationally efficient, operable on one general GPU, have modest memory requirements (14 GB for 7B models and 27 GB for 13B models), and exceed both GENRE and the voting system in inferring speed.

ARTICLE HISTORY
Received 20 February 2024
Accepted 11 September 2024

KEYWORDS
Geoparsing; toponym resolution; geocoding; large language model

1. Introduction

In this era, characterized by a vast array of semi-structured and unstructured natural language texts, a hidden treasure trove of geographic information awaits discovery. This information is embedded in a wide range of sources, from social media posts and news articles to scientific publications and historical documents, often in the form of

CONTACT Xuke Hu xuke.hu@dlr.de

Figure 1. Example of toponym ambiguity. Around 400 different places worldwide are named '*Victoria Park*', such as the park in London, Ontario, the park in London Borough of Tower Hamlets, London, UK, and the neighborhood in Los Angeles, California.

toponyms and place names (Hu *et al*. 2023b). This wealth of information offers a unique geospatial lens that deepens our understanding of our world (Hu and Adams 2021). Extracting the geographic information from texts, also named geoparsing, is important for a variety of applications across domains such as spatial humanities (Gregory *et al*. 2015), geographic search (Purves *et al*. 2018), disaster management (Zhang *et al*. 2021), urban planning (Milusheva *et al*. 2021), and epidemiological studies (Scott *et al*. 2019). Geoparsing consists of two steps: toponym recognition, which is to recognize toponyms mentioned in texts, and toponym resolution or geocoding, which is to determine the geospatial representation or geo-coordinates of the toponyms. While significant progress has been made in toponym recognition (Hu *et al*. 2022a, 2022b, 2023c), toponym resolution remains a complex challenge due to the inherent ambiguity of toponyms. For instance, in the example of Figure 1, there are nearly 400 global locations named '*Victoria Park*', according to data retrieved from OpenStreetMap (OSM).[1]

The resolution of toponyms has been addressed using two distinct methods. The first method, traditional toponym resolution, specifically targets toponyms or location entities. On the other hand, the second method employs entity linkers, which extend beyond mere toponym resolution by associating a broader range of entities—including Person, Organization, and Location—with corresponding entries in knowledge bases (KBs) such as Wikipedia (2004), Wikidata (Vrandečić and Krötzsch 2014), and DBpedia (Auer *et al*. 2007). Recent advancements in deep learning have significantly improved toponym resolution accuracy. Notable examples include CamCoder (Gritta *et al*. 2018) and sophisticated entity linkers such as BLINK (Wu *et al*. 2020b), GENRE (De Cao *et al*. 2021), and ReFinED (Ayoola *et al*. 2022). Furthermore, we have proposed an ensemble method (Hu *et al*. 2023a), known as a voting system to further enhance the accuracy. It integrates seven distinct approaches including GENRE, BLINK, and CamCoder. Despite these advancements, the average accuracy of the most advanced method at predicting locations within 161 km (*Accuracy@161km*) remains at 0.84, as documented in our previous work. Therefore, there is still a need to further improve the accuracy of these methods.

In the rapidly evolving field of natural language processing, large language models (LLMs) such as GPT-4 have marked a transformative era, greatly influencing both academic research and practical application development across various sectors, including

geospatial science (Ji and Gao 2023, Hochmair *et al*. 2024). There is also a growing interest and application of LLMs in geoparsing, a subfield of geospatial science. Studies like (Hu *et al*. 2023c) have employed GPT-based models, including GPT-3 and ChatGPT, to effectively identify location mentions from social media data, a critical initial step in geoparsing. Unlike toponym recognition, toponym resolution depends heavily on extensive geo-knowledge, typically found in comprehensive gazetteers like OSM. LLMs, despite their advanced capabilities, are not equipped with complete geo-knowledge (Mai *et al*. 2024). Moreover, large-size models like GPT4, which has about 1.76 trillion parameters, necessitate significant computational resources and energy consumption. This presents a significant limitation for their practical application in geoparsing, where efficient and cost-effective processing of large volumes of text data is crucial, such as in creating geographic indices for daily generated terabytes of documents to support geographic search.

This study is motivated by the question: *How can we develop an enhanced method for toponym resolution that is both accurate and computationally efficient?* In response, we propose combining lightweight (e.g., 7 or 13 billion parameters) and open-source LLMs with geo-knowledge for toponym resolution. This method not only aims to elevate accuracy but also ensures compatibility with standard computing hardware and adheres to open-source principles. Specifically, we first fine-tune pre-trained models such as Mistral (7B) (Jiang *et al*. 2023), Baichuan2 (7B) (Baichuan 2023), Llama2 (7B & 13B) (Touvron *et al*. 2023), and Falcon (7B) (Penedo *et al*. 2023) to estimate toponyms' unambiguous reference, including city, state, and country information, such as rendering *'Paris, Texas, United States'* for *'Paris'* in the text *'I live in Paris, a city of TX'*. This allows us to interpret the intended meaning of toponyms within texts. However, comprehensive toponym resolution requires access to extensive properties beyond mere unambiguous reference, including geo-coordinates and even detailed geospatial representations (e.g., polygons), population statistics, type, and administrative level information, which are not typically provided by standard LLMs. Moreover, instances occur where the references generated by the fine-tuned models are inaccurate. For instance, the reference *'Dean Woods Road, Metcalfe County, Kentucky, United States'* provided for *'Dean Woods Road'* is not existing, whereas the correct reference should be *'Dean Woods Road, Adair County, Kentucky, United States'*. To address these limitations, we enhance our method by combining three geocoding services including Nominatim,[2] GeoNames,[3] and ArcGIS.[4] This step can not only fixes inaccuracies in generated references (e.g., correctly geocode the inaccurate reference *'Dean Woods Road, Metcalfe County, Kentucky, United States'*) but also acquires geo-coordinates, detailed geospatial representations, and additional toponym attributes. Our method is rigorously benchmarked against the latest and commonly used 20 methods for toponym resolution across 7 diverse datasets. This paper's central contribution is proposing an advanced toponym resolution method that leverages the capabilities of lightweight and open-source LLMs and geo-knowledge, which can run efficiently on a single general GPU.

The paper is structured as follows: Section 2 provides an overview of the existing literature on deep learning-based entity linking and the use of LLMs in geospatial science. Section 3 details our proposed method. In Section 4, we present the

experimental evaluation and findings. Section 5 discusses bias issues. Finally, Section 6 concludes the paper and outlines potential avenues for future research.

2. Related works

In our prior study (Hu *et al*. 2023a), we reviewed traditional toponym resolution approaches, which were classified into three groups: rules, learning and ranking, and learning and classification, but did not address entity linkers. This section will focus on an in-depth review of the latest advancements in deep learning-based entity linkers. Additionally, we will discuss utilizing LLMs, especially generative AI, in geospatial science and in its subfield, geoparsing. For an in-depth examination of techniques and theoretical foundations of LLMs, refer to the survey papers by Chang *et al*. (2024) and Min *et al*. (2023).

2.1. Deep learning-based entity linking

Entity linking (Carmel *et al*. 2014), a generic task in text understanding, consists of two main steps: entity recognition that detects named entities, such as Persons, Organizations, and Locations, in texts, and entity disambiguation which involves associating mentioned entities in texts with their corresponding entries in knowledge bases like Wikipedia. Toponym resolution is the specialized form of entity disambiguation.

Guo and Barbosa (2018) presented the walking named entity disambiguation method, which employs information-theoretic semantic relatedness and random walks on disambiguation graphs for named entity disambiguation. This approach includes a revised iterative algorithm and a new learning-to-rank method. Yang *et al*. (2019) introduced dynamic context augmentation (DCA) for entity linking, sequentially accumulating and utilizing context from previously linked entities within a document. This approach, employing supervised and reinforcement learning, incorporates entity properties and relationships and uses attention mechanisms to manage relevance and reduce error propagation.

To improve model performance, numerous studies have employed unsupervised or semi-supervised techniques, such as training models on unlabeled data from Wikipedia and Wikidata. For instance, Orr *et al*. (2020) proposed Bootleg, a self-supervised entity disambiguation system. By extracting relational and contextual information from Wikipedia and Wikidata, the system self-learns entity and relation embeddings, particularly focusing on rare or *'tail'* entities. Wu *et al*. (2020a) introduced BLINK (Bi-encoder for Linking Knowledge), a scalable zero-shot entity linker. BLINK uses a two-stage process with a bi-encoder and cross-encoder for entity linking. The model's bi-encoder is pre-trained on Wikipedia and further trained on specific datasets such as the WikilinksNED (Onoe and Durrett 2020) unseen-mentions dataset with around 2.2 M examples. GENRE (De Cao *et al*. 2021), a system for autoregressive entity retrieval, is pre-trained on the BART (Lewis *et al*. 2019) language model and BLINK dataset with 9 M document-mention-entity triples from Wikipedia. It is further fine-tuned on datasets like AIDA-CoNLL (Hoffart *et al*. 2011) for entity disambiguation.

Barba *et al*. (2022) presented EXTEND, which treats entity disambiguation as text extraction rather than classification tasks and introduced two Transformer-based models. Yamada *et al*. (2022) introduces LUKE, an entity disambiguation technique based on BERT (Devlin *et al*. 2019), which combines words and entities as input tokens. It is trained using a large entity-annotated corpus obtained from Wikipedia. The model's training task involves predicting masked entities, similar to BERT's masked language model objective. Ayoola *et al*. (2022) proposed ReFinED, which combines mention detection, fine-grained entity typing, and entity disambiguation in a single forward pass. It targets Wikidata, enabling it to link to more entities than models primarily relying on Wikipedia.

The utilization of advanced entity linkers, notably those pre-trained on extensive Wikipedia or Wikidata documents, has exhibited superior proficiency in disambiguating entities, including toponyms. Nonetheless, the evaluation results in our previous research (Hu *et al*. 2023a) also illuminate the room for optimization in the performance of existing entity linkers for toponym resolution tasks.

2.2. LLMs for geospatial science

The advent of LLMs has substantially enhanced the ability of machines to understand complex user queries and boosted language processing effectiveness, thereby benefiting numerous domains, including geospatial science. Many studies have investigated the potentials and constraints of LLMs within this domain (Ji and Gao 2023, Mooney *et al*. 2023, Tao and Xu 2023, Xie *et al*. 2023, Yin *et al*. 2023, Hochmair *et al*. 2024). For example, Xie *et al*. (2023) examined the application limits of AI foundation models in geospatial contexts and advocated for tailored geo-foundation models. It spotlights the divergence between conventional data types, such as natural language texts and video, for which AI models are designed, and geospatial data, pointing out the necessity for innovation in this domain to advance related fields. Ji and Gao (2023) assessed how well LLMs interpret and represent spatial concepts described in texts. It reveals that while these models can grasp basic spatial relationships, they struggle with more complex geometric reasoning and specific spatial tasks, such as distance measuring, suggesting a gap that needs to be bridged for advanced geospatial artificial intelligence (GeoAI) functions. Hochmair *et al*. (2024) compared the performance of ChatGPT-4, Bard, Claude-2, and Copilot on various geospatial tasks, such as spatial literacy, GIS concepts, and mapping. Results indicate that ChatGPT-4 outperformed other chatbots across most tasks. Juhász *et al*. (2023) explored the potential of using GPT-3.5-turbo to enrich OpenStreetMap (OSM) by suggesting the most appropriate tagging (e.g., *'highway'='primary'*, *'lanes'= 3*) for each road in OSM based on derived descriptions of Mapillary images. Tao and Xu (2023) used ChatGPT to design thematic maps given public geospatial data and to create mental maps based on textual descriptions of geographic space. Yin *et al*. (2023) introduced a benchmark to assess the GPT-3 model's performance in geocoding address parsing, comparing it with three transformer-based models and one LSTM-based model. The dataset included 21 input errors/variations from real user logs and diverse address formatting across the U.S. The

results showed that the Bidirectional LSTM-CRF model slightly outperformed the transformer-based models, including GPT-3.

Some research has also been conducted on applying LLMs in geoparsing, a subfield of geospatial science. For instance, Li *et al*. (2023a) introduced GeoLM, a geospatial language model that was pre-trained using data from sources like Wikipedia, Wikidata, and OpenStreetMap. GeoLM can be adapted to support multiple downtown tasks, such as toponym recognition, resolution, relation extraction, and geo-entity typing tasks. In the relation extraction task, GeoLM outperforms GPT-3.5. Mai *et al*. (2022) explored the potential of leveraging LLMs, like GPT-2 and GPT-3, for GeoAI. They demonstrate that LLMs outperform task-specific models in geospatial semantics tasks, including toponym recognition and location description recognition. Hu *et al*. (2023c) investigated the utility of various GPT models, including GPT-2, GPT-3, ChatGPT, and GPT-4, for the extraction of location descriptions from social media during disaster events. They compared these models against traditional Named Entity Recognition (NER) tools and a fine-tuned BERT model. The study utilized a dataset of annotated tweets from Hurricane Harvey. The results demonstrate that geo-knowledge-guided GPT models significantly outperformed NER tools and the fine-tuned BERT model in recognizing both complete location descriptions and their associated categories.

These studies underscore the considerable potential of LLMs in geoparsing. On one hand, utilizing much larger models, such as GPT-3.5 and 4, can substantially enhance performance. On the other hand, relying on large-sized models also limits the applicability as they cannot be deployed on standard hardware and consume considerable energy.

3. Proposed approach

Our approach, outlined in Figure 2, is a two-stage process involving training (fine-tuning) and geocoding. We fine-tune four LLMs, including Mistral (7B), Baichuan2 (7B), Llama2 (7B & 13B), and Falcon (7B) in the training phase using one toponym resolution dataset. The models are trained to interpret a given toponym within its narrative context as the input and generate the unambiguous reference (e.g., city, county, state, country) as the output. During geocoding, we fix the inaccurately generated

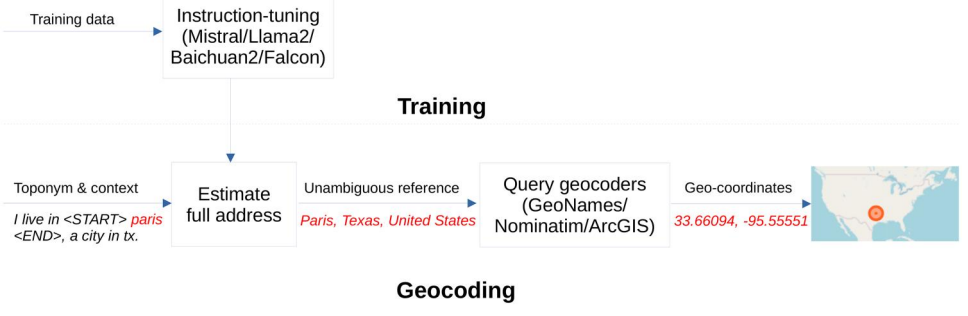


Figure 2. Workflow of the proposed approach. In the example, '*Paris*' is the target toponym.

references and determine toponyms' geo-coordinates or geospatial representations by querying the reference against three geocoders—Nominatim, GeoNames, and the ArcGIS geocoder. Detailed discussions of each stage follow.

3.1. Training

During the training phase, we first prepare the training data, followed by fine-tuning the four distinct models. The LGL[5] (Local-Global Lexicon) corpus is our main training data, which was created by Lieberman *et al*. (2010), containing 588 human-annotated news articles published by 78 local newspapers. We selected this corpus due to its high quality and the abundance of ambiguous toponyms stemming from its diverse local newspaper sources. Annotations in the corpus include the GeoNames ID for each toponym, which we utilize to extract their unambiguous reference. For Mistral, Llama2, and Falcon, we prepare it as the standard format of Alpaca instruction-tuning dataset,[6] which encompasses three parts: **Instruction**, **Input**, and **Output**. Their definitions are as follows:

- **Instruction** is a directive to the model, i.e.,'Identify the unambiguous reference of {toponym} (marked with «START» and «END») in the text.'
- **Input** refers to the text where the toponym is embedded, with the toponym itself enclosed in «START» and «END» tags.
- **Output** corresponds to the unambiguous reference for the marked toponym, including the city, county, state, and country information. A correctly identified reference allows us to deduce the precise geographic meaning of the toponym.

Figures 3 and 4 illustrate two training examples. The LGL corpus is characterized by predominantly lengthy texts, often exceeding 200 words, and primarily features locations within the US. To broaden the diversity and improve the representation within our training data, we augmented it with an additional 145 training examples generated by ChatGPT. These examples contain shorter texts and cover locations globally. Each generated example was subsequently manually reviewed and adjusted for accuracy. Overall, our training corpus includes a total of 4,291 examples. Note that, for

Instruction: Identify the unambiguous reference of Hurricane (marked with «START» and «END») in the text.

Input: Sheriff's deputies and «START» **Hurricane** «END» police found the two.'The guys were down in the area for about 10 minutes when they found them near the tracks,' the dispatcher said. Neither of the teenagers were hurt and both are safe, the dispatcher said. Keya Phillips and Seth Pettry, both 14, were found Monday after a tip came in that the two were hiding under a railroad bridge in Hurricane, according to a Putnam County dispatcher.

Output: Hurricane, Putnam County, West Virginia, United States

Figure 3. Training example illustrating unambiguous reference estimation for *'Hurricane'*.

Instruction: Identify the unambiguous reference of Scottish Highlands (marked with ≪START≫ and ≪END≫) in the text.

Input: The ≪START≫ **Scottish Highlands** ≪END≫ are a rugged and breathtaking part of Scotland. My recent hiking expedition took me through the misty glens and towering peaks. I enjoyed the tranquility of Loch Ness and the hospitality of local Highland communities. The Scottish Highlands are a paradise for nature enthusiasts.

Output: Scottish Highlands, Scotland, United Kingdom

Figure 4. Training example illustrating unambiguous reference estimation for *'Scottish Highlands'*.

Baichuan2, a different format[7] was required. Therefore, we further converted the Alpaca instruction-tuning dataset to meet the Baichuan2 format requirements.

We found that the formulation of training examples influences the model's performance. For instance, modifications to the instruction field, such as omitting the explicit mention of the target toponym (e.g., *'Identify the unambiguous reference of the toponym marked with «START» and «END» in the text.'*) resulted in a significant decrease in model effectiveness. Furthermore, we attempted to fine-tune the model to directly generate the geo-coordinates (latitude and longitude) of toponyms. However, this approach did not yield satisfactory results. The model either generated inaccurate geo-coordinates or the geo-coordinates of the same toponym varied each time. This could be due to the fact that LLMs do not possess precise knowledge of the geo-coordinates of toponyms. Moreover, a mere latitude and longitude coordinate pair is often insufficient. Frequently, additional information about the toponym, such as its type, administrative levels, and complete geospatial representation, for instance, as a polygon or line segment, is required. In contrast, the parental administrative units of toponyms are more likely to be included in the knowledge of LLMs. Therefore, we opted to instruct the model to output the parental administrative units (e.g., city, state, and country information) of a toponym, and subsequently utilize geocoders to acquire the accurate geo-coordinates and other relevant information of the toponyms.

In our methodology, we utilized the Low-Rank Adaptation (LoRA) (Hu *et al*. 2021) technique for fine-tuning LLMs on standard GPUs with limited resources. This method has been validated for its effectiveness in numerous downstream tasks (Lermen *et al*. 2023, Nguyen *et al*. 2023, Li *et al*. 2023b). Rather than adjusting all the weights in the weight matrix of a pre-trained model, LoRA fine-tunes two smaller matrices that approximate the larger matrix, forming the so-called LoRA adapter. Once fine-tuned, this adapter is integrated into the pre-trained model for inference purposes. This study applied this technique to four open-source LLMs.

- **Mistral** (Jiang *et al*. 2023), released in September 2023 by Mistral AI. Currently, it comprises several versions, including a 7B model.
- **Baichuan2** (Baichuan 2023), released by Chinese Baichuan Intelligent company in October 2023, is a series of large-scale multilingual language models containing 7 billion and 13 billion parameters.

- **Llama2** (Touvron *et al*. 2023), released by Meta in July 2023, includes 12 models ranging from 7B to 70B parameters.
- **Falcon** (Penedo *et al*. 2023), released in May 2023, is a language model family by the Technology Innovation Institute, including three versions: 7B, 40B, and 180B.

3.2. Geocoding

In this phase, our process begins with employing the fine-tuned model to deduce the unambiguous reference of a toponym from its contextual information. Subsequently, this inferred reference is used to query geocoders to ascertain its geographical coordinates. This step not only rectifies inaccuracies in generated references but also acquires geo-coordinates, detailed geospatial representations, and additional toponym attributes. For instance, consider the reference *'Dean Woods Road, Metcalfe County, Kentucky, United States'*, where the county information is erroneous. Despite this discrepancy, the reference can still be geocoded to a location near its true geographic position. Multiple geocoding services exist, such as Nominatim, GeoNames, and ArcGIS geocoder. Each of these services, while useful, is not without its limitations, often encountering difficulties in geocoding a reference accurately due to either stringent requirements for precise reference inputs or the incomplete global representation of locations. To mitigate these limitations, we employ a strategy of sequential querying among the three geocoders, prioritizing them in the order of GeoNames, Nominatim, and then the ArcGIS geocoder.[8] This approach ensures that if one service fails to yield a result, the next geocoder in the sequence is immediately queried, thereby enhancing our geocoding process's overall reliability. The Google Maps API[9] is widely recognized for its comprehensive coverage and reliability in geocoding. However, it is a costly service, despite offering a monthly free access limit of 40,000 requests (valued at 200 $). Given that our test datasets contain nearly 80,000 toponyms, which will be discussed in subsequent sections. Each of the fine-tuned LLMs generates different references for the same toponyms, resulting in a high demand for geocoding requests during experiments. Furthermore, we are continuously adjusting the fine-tuning strategies to improve each LLM and then testing them, which would further increase the demand for geocoding requests on a daily basis. Therefore, due to its cost, the Google Maps API has not been utilized in our study.

4. Experiments and evaluation

In this section, we initially set the parameters. This is followed by a brief introduction to the test datasets, evaluation metrics, and an overview of 20 existing approaches which can be used to resolve toponyms. We then compare our approaches with these existing approaches, focusing on their accuracy and computational efficiency.

4.1. Parameter setting

In our experimental setup, the training corpus was partitioned into training and evaluation subsets at a ratio of 9:1. The LoRA technique was pivotal in our approach, with

the LoRA attention dimension set to 8 and the LoRA alpha, the scaling parameter, fixed at 16. Additionally, we set the LoRA dropout rate to 0.1. Optimization during fine-tuning was achieved using the *AdamW* optimizer with a learning rate of 0.003. We set these parameters based on numerical experimental results to ensure optimal performance. It is crucial to acknowledge that the prediction outcomes from the model are not perfectly indicative of the final geocoding accuracy. For example, given a gold standard annotation such as '*Paris, Lamar County, Texas, United States*', variations in prediction like '*Paris, Texas, United States*' or '*Paris, Texas, US*'—though not exact matches—are still valid interpretations of '*Paris*' in the sentence '*I live in Paris, a place in Tx*'. Consequently, training or evaluation losses do not fully mirror the true accuracy and performance of the trained model. To select the most effective model, we examined models from steps near the point where the training or evaluation loss stopped decreasing.

The fine-tuning was conducted on an NVIDIA Tesla V100 GPU equipped with 40GiB RAM, while the actual GPU memory consumption varied depending on the model's size. Specifically, for the 7B models, approximately 14 GB of GPU memory was utilized, whereas the larger 13B model required about 27 GB of GPU memory.

4.2. Test data for toponym resolution

We utilized 7 public datasets as test data, details of which are summarized in Table 1. The geographical spread of the toponyms in the test datasets is depicted in Figure 5. Note that different studies employ varying definitions of toponyms (Wang and Hu 2019). Gritta *et al*. (2020) proposed a taxonomy that classifies toponyms into multiple types, including literal (e.g., 'Earthquake in *Turkey* is serious.'), demonyms (e.g., *Canadian*), metonymy (e.g., '*Mexico* changed the law.'), and languages (e.g., *Spanish* and *Chinese*). This study utilizes datasets that encompass all types of toponyms as defined by Gritta *et al*. (2020). The 7 datasets employed in our study are detailed as follows:

- **TR-News**: Developed by Kamalloo and Rafiei (2018), this dataset comprises news articles from various sources.[10]
- **GeoWebNews**: Assembled by Gritta *et al*. (2018), it includes news articles collected during the first week of April 2018.[11]
- **GeoCorpora**: Curated by Wallgrün *et al*. (2018), this dataset features tweets from various global events in 2014 and 2015.[12]

Table 1. Summary of the 7 test datasets.

Name	Text Count	Toponym Count	Type	KB/Gazetteer
TR-News	118	1,319	News	GeoNames
GeoWebNews	200	2,601	News	GeoNames
GeoCorpora	6,648	3,100	Tweet	GeoNames
WikToR	5,000	25,242	Wiki article	Wikipedia
WOTR	1,644	11,795	History	GeoNames
CLDW	62	34,713	History	GeoNames
NCEN	455	4,595	History	Wikipedia

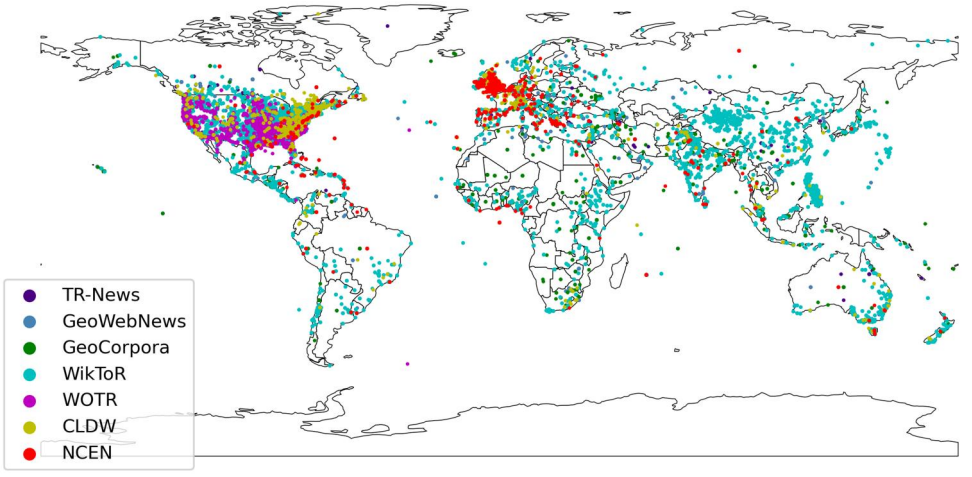


Figure 5. Geographical spread of the 83,365 toponyms from the 7 datasets.

- **WikToR**: Created automatically by Gritta *et al*. (2018), it consists of 5,000 Wikipedia articles, rich in ambiguous toponyms.[13]
- **WOTR**: This dataset, crafted by DeLozier *et al*. (2016), is based on the Official Records of the War of the Rebellion.[14]
- **CLDW** (Corpus of Lake District Writing): Formulated by Rayson *et al*. (2017), it encompasses writings about the English Lake District from the seventeenth to the early twentieth century.[15]
- **NCEN** (Nineteenth-Century English Newspapers): Created by Ardanuy *et al*. (2022), it comprises news articles published in England between 1780 and 1870.[16]

It is important to note that entity linkers typically employ Wikipedia as the primary target KB. However, most datasets in our study link toponyms to GeoNames, with exceptions being WikToR and NCEN. A significant challenge we encountered pertains to the inconsistent geocoding of certain coarse-grained locations, such as countries, between Wikipedia and GeoNames. For instance, '*United States*' is geocoded as (40, -100) in Wikipedia and (39.76, -98.5) in GeoNames, while '*China*' is represented as (35, 103) and (35, 105) in Wikipedia and GeoNames, respectively. Such discrepancies in the datasets pose a risk of incorrect evaluations. We excluded 29 frequently misaligned places from the evaluation process, as detailed in our previous study (Hu *et al*. 2023a). One potential approach to address the limitations of point-based evaluation is to use toponyms' geospatial representations, such as polygons or line segments, which can provide more accurate evaluations. However, this method imposes high demands on both the datasets and the methods used. The datasets used in our study only provide geo-coordinates or a link to Wikipedia or GeoNames, from which we cannot obtain detailed geospatial representations beyond just geo-coordinates for most toponyms. Similarly, most methods link toponyms to either GeoNames or Wikipedia, making geospatial representation-based evaluation infeasible at present.

Additionally, our previous study (Hu *et al*. 2023a) used 12 publicly available datasets. In this paper, we have selected 7 of these 12 datasets, excluding TUD-Loc-2013

(Katz and Schill 2013), NEEL,[17] GeoVirus (Gritta *et al*. 2018), SemEval (Weissenbacher *et al*. 2019), and LGL. This decision was based on our observation that the former four datasets predominantly contain simple, coarse-grained, and unambiguous place names, such as country names. The effectiveness of resolving these names largely depends on the gazetteers or knowledge bases employed, which may present inconsistencies in the coordinates of coarse-grained places. As LGL serves as the training set in our approach, it has also been omitted from the method evaluation to avoid bias.

4.3. Evaluation metrics

Our evaluation incorporates three principal metrics, as defined in Gritta *et al*. (2020), crucial for a comprehensive analysis of geocoding accuracy and error:

- **Accuracy@161km:** This metric calculates the percentage of geocoding errors within a 100-mile (161 km) range.
- **Mean Error (ME):** This metric computes the average distance error for all toponyms.
- **Area Under the Curve (AUC):** The AUC for toponym resolution quantifies geocoding accuracy by using the Trapezoid Rule[18] to integrate the area under a curve of logarithmically adjusted errors, diminishing the influence of outliers for a more balanced evaluation.

4.4. Compared approaches

Table 2 enumerates various representative approaches. The Voting approach integrates seven distinct approaches—GENRE, BLINK, LUKE, CamCoder, Edinburgh geoparser (Grover *et al*. 2010), CBH, and SHS—into a unified voting system. For detailed descriptions of the voting system and other approaches, refer to our previous study (Hu *et al*.

Table 2. Summary of 20 representative approaches.

Name	Method Type
Entity-Fishing[20]	entity linker
MulRel-NEL (Le and Titov 2018)	entity linker
DCA (Yang *et al*. 2019)	entity linker
BLINK (Wu *et al*. 2020b)	entity linker
Bootleg (Orr *et al*. 2020)	entity linker
GENRE (De Cao *et al*. 2021)	entity linker
ExtEnD (Barba, Procopio, and Navigli2022)	entity linker
LUKE (Yamada *et al*. 2022)	entity linker
ReFinED (Tom Ayoola 2022)	entity linker
Nominatim	TR (rule)
ArcGIS	TR (rule)
Population (Speriosu and Baldridge 2013)	TR (rule)
Adaptive learning (Lieberman and Samet 2012)	TR (learning & ranking)
CLAVIN[21]	TR (rule)
TopoCluster (DeLozier, Baldridge, and London 2015)	TR (learning & classification)
Mordecai (Halterman 2017)	TR (rule)
CBH, SHS, CHF (Kamalloo and Rafiei 2018)	TR (rule)
CamCoder (Gritta *et al*. 2018)	TR (learning & classification)
Voting (Hu *et al*. 2023a)	hybrid

TR denotes toponym resolution.

2023a). As a standard practice, we assume that all toponyms are accurately identified, allowing us to input the gold-standard toponyms directly into the entity disambiguation or toponym resolution processes. For entities associated with Wikipedia or Wikidata, we extract their geographical coordinates based on the available geo-properties. In instances where associated Wikipedia or Wikidata entities lack geographical annotations, we assign (0,0) coordinates, indicating an invalid estimation. For such cases, we set a distance error of 20,039 km, the maximum possible error on Earth, as per Gritta *et al*. (2020).

In addition to the existing approaches, we have expanded our evaluation to include a much larger model with fine-tuning, namely Llama2 (70B). This is done to provide a more comprehensive assessment of the impact of model size on the accuracy of the proposed method. The fine-tuning and testing process for Llama2 (70B) is consistent with the approach used for the other five lightweight models.

4.5. Results

We merge toponyms from the 7 test datasets, with toponym counts ranging from 1,319 to over 30,000, and then collectively calculate the three metrics, *Accuracy@161km*, *ME*, and *AUC*, ensuring a more equitable evaluation. The comparative performance of existing methods and our proposed methods is depicted in Figure 6. In the figure and the subsequent ones, as well as the tables, FT is used as an abbreviation for 'fine-tuned'.

Our proposed methods, which are based on fine-tuned models, exhibit significant improvements over the existing ones. The method that fine-tunes a much larger model, Llama2 (70B), is the best performer. However, the increase in *Accuracy@161km*, from 0.91 achieved by the fine-tuned Mistral (7B) model to 0.93, is not substantial, given the significant increase in size from 7B to 70B. While among the lightweight models (7B or 13B), the method based on fine-tuned Mistral (7B) stands out. It surpasses the best prior approach, the voting approach, by 7%, 64%, and 10% in the aforementioned metrics. When compared to the best individual method, GENRE, it shows even more significant improvements of 17%, 90%, and 32% in the same metrics. In 91% of the cases, the distance error is below 161 km, which satisfies the

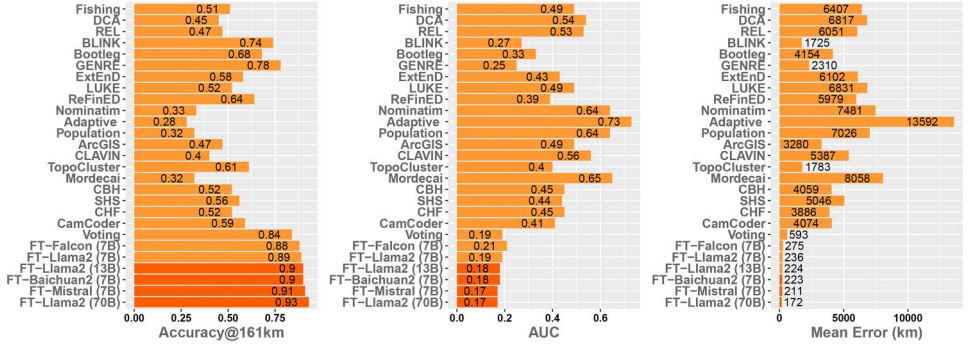


Figure 6. *Accuracy@161km* (↑), *AUC* (↓), and *ME* (↓) for each approach on the complete test datasets. The top 3 scores are highlighted.

requirements of many practical applications. For instance, this allows for the search of relevant content, such as news, in the vicinity of a certain city. Additionally, it is observed that the Llama2 (13B)-based approach marginally outperforms its 7B counterpart. Among the existing techniques, the voting approach, combining seven individual methods, emerges as the most accurate, followed by GENRE and BLINK—both well-regarded entity linkers.

The comprehensive results for both established and novel methods across each dataset are detailed in Tables 3–5. Our proposed approaches demonstrate superior performance over existing techniques in most assessed metrics within the 7 datasets. An outlier is the WOTR dataset, where the voting approach equals the performance of the Mistral (7B)-based approach and surpasses the other methods. We observe that the proposed approaches perform particularly well on the WikToR dataset, achieving an impressive accuracy of 0.98. This is largely due to the dataset's composition, which includes Wikipedia articles about places where higher-level administrative units are mentioned within the text. LLMs are adept at extracting key surface information from the text. For example, in Figure 7, the higher-level administrative units of '*Santa Cruz*' (i.e., '*Davao del Sur, Philippines*') are included in the text, allowing the model to accurately infer its unambiguous reference. Moreover, in numerous scenarios, the fine-tuned models demonstrate the capability to deduce unambiguous references even without explicit mention of higher-level administrative units. This ability stems from their vast pre-trained data, which encompasses basic geographic knowledge. This is exemplified in Figure 8, where the toponym '*OleMiss*' is correctly resolved by the fine-tuned model without explicit higher-level administrative units in the texts. These two strengths significantly contribute to the superior performance of our proposed methods in toponym resolution.

Despite these strengths, a noted limitation of the models is their occasional propensity to generate fictitious references. This issue is caused mainly by their incomplete geographic knowledge, which, unlike exhaustive gazetteers such as OpenStreetMap, may encompass limited geographical locations. Consequently, the models sometimes produce references that are plausible in structure yet factually incorrect. Illustrative examples of this can be found in Figures 9 and 10, where the fine-tuned Mistral models erroneously generate references like '*Kiri Kiri Prison, Auckland, New Zealand*' and '*Bellamy, Jefferson County, West Virginia, United States*'. These inaccuracies

Instruction: Identify the unambiguous reference of Santa Cruz (marked with «START» and «END») in the text.

Input: Santa Cruz is a first-class municipality in the province of Davao del Sur, Philippines. It has a population of 81,093 people as of 2010. The Municipality of «START» **Santa Cruz** «END» is part of Metropolitan Davao. Santa Cruz is politically subdivided into 18 barangays. Of the 18 barangays, 7 are uplands, 9 are upland-lowland and coastal and 2 are lowland-coastal...

Output (by Mistral): Santa Cruz, Davao del Sur, Philippines

Figure 7. An example of correctly estimating unambiguous references by the fine-tuned Mistral model, where higher-level administrative units are mentioned in the text.

Instruction: Identify the unambiguous reference of OleMiss (marked with «START» and «END») in the text.

Input: Only 2 hours until kickoff and the Grove is packed! Let's win the day REBELS! #AreYouReady!!! # «START» **OleMiss** «END» #WinTheDay!

Output (by Mistral): University of Mississippi, Oxford, Lafayette County, Mississippi, United States

Figure 8. An example of correctly estimating unambiguous references by the fine-tuned Mistral model, where higher-level administrative units are not mentioned in the text.

Table 3. *Accuracy@161km* for each dataset (GC denotes GeoCorpora). Numbers in **bold** signify the best scores.

	TR-News	NCEN	GWN	GC	WikToR	WOTR	LDC
Fishing	0.6	0.63	0.56	0.45	0.35	0.54	0.6
DCA	0.64	0.66	0.6	0.62	0.21	0.51	0.56
REL	0.66	0.51	0.65	0.72	0.27	0.54	0.56
BLINK	0.76	0.82	0.75	0.75	0.68	0.74	0.78
Bootleg	0.74	0.68	0.7	0.69	0.7	0.65	0.66
GENRE	0.82	0.82	0.8	0.79	0.88	0.77	0.69
ExtEnD	0.71	0.64	0.68	0.68	0.57	0.64	0.55
LUKE	0.71	0.55	0.74	0.57	0.48	0.42	0.55
ReFinED	0.36	0.48	0.26	0.7	0.74	0.76	0.55
Nominatim	0.68	0.7	0.66	0.74	0.21	0.52	0.24
ArcGIS	0.68	0.71	0.67	0.77	0.24	0.55	0.55
Adaptive	0.66	0.41	0.6	0.54	0.15	0.36	0.29
Population	0.72	0.6	0.62	0.71	0.22	0.43	0.27
CLAVIN	0.71	0.67	0.66	0.77	0.22	0.5	0.41
TopoCluster	0.67	0.72	0.69	0.71	0.24	0.61	0.72
Mordecai	0.68	0.58	0.61	0.66	0.15	0.42	0.36
CBH	0.77	0.57	0.65	0.36	0.43	0.54	0.59
SHS	0.69	0.4	0.57	0.73	0.76	0.43	0.44
CHF	0.77	0.48	0.65	0.75	0.44	0.52	0.56
CamCoder	0.67	0.62	0.6	0.72	0.67	0.47	0.54
Voting	0.86	0.87	0.83	0.84	0.91	0.81	0.8
FT-Falcon (7B)	0.95	0.85	0.87	0.89	0.95	0.75	0.86
FT-Llama2 (7B)	0.96	0.87	0.9	0.9	0.97	0.75	0.88
FT-Llama2 (13B)	0.96	0.88	0.9	0.9	0.98	0.79	0.87
FT-Baichuan2 (7B)	0.94	0.86	0.89	0.89	0.98	0.79	0.88
FT-Mistral (7B)	0.93	0.88	0.89	0.89	0.98	0.81	0.89
FT-Llama2 (70B)	**0.97**	**0.92**	**0.9**	**0.92**	**0.99**	**0.84**	**0.91**

predominantly occur with less prominent locations, such as *'Bellamy'*, a community in the US, which might not be well represented in the model's pre-trained data. To mitigate this effect, deeper integration of LLMs with comprehensive geographic knowledge. For instance, providing the models with a broader range of candidate references for each toponym (e.g., *'Bellamy'* and *'Coatopa'*) during fine-tuning and inference might significantly enhance the accuracy of reference predictions.

4.6. Few-shot prompting vs. fine-tuning

In this section, we examine the effectiveness of few-shot prompting in toponym resolution tasks. We have selected three chat models for this purpose: the original Llama2 (7B), Llama2 (13B), and Mistral (7B). We use a prompt-based approach, asking the

Table 4. *AUC* for each dataset (GC denotes GeoCorpora). Numbers in **bold** signify the best scores.

	TR-News	NCEN	GWN	GC	WikToR	WOTR	LDC
Fishing	0.45	0.36	0.46	0.6	0.55	0.49	0.44
DCA	0.41	0.34	0.45	0.44	0.69	0.52	0.48
REL	0.4	0.5	0.42	0.36	0.62	0.48	0.52
BLINK	0.3	0.17	0.33	0.31	0.25	0.29	0.28
Bootleg	0.33	0.32	0.38	0.39	0.24	0.38	0.39
GENRE	0.25	0.16	0.28	0.28	0.12	0.27	0.35
ExtEnD	0.36	0.36	0.39	0.4	0.38	0.41	0.49
LUKE	0.35	0.45	0.34	0.49	0.49	0.6	0.47
ReFinED	0.61	0.52	0.71	0.37	0.24	0.3	0.49
Nominatim	0.39	0.34	0.4	0.37	0.66	0.47	0.76
Adaptive	0.33	0.63	0.4	0.47	0.84	0.7	0.73
Population	0.26	0.45	0.37	0.31	0.67	0.56	0.72
ArcGIS	0.39	0.31	0.4	0.35	0.64	0.45	0.44
CLAVIN	0.26	0.37	0.33	0.23	0.66	0.49	0.58
TopoCluster	0.37	0.32	0.38	0.37	0.63	0.42	0.32
Mordecai	0.31	0.46	0.38	0.34	0.77	0.57	0.65
CBH	0.21	0.46	0.33	0.65	0.46	0.48	0.44
SHS	0.29	0.6	0.41	0.28	0.25	0.55	0.56
CHF	0.21	0.53	0.33	0.26	0.46	0.49	0.45
CamCoder	0.31	0.43	0.38	0.27	0.3	0.52	0.47
Voting	0.21	**0.14**	0.26	0.23	0.1	**0.25**	0.24
FT-Falcon (7B)	0.07	0.24	0.15	0.15	0.12	0.32	0.26
FT-Llama2 (7B)	**0.06**	0.22	**0.12**	0.13	0.1	0.32	0.21
FT-Llama2 (13B)	0.07	0.2	**0.12**	0.13	0.1	0.3	0.22
FT-Baichuan2 (7B)	0.08	0.22	0.14	0.15	0.1	0.3	0.21
FT-Mistral (7B)	0.09	0.2	0.14	0.15	0.1	0.29	0.2
FT-Llama2 (70B)	**0.06**	0.17	**0.12**	**0.12**	**0.09**	0.28	**0.19**

Table 5. *Mean Error* in kilometers for each dataset (GC denotes GeoCorpora). Numbers in **bold** signify the best scores.

	TR-News	NCEN	GWN	GC	WikToR	WOTR	LDC
Fishing	5424	6590	5371	9729	5104	7016	6987
DCA	3510	5750	5143	5158	7119	6479	7181
REL	3364	7117	4286	3502	5546	4700	7194
BLINK	1655	1776	2243	1577	1217	1040	2323
Bootleg	2943	5825	4494	4425	1483	4501	6050
GENRE	645	894	1088	684	1006	492	4383
ExtEnD	3099	6530	4444	4618	4505	5190	7939
LUKE	3431	8479	2558	6812	7292	9651	5776
ReFinED	5970	10078	8715	3588	3836	2776	8280
Nominatim	1412	3722	1982	1731	3894	2252	13509
Adaptive	5842	11343	7355	8667	15646	11783	13942
Population	1935	5744	4910	3683	4179	5521	10490
ArcGIS	1537	2351	1661	1224	3884	2166	3565
CLAVIN	2862	4622	4705	2777	4171	3929	7262
TopoCluster	1213	1311	1069	1036	3819	1338	1417
Mordecai	3600	6257	5041	4967	9298	6292	10032
CBH	1593	5916	4116	11930	1037	4032	5628
SHS	2375	6828	5237	3154	1078	5219	8133
CHF	1285	6019	3809	2985	1264	3987	5823
CamCoder	3022	5332	5032	3506	501	6067	6101
Voting	403	542	646	460	273	389	928
FT-Falcon (7B)	162	445	409	327	80	353	369
FT-Llama2 (7B)	120	418	293	247	37	303	346
FT-Llama2 (13B)	137	314	**259**	264	36	257	344
FT-Baichuan2 (7B)	295	437	458	333	32	240	311
FT-Mistral (7B)	261	457	371	323	34	251	285
FT-Llama2 (70B)	**61**	**172**	319	**183**	**24**	**180**	277

Instruction: Identify the unambiguous reference of Kiri Kiri prison (marked with ≪START≫ and ≪END≫) in the text.

Input: RT MrAyeDee: Very sketchy info at this point but: A protest was staged by the inmates at ≪START≫ **Kiri Kiri prison** ≪END≫ earlier today and in the process.

Output (by Mistral): Kiri Kiri Prison, Auckland, New Zealand

Figure 9. Fictitious reference generated by the fine-tuned Mistral model, where the correct reference is '*Kiri Kiri Prison, Lagos, Nigeria*'.

Instruction: Identify the unambiguous reference of Bellamy (marked with ≪START≫ and ≪END≫) in the text.

Input: Rescue 14 and engine 4 manned 7 times sent to assist Coatopa fire and ≪START≫ **Bellamy** ≪END≫ fire with a field and truck fire. See smoke from the road.

Output (by Mistral): Bellamy, Jefferson County, West Virginia, United States

Figure 10. Fictitious reference generated by the fine-tuned Mistral model, where the correct reference is '*Bellamy, Alabama, USA*'.

models to determine the unambiguous reference of a toponym. To ensure the models understand our intent, we include three examples of the desired output in the prompt. Figure 11 illustrates the prompt that we have designed for this task. The placeholder '{}' in the prompt will be substituted with the target toponym for each query. Note that, we experimented with numerous prompts, and the final one presented is the most effective according to our testing. After generating the output, we further process it to extract the precise and structured reference, as there is frequently additional explanatory information included in the output, such as *'The unambiguous reference of Jena in the text is (Jena, Germany)'*, a repetition of a toponym in the output such as *'(Germany, Germany)'*, or *'Unknown'* placeholder in the reference. The final reference is then converted to geographical coordinates by querying three geocoders, a process the same as our proposed methods. The overall results of the few-shot model-based approaches and the fine-tuned model-based approaches on the entail test datasets are presented in Table 6.

The results of our study indicate that few-shot model-based strategies do not perform as well as their fine-tuned counterparts. This discrepancy in performance can be attributed to the fact that few-shot models may sometimes struggle to fully comprehend the true intent of the task at hand. For example, the Llama2 (13B) model returns *'Paris, United States'* as the reference for *'Paris'*, despite the fact that there are numerous places named *'Paris'* within the United States, making the reference still ambiguous; The Mistral (7B) model occasionally misidentifies the target toponym, such as returning *'Edmonton, Ky, United States'* for the target *'Ky'*. The Llama2 (7B) model is less reliable, sometimes just repeating the prompt such as outputting *"Boulevard*

Prompt for unambiguous reference estimation:
You are a geographic knowledge expert tasked with estimating the unambiguous reference of the toponym '{}' enclosed within ≪START≫ and ≪END≫ tags in the given text. The reference should be presented in the following format: ({}'s formal name, {}'s parental administrative units). For instance, the parental administrative units of a town should comprise the city, state, and country. Conversely, a country's reference does not require parental administrative units, as it represents the highest level. In cases where specific information for certain parental administrative units is unavailable, please use "Unknown" as a placeholder. Please refrain from including any introductory or explanatory information, and focus solely on providing the reference in the requested format.
Examples:
Input: "I recently visited the ≪START≫ Scottish Highlands ≪END≫ and was amazed by the stunning scenery."
Output: (Scottish Highlands, Scotland, United Kingdom)

Input: "I had a great time exploring ≪START≫ United States ≪END≫ last summer."
Output: (United States)

Input: "I live in ≪START≫ Tx ≪END≫ , US"
Output: (Texas, United States)

Figure 11. Prompt for estimating toponyms' unambiguous reference.

Table 6. Comparison of fine-tuned (FT) and few-shot (FS) prompting models' performance on toponym resolution tasks.

	Accuracy@161km (↑)	AUC (↓)	ME(↓)
FS-Llama2 (13B)	0.88	0.20	276
FT-Llama2 (13B)	0.9	0.18	224
FS-Mistral (7B)	0.86	0.21	429
FT-Mistral (7B)	0.91	0.17	211
FS-Llama2 (7B)	0.82	0.24	617
FT-Llama2 (7B)	0.89	0.19	236

Voltaire's formal name, Boulevard Voltaire's parental administrative units" for the query '*Boulevard Voltaire*'. In contrast, fine-tuning can improve a model's ability to generate the desired output. Our findings highlight the importance and necessity of fine-tuning for effective and accurate toponym disambiguation.

4.7. Impact of geocoders

In this section, we explore the influence of utilizing different geocoders and their combinations to convert estimated unambiguous references into geo-coordinates, using results obtained by the Mistral model as a case study. Eight approaches were employed to geocode references inferred by Mistral: GeoNames (referred to as G), Nominatim (referred to as N), Photon[19] which is an open-source OSM-based geocoder (referred to as P), ArcGIS geocoder (referred to as A), combining GeoNames and Nominatim (referred to as G + N), combining GeoNames and ArcGIS geocoder (referred

Table 7. Comparison of accuracy across different geocoders on Mistral's results.

	G	N	A	P	G + N	G + A	G + P	G + N + A
Accuracy@161km	0.65	0.60	0.89	0.88	0.86	0.90	0.89	0.91
ME	6323	7234	273	294	1601	269	282	211
AUC	0.42	0.46	0.17	0.22	0.22	0.17	0.20	0.17

to as G + A), combining GeoNames and Photon (referred to as G + P), and combining Geonames, Nominatim, and ArcGIS (referred to as G + N + A). The combination methods imply that if one geocoder fails to geocode a reference, the subsequent one will be applied. The results are presented in Table 7.

Both GeoNames and Nominatim geocoders have strict requirements for the input, limited to exact matching, which leads to low accuracy. In contrast, ArcGIS and Photon support fuzzy matching. For example, when provided with the correct reference *'Bantam, Litchfield County, Connecticut, United States'*, Nominatim still fails to geocode it but successfully geocodes *'Bantam, Connecticut, United States'*, where the county information is omitted. Similarly, for *'Dean Woods Road, Metcalfe County, Kentucky, United States'*, where the county *'Metcalfe County'* is incorrect, both Nominatim and Geonames fail to geocode it. However, ArcGIS successfully geocodes these references or returns nearby locations. Conversely, when the input reference is correct and has exact matches with the GeoNames and Nominatim, they tend to produce more accurate results than ArcGIS. For instance, the reference *'Zhejiang, China'*, representing Zhejiang Province in China, is erroneously geocoded by ArcGIS as *'Zhejiang, Baise, Guangxi, China'*, a village in Guangxi Province. Nevertheless, Nominatim and GeoNames accurately geocode this reference. This explains why employing them prior to ArcGIS can further enhance performance.

4.8. Place category

In our research, we examine the resolution abilities of existing methods and our new LLM-based techniques on four kinds of geographic entities: administrative units (like countries, states, and counties), Points of Interest (POIs), including parks, churches, and hospitals, traffic ways (such as streets, highways, and bridges), and natural landmarks (examples being rivers, beaches, and hills). We identified a total of 5,272 administrative units, with examples including *'Germany'*, *'Wuhan city'*, and *'Ferguson'*; 482 POIs such as *'Montreal-Pierre Elliott Trudeau International Airport'*, *'T.J. Health Hospital'*, and *'Saint Peter and Saint Paul Coptic Orthodox Church'*; 1,324 natural features like *'Oulart Hill'*, *'Rich Mountain'*, and *'Eel River'*; and 314 traffic ways, for instance, *'Chapman Highway'*, *'14th street bridge'*, and *'3 Aurangzeb Road'*. We utilized the GeoNames IDs of toponyms in the datasets to determine their categories. To increase the number of fine-grained places, we searched for toponyms without GeoNames IDs, which however, contained keywords such as *'road'*, *'street'*, *'bridge'*, *'highway'*, *'school'*, *'hospital'*, *'airport'*, and *'church'*. These toponyms were then manually verified and categorized.

Next, we calculated the *Accuracy@161km* for each type of location. As shown in Figure 12, many of the tested approaches effectively resolve coarse-grained categories such as administrative units, with eight correctly resolving over 80% of administrative units. However, accurately resolving fine-grained locations like POIs, natural features,

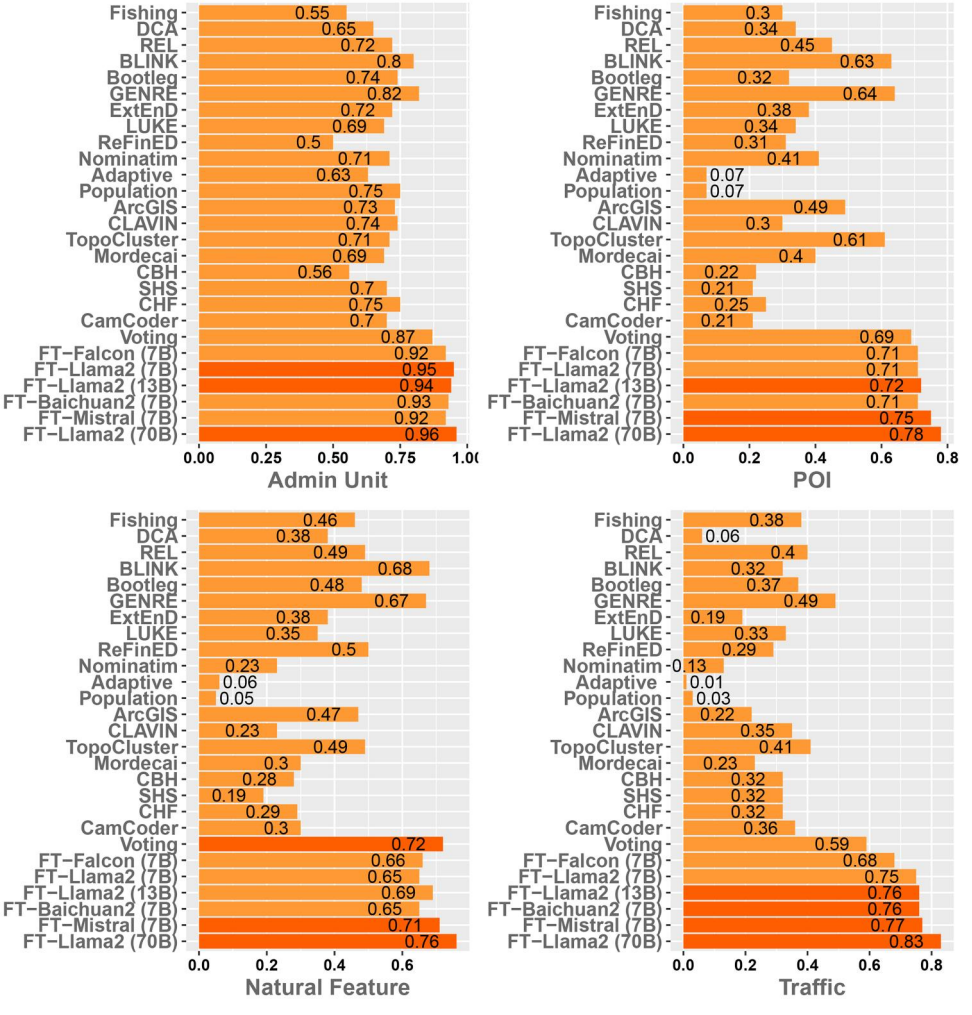
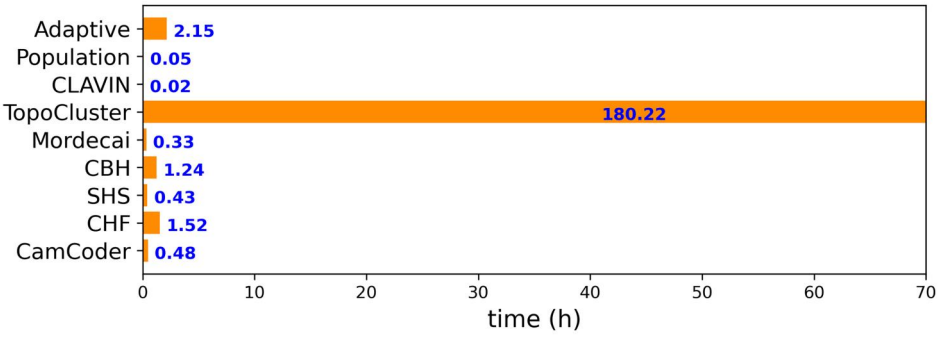


Figure 12. *Accuracy@161km* on four place types with 5,272 admin units, 482 POIs, 1,324 natural features, and 314 traffic ways. The top 3 scores are highlighted.

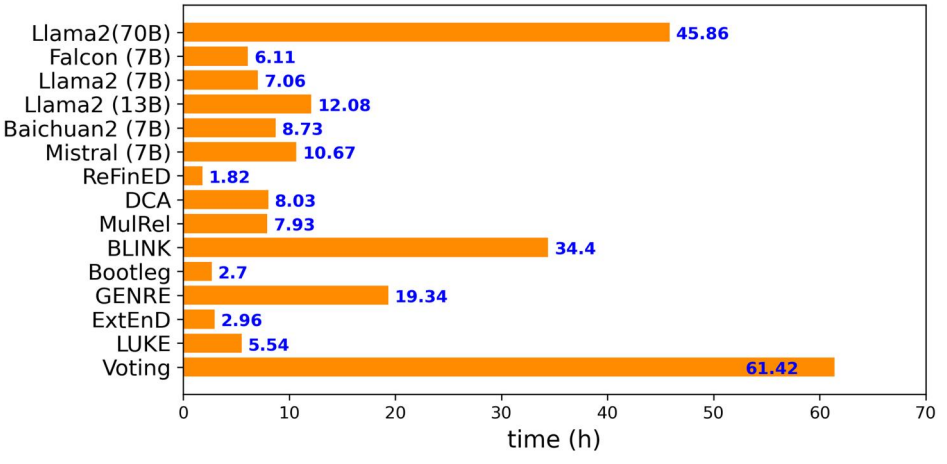
and traffic ways is more challenging, with only six, three, and five methods, respectively, achieving more than 70% accuracy. One contributing factor is incomplete geographic knowledge, particularly regarding fine-grained places within LLMs. The findings show that excluding the Llama2 (70B) based approach, our fine-tuned models perform best for administrative units, POIs, and traffic ways, while the voting method is most effective for natural features. Nevertheless, there is still room for improvement in accurately resolving these fine-grained places.

4.9. Computational efficiency

While accuracy is crucial for geoparsing methods, computational speed is equally important, particularly for applications prioritizing quick processing over higher accuracy. For instance, web search engines require the rapid geocoding of vast



(a) Time consumption for the approaches running on the Intel Core i7-8650U CPU



(b) Time consumption for the approaches running on GPU. Apart from Llama2 (70B), running on four Tesla A100 GPUs, the others run on one Tesla V100 GPU.

Figure 13. Time consumption for each approach running on the complete test datasets.

volumes of documents to support geo-search functionalities. Consequently, we have comprehensively analyzed the computational efficiency of the studied approaches.

This investigation involved running each approach on the complete datasets, focusing on measuring the time consumed during operation, excluding the training phase. The recorded times, as depicted in Figure 13, provide insights into the efficiency of each method. The traditional toponym resolution approaches were tested on a Dell laptop equipped with an Intel Core i7-8650U CPU (1.90 GHz 8-Core) and 16 GB RAM. The fine-tuned models and the deep learning-based ELs, which usually require a GPU execution environment, were run on an NVIDIA Tesla V100 GPU. Llama2 (70B) was run on four Tesla A100 GPUs. It is important to note that for our proposed approaches, we measured only their inference time—the duration required to estimate the unambiguous reference of a toponym—excluding the time taken for subsequent geocoder queries, which varies based on the chosen geocoder and its deployment (local or remote).

Our analysis reveals a notable variance in time efficiency among different toponym resolution approaches. TopoCluster requires approximately 180 hours for completion. In stark contrast, CLAVIN stands out for its speed, accomplishing the task in just about 1 minute. The time consumption for a voting ensemble is cumulative, summing up to 61 hours based on the duration of each incorporated approach. Our proposed methods based on lightweight models exhibit enhanced efficiency, particularly with 7B models, which take between 6 and 11 hours, averaging 0.3 to 0.5 seconds per toponym. Our methods are faster than GENRE and BLINK, the most accurate individual approaches when running on identical hardware platforms. In comparison, the Llama2 (70B) model takes nearly 45 hours and requires significantly more computing resources.

5. Discussion

5.1. Popularity and population bias

Popularity and population bias are common issues in existing toponym resolution approaches, as they tend to favor places with larger populations or greater popularity during disambiguation to achieve the highest accuracy. In this section, we will discuss whether the proposed approaches have addressed these biases. In Table 2, Nominatim, Population (GeoNames-based), and ArcGIS are representative methods that select the most popular or the one with the largest population from a list of returned results given a toponym. As Table 3 illustrates, these methods yield good results (over 0.6) on datasets such as TR-News, NCEN, GeoWebNews, and GeoCorpora, which primarily contain popular places, such as Germany (country). However, when processing datasets with a higher degree of ambiguity, such as WikToR, WOTR, and LDC, their accuracy decreases significantly, particularly on WikToR, with an accuracy below 0.25. This is due to the presence of numerous ambiguous toponyms and unpopular places in these datasets. For instance, in WikToR, each toponym refers to multiple distinct places across the world, such as Santa Maria, California, US; Santa Maria, Bulacan, Philippines; Santa Maria, Ilocos Sur, Philippines; and Santa Maria, Romblon, Philippines for 'Santa Maria', or Paris, France; Paris, Tx, US; Paris, Wisconsin, US; and Paris, Ontario, Canada for 'Paris'. This demonstrates that the three geocoders exhibit a bias towards popularity and population, which explains their low performance on the three ambiguous and challenging datasets. In contrast, the proposed Mistral-based approach achieves scores of 0.98, 0.81, and 0.89 on the three datasets, indicating that our proposed method has effectively addressed the popularity and population bias.

To further conduct a quantitative analysis assessing whether our approach exhibits population bias, we select toponyms from our test datasets whose GeoNames IDs have been annotated. Utilizing the GeoNames ID, we can retrieve the population data for these toponyms. We then categorize these toponyms into 11 groups according to their population size, ranging from 1 to 5,000, 5,000 to 10,000, and so forth. We utilize the Mistral-based approach's estimation and calculate the proportion of toponyms within each group that exhibit a distance error of less than 161 km. Figure 14 depicts the relationship between the population group and the *Accuracy@161km*, where the

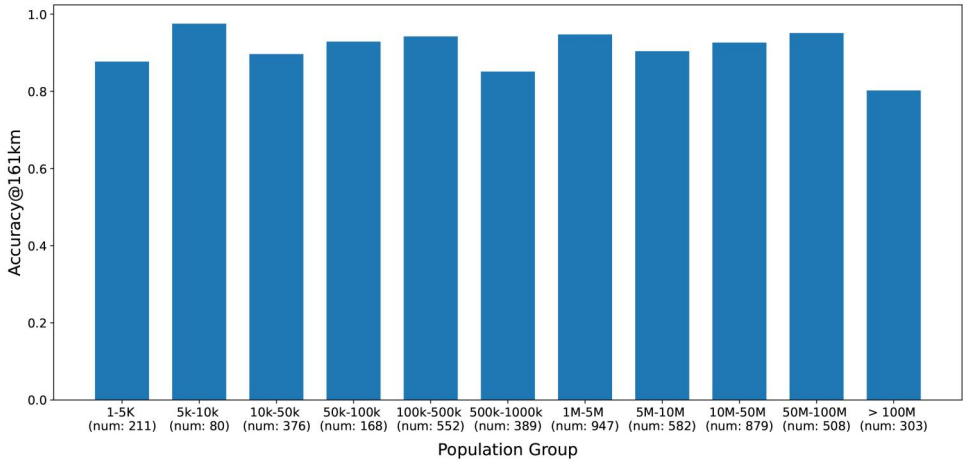


Figure 14. Relationship between population size and *Accuracy@161km* for the Mistral-based approach, with the number of toponyms in each group indicated by the variable 'num'.

variable 'num' under each bar represents the number of toponyms in the corresponding group. The *Accuracy@161km* for each group remains consistent, with all groups achieving a score above 0.8. For example, both the group with a population under 5,000 and the group with a population between 5 and 10 million record an *Accuracy@161km* of approximately 0.85. These findings suggest that our proposed approach does not exhibit a bias towards populations.

5.2. Geographic bias

Many toponym resolution approaches exhibit geographic bias, resulting in inconsistent performance across different regions, with some approaches favoring certain regions over others (Liu *et al*. 2022). We believe that our proposed approaches have addressed this issue. As shown in Figure 5, each dataset's toponyms have different geographic distributions across the globe. However, our proposed Mistral-based approach achieves high Accuracy@161km scores, with all scores above 0.88, as shown in Table 3, except for the WOTR dataset, which focuses on historical toponyms in the US. Furthermore, the toponyms in the WikToR dataset are almost equally distributed globally, and the Mistral-based approach achieves an Accuracy@161km of 0.98 on this dataset. These results suggest that our approaches do not exhibit geographic bias and perform consistently well across different regions.

6. Conclusion

This study presents a novel approach to toponym resolution that combines lightweight, open-source LLMs (e.g., Mistral, Baichuan2, Llama2, and Falcon) and geo-knowledge. The efficacy of our approach is validated through extensive testing on the 7 public datasets, which encompass four distinct types of text. The results clearly demonstrate the superiority of our proposed method, elevating the performance of toponym resolution to a new benchmark. Moreover, the fine-tuned models showcase remarkable computational efficiency, maintaining manageable GPU memory usage: 14 GB for the 7B models and

outperforming the existing state-of-the-art approaches—including a voting system and two deep learning-based entity linkers (namely, GENRE and BLINK)—in terms of speed. On average, the 7B models can infer the unambiguous reference of a toponym in 0.3 to 0.5 seconds, satisfying the requirements of many applications.

Looking forward, our upcoming research will concentrate on diminishing the size of these models to boost processing efficiency, possibly by employing knowledge distillation techniques (West *et al*. 2021). Furthermore, we intend to delve into the deeper fusion of geo-knowledge with LLMs to enhance the accuracy. We also plan to extend the models' capabilities to accommodate multilingual contexts, not limited to English, thereby widening their applicability in various international settings.

Notes

1. https://www.openstreetmap.org/.
2. https://nominatim.org/.
3. https://www.geonames.org/.
4. https://developers.arcgis.com/documentation/mapping-apis-and-services/geocoding/.
5. https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/lgl.xml.
6. https://github.com/tatsu-lab/stanford/_alpaca.
7. https://github.com/baichuan-inc/Baichuan2/tree/main/fine-tune/data.
8. https://geocoder.readthedocs.io/providers/ArcGIS.html/#geocoding. Using this API, we can access the ArcGIS geocoder without a key or token.
9. https://developers.google.com/maps/documentation/geocoding/overview.
10. https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/TR-News.xml.
11. https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/tree/master/data.
12. https://github.com/geovista/GeoCorpora.
13. https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/WikToR.xml.
14. https://github.com/barbarainacioc/toponym-resolution/tree/master/corpora/WOTR.
15. https://github.com/UCREL/LakeDistrictCorpus.
16. https://bl.iro.bl.uk/concern/datasets/f3686eb9-4227-45cb-9acb-0453d35e6a03.
17. http://microposts2016.seas.upenn.edu/challenge.html.
18. https://docs.scipy.org/doc/numpy/reference/generated/numpy.trapz.html.
19. https://github.com/komoot/photon.
20. https://github.com/kermitt2/entity-fishing.
21. https://github.com/Novetta/CLAVIN.

Author contribution

Xuke Hu: Writing – original draft, Software, Methodology, Data curation, Conceptualization; Jens Kersten: Writing – review & editing; Friederike Klan: Writing – review & editing; Sheikh Mastura Farzana: Software.

Disclosure statement

No potential conflict of interest was reported by the author(s). The authors employed ChatGPT and Mistral Chat to polish the language. Following this, the manuscript underwent a thorough review and necessary modifications by the authors, who assume complete responsibility for the final content.

Funding

This work was supported by OpenSearch@DLR.

Notes on contributors

Xuke Hu is a permanent researcher at the DLR's Institute of Data Science. He earned his PhD in Geoinformation from Heidelberg University in 2020. His primary research interests include GeoAI, VGI, indoor localization and mapping, with a recent focus on the extraction and analysis of geographic information embedded in big text data, such as news articles, social media data, and historical documents.

Jens Kersten has a background in geodesy, remote sensing and computer vision. At DLR's Institute of Data Science, he leads a group focusing on multimodal and geospatial information retrieval. His research interests focus on acquiring, analyzing and linking textual data from heterogeneous sources to obtain application-specific information from big data for monitoring and decision making.

Friederike Klan is heading the Data Acquisition and Mobilization Department at the DLR Institute of Data Science. She has a scientific background in computer science with a specialization on data acquisition, preparation, management and provision. The focus of her work is on the development of innovative methods for collecting data, ensuring its quality, making it usable and deriving information from it - from intelligent data acquisition with mobile applications to the development of effective approaches for sharing data in data ecosystems.

Sheikh Mastura Farzana is a researcher at the German Aerospace Center, focusing on Geographic Information Retrieval and associated technologies. Her research interests encompass a range of topics including Geoparsing Web Data, Scalable Geographic Information Retrieval, and Geoparsing Multilingual Data. She holds a Master's degree in Computer Science from the University of Bonn, Germany, and a Bachelor's degree in Computer Science and Engineering from BRAC University, Bangladesh.

ORCID

Xuke Hu http://orcid.org/0000-0002-5649-0243

Data and code availability

The code and data supporting this study's findings are available on GitHub with the link https://github.com/uhuohuy/LLM-geocoding.

References

Ardanuy, M.C., *et al.*, 2022. A dataset for toponym resolution in nineteenth-century English newspapers. *Journal of Open Humanities Data*, 8 (1), 1–7.

Auer, S., *et al.*, 2007. DBpedia: A nucleus for a web of open data. *In*: K. Aberer, *et al.*, eds. *The semantic web*. ISWC ASWC 2007 2007. Lecture Notes in Computer Science, vol. 4825. Berlin, Heidelberg: Springer, 722–735. https://doi.org/10.1007/978-3-540-76298-0_52.

Ayoola, T., *et al.*, 2022. Refined: An efficient zero-shot-capable approach to end-to-end entity linking. *In*: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, Seattle, United States, 209–220.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305* https://arxiv.org/abs/2309.10305, 1–28.

Barba, E., Procopio, L., and Navigli, R., 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.

Carmel, D., *et al.*, 2014. ERD'14: entity recognition and disambiguation challenge. *ACM Sigir Forum*, 48, 63–77.

Chang, Y., *et al.*, 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15 (3), 1–45.

De Cao, N., *et al.*, 2021. Autoregressive entity retrieval. *In*: *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*, Virtual Event, Austria, 1–20, https://openreview.net/forum?id=5k8F6UU39V.

DeLozier, G., *et al.*, 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)*, Berlin, Germany, 188–198.

DeLozier, G., Baldridge, J., and London, L., 2015. Gazetteer-independent toponym resolution using geographic word profiles. *In*: *Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Texas, 2382–2388.

Devlin, J., *et al.*, 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In*: *Proceedings of NAACL-HLT 2019*, Minneapolis, Minnesota, 4171–4186.

Gregory, I., *et al.*, 2015. Geoparsing, GIS, and textual analysis: current developments in spatial humanities research. *International Journal of Humanities and Arts Computing*, 9 (1), 1–14.

Gritta, M., *et al.*, 2018. What's missing in geographical parsing? *Language Resources and Evaluation*, 52 (2), 603–623.

Gritta, M., Pilehvar, M., and Collier, N., 2018. 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, 1285–1296.

Gritta, M., Taher Pilehvar, M., and Collier, N., 2020. A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics. *Language Resources and Evaluation*, 54 (3), 683–712.

Grover, C., *et al.*, 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 368 (1925), 3875–3889.

Guo, Z., and Barbosa, D., 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9 (4), 459–479.

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

Hochmair, H.H., Juhasz, L., and Kemp, T., 2024. Correctness comparison of ChatGPT-4, bard, claude-2, and copilot for spatial tasks. Transactions in GIS, 1–13. https://doi.org/10.1111/tgis.13233.

Hoffart, J., *et al.*, 2011. Robust disambiguation of named entities in text. *In*: *Proceedings of the 2011 conference on empirical methods in natural language processing*, Edinburgh, Scotland, UK, 782–792.

Hu, E.J., *et al.*, 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv: 2106.09685*, 1–26.

Hu, X., *et al.*, 2022a. GazPNE: Annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and synthetic data by rules. *International Journal of Geographical Information Science*, 36 (2), 310–337.

Hu, X., *et al.*, 2022b. GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models. *IEEE Internet of Things Journal*, 9 (17), 16259–16271.

Hu, X., *et al.*, 2023a. How can voting mechanisms improve the robustness and generalizability of toponym disambiguation? *International Journal of Applied Earth Observation and Geoinformation*, 117, 103191.

Hu, X., *et al.*, 2023b. Location reference recognition from texts: A survey and comparison. *ACM Computing Surveys*, 56 (5), 1–37.

Hu, Y., *et al.*, 2023c. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37 (11), 2289–2318.

Hu, Y., and Adams, B., 2021. Harvesting big geospatial data from natural language texts. *In*: M. Werner and Y.Y. Chiang, eds. *Handbook of big geospatial data*. Cham: Springer, 487–507. https://doi.org/10.1007/978-3-030-55462-0_19.

Ji, Y., and Gao, S., 2023. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations. *In:* Proceedings *of the 12th International Conference on Geographic Information Science (GIScience 2023),* Leeds, *U*K, 1–6.

Jiang, A.Q., *et al.*, 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 1–9.

Juhász, L., *et al.*, 2023. ChatGPT as a mapping assistant: A novel method to enrich maps with generative AI and content derived from street-level photographs. *arXiv preprint arXiv: 2306.03204*, 1–12.

Kamalloo, E., and Rafiei, D., 2018. A coherent unsupervised model for toponym resolution. *In*: *Proceedings of the 2018 World Wide Web Conference*, Lyon, France, 1287–1296.

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

Le, P., and Titov, I., 2018. Improving entity linking by modeling latent relations between mentions. *In*: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 1595–1604.

Lermen, S., Rogers-Smith, C., and Ladish, J., 2023. LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B. *In*: *Proceedings of ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, Vienna, Austria, 1–11.

Lewis, M., *et al.*, 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *In*: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 7871–7880.

Li, Z., *et al.*, 2023a. GeoLM: Empowering language models for geospatially grounded language understanding. *arXiv preprint arXiv:2310.14478*.

Li, Z., *et al.*, 2023b. Label supervised llama finetuning. *In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 5227–5240.

Liu, Z., *et al.*, 2022. Geoparsing: Solved or biased? An evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3, 1–13.

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

Lieberman, M.D., Samet, H., and 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)*, 201–212. IEEE.

Mai, G., *et al.*, 2022. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–4.

Mai, G., *et al.*, 2024. On the opportunities and challenges of foundation models for geospatial artificial intelligence (Vision Paper). *ACM Transactions on Spatial Algorithms and Systems*, 10 (2), 1–46.

Milusheva, S., *et al.*, 2021. Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning. *PloS One*, 16 (2), e0244317.

Min, B., *et al.*, 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56 (2), 1–40.

Mooney, P., *et al.*, 2023. Towards understanding the geospatial skills of chatgpt: Taking a geographic information systems (gis) exam. *In*: *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 85–94.

Nguyen, T.T., Wilson, C., and Dalins, J. 2023. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*, 1–8.

Onoe, Y., and Durrett, G., 2020. Fine-grained entity typing for domain independent entity linking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 8576–8583.

Orr, L., *et al.*, 2020. Bootleg: Chasing the tail with self-supervised named entity disambiguation. *arXiv preprint arXiv:2010.10363*, 1–25.

Penedo, G., *et al.*, 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *In*: *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 79155–79172.

Purves, R.S., *et al.*, 2018. Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval*, 12 (2–3), 164–318.

Rayson, P., *et al.*, 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.

Scott, P., *et al.*, 2019. Global biogeography and invasion risk of the plant pathogen genus Phytophthora. *Environmental Science & Policy*, 101, 175–182.

Speriosu, M., and Baldridge, 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)*, 1466–1476.

Tao, R., and Xu, J., 2023. Mapping with chatgpt. *ISPRS International Journal of Geo-Information*, 12 (7), 284.

Touvron, H., *et al.*, 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 1–77.

Vrandečić, D., and Krötzsch, M., 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57 (10), 78–85.

Wallgrün, J.O., *et al.*, 2018. GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32 (1), 1–29.

Wang, J., and Hu, Y., 2019. Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23 (6), 1393–1419.

Weissenbacher, D., *et al.*, 2019. Semeval-2019 task 12: Toponym resolution in scientific papers. *In*: *Proceedings of the 13th International Workshop on Semantic Evaluation*, 907–916.

West, P., *et al.*, 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

Wikimedia Foundation Inc Encyclopedia online. *Wikipedia: The free encyclopedia* [online]. Available from: http://en.wikipedia.org/wiki/Wikipedia [Accessed 30 January 2024].

Wu, L., *et al.*, 2020a. Scalable zero-shot entity linking with dense entity retrieval. *In*: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6397–6407.

Wu, L., *et al.*, 2020b. Zero-shot entity linking with dense entity retrieval. In *EMNLP*.

Xie, Y., *et al.*, 2023. Geo-foundation models: Reality, gaps and opportunities. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, 1–4.

Yamada, I., *et al.*, 2022. Global entity disambiguation with BERT. *In*: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Association for Computational Linguistics, 3264–3271.

Yang, X., *et al.*, 2019. Learning dynamic context augmentation for global entity linking. *In*: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 271–281.

Yin, Z., Li, D., and Goldberg, D.W., 2023. Is ChatGPT a game changer for geocoding-a benchmark for geocoding address parsing techniques. *In*: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data*, 1–8.

Zhang, Y., *et al.*, 2021. Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data. *Journal of Hydrology*, 603, 127053.