

UniTopRank: a scalable and language-independent method for toponym resolution

Xuke Hu, Yao Sun , Tobias Hecking , Jens Kersten & Friederike Klan

To cite this article: Xuke Hu, Yao Sun , Tobias Hecking , Jens Kersten & Friederike Klan (25 Mar 2026): *UniTopRank*: a scalable and language-independent method for toponym resolution, International Journal of Geographical Information Science, DOI: [10.1080/13658816.2026.2645831](https://doi.org/10.1080/13658816.2026.2645831)

To link to this article: <https://doi.org/10.1080/13658816.2026.2645831>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 25 Mar 2026.



[Submit your article to this journal](#)



Article views: 1139



[View related articles](#)



[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

RESEARCH ARTICLE



UniTopRank: a scalable and language-independent method for toponym resolution

Xuke Hu^a , Yao Sun^b, Tobias Hecking^c, Jens Kersten^a and Friederike Klan^a

^aInstitute of Data Science, German Aerospace Center (DLR), Jena, Germany; ^bRemote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany; ^cInstitute for Software Technology, German Aerospace Center (DLR), Cologne, Germany

ABSTRACT

Toponym resolution is essential for extracting geographic information from unstructured texts. Although recent deep learning and large language model (LLM)-based methods achieve strong performance, they typically require substantial computational resources and have slow inference speeds, limiting their practicality for resource-constrained, large-scale or time-sensitive applications. Moreover, most existing approaches are developed for English or a limited set of languages, restricting their cross-lingual applicability. To fill the gaps, we propose *UniTopRank*, a universal, scalable, and language-independent method based on rule-driven ranking. It operates in two steps: first, candidate locations for all toponyms in a text are retrieved from gazetteers; second, the candidates are ranked using name similarity, population, administrative level, and spatial relationships among toponyms through beam search. We evaluate *UniTopRank* on 25 datasets in 14 languages and 7 text types, covering 186,100 toponyms worldwide. For 11 English datasets, we compare it with 21 existing methods, including deep learning- and LLM-based models, and for non-English datasets, with 12 representative approaches. Results show that *UniTopRank* attains accuracy comparable to state-of-the-art methods—though not the best—and generalizes well across different languages, while being significantly faster and much more resource-efficient. Running entirely on standard CPUs, it is well-suited for large-scale, multilingual, or resource-constrained applications.

ARTICLE HISTORY

Received 8 November 2025
Accepted 11 March 2026

KEYWORDS

Geoparsing; geocoding; toponym ranking; toponym resolution; language-independent

1. Introduction

Geospatial information describes entities in relation to geographic space and is typically collected through ground surveying, mobile mapping, laser scanning, remote sensing, GNSS, and Volunteered Geographic Information (Li *et al.* 2016). Yet a huge

CONTACT Xuke Hu  xuke.hu@dlr.de

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/13658816.2026.2645831>.

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

amount of geospatial information is still buried in vast and unstructured natural language text. Each day, 500 million Twitter/X posts,¹ 100 million Weibo posts,² and 2 million news articles³ are generated, while about 252,000 new websites⁴ are created, supplemented by a continual stream of scientific papers, organizational reports, and newly digitized historical documents. Buried within these materials are countless toponyms, location expressions, and spatial details about people, events, environments, and societies—information that existing physical sensors or cartographic data often miss. These range from real-time resource and rescue requests during disasters (Zou *et al.* 2023) to landscape descriptions written centuries before the first aerial photograph (Gregory and Ell 2007). This information provides a new spatiotemporal view of our world (Hu and Adams 2021) and supports various applications, such as geographic search (Hendriksen *et al.* 2024), emergency response (Zheng, Juanle, and Jie 2025), spatial humanities (McDonough, Moncla, and Van de Camp 2019), hydrology analysis (Stein *et al.* 2024), urban analytics (Reades *et al.* 2025), biodiversity informatics (Nowak, Słupecka, and Jackowiak 2021), and disease surveillance (Allen *et al.* 2017).

However, unlike other data sources, geospatial information in free text is implicit. Extracting it requires advanced NLP methods, often named geoparsing. Geoparsing has two main steps: toponym recognition, which recognizes toponym mentions from free text, and toponym resolution, which links each mention to a real place or a gazetteer entry. While toponym recognition has advanced considerably (Akbik *et al.* 2019; Hu *et al.* 2022a, 2022b, 2023b), toponym resolution remains challenging due to (1) diverse **toponym variants**—ranging from misspellings, alternative or colloquial names, abbreviations, and word omissions to historical or temporally shifted variants and multilingual or cross-linguistic variants, and (2) severe **geo/geo-ambiguity**. For example, querying “Heidelberg” in Nominatim returns more than 300 distinct results, such as *Heidelberg, Baden-Württemberg, Germany*; *Heidelberg, Wellesley, Ontario, Canada*; and *Heidelberg Drive, Baldwin County, Alabama, United States*. These challenges are further magnified in under-represented languages and regions where geographic data is sparse.

Past research has explored rule-based and machine learning-based approaches for toponym resolution (Leidner and Lieberman 2011; De Bruijn *et al.* 2018; Kamalloo and Rafiei 2018; Karimzadeh *et al.* 2019). In recent years, deep neural models and large language models (LLMs) have achieved remarkable accuracy (Gomes, Purves, and Volpi 2024; Hu *et al.* 2026). However, these models typically require GPUs, large memory resources, or exhibit high latency, which limits their practicality as lightweight components within larger processing pipelines. In many real-world scenarios, efficiency, ease of integration, and high throughput are often more critical than achieving peak accuracy (Belliardo, Kalimeri, and Mejova 2023; Stein *et al.* 2024; Syed *et al.* 2025). For instance, in the EU Horizon project *OpenWebSearch.eu*,⁵ around 2 TB of web data are collected daily, which must be geolocated to support location-based web search, while only CPU resources are available (Farzana and Hecking 2024). Consequently, heuristic and rule-based resolution methods remain valuable, as they are much faster, more affordable, and easier to integrate for large-scale deployment, despite the existence of more accurate deep learning alternatives.

Geoparsing is often applied in large-scale multilingual environments, such as the European Commission’s Europe Media Monitor (EMM),⁶ which continuously analyses over 220,000 news articles daily across more than 20 languages. Nevertheless, most existing toponym resolution methods have been developed for a single language, predominantly English. Although several studies have extended their approaches to other languages, such as German (Ardanuy and Sporleder 2017), Italian (Avvenuti *et al.* 2018), Finnish (Leppämäki, Toivonen, and Hiippala 2024), Chinese (Chen *et al.* 2026), and Spanish (Aldana-Bobadilla *et al.* 2020), these efforts remain largely language-specific or have been evaluated in limited linguistic settings. Moncla *et al.* (2014) showed that the geocoding step can, in principle, operate in a language-independent manner through spatial reasoning and gazetteer-based inference, and evaluated this approach on three Romance languages: French, Spanish, and Italian. To date, no approach has been systematically validated or proven effective across a truly diverse set of languages.

These challenges motivate the following key question of this study:

RQ: *How can we design a toponym resolution method that attains multilingual accuracy comparable to state-of-the-art methods, while remaining computationally efficient on standard CPUs?*

To address this research question, we propose a rule-based toponym ranking strategy designed for scalability and efficiency. Although rule-based approaches may seem less attractive in the current AI-dominated era, they remain useful when transparency, high speed, and low computational cost are essential. Given all detected toponyms in a text, our method first queries gazetteers such as GeoNames to retrieve candidate locations for each toponym. The candidates are then ranked based on their intrinsic attributes and spatial relationships. Specifically, the ranking integrates four complementary factors: (i) name similarity, (ii) population, (iii) administrative level, and (iv) spatial coherence across toponyms within the text. Because all features are derived from gazetteer metadata rather than language-specific textual context, the method is inherently language-independent. Like many previous studies, we also questioned the generalizability of rule-based methods—a well-known limitation of such methods. However, our experiments indicate that the proposed method generalizes well in practice. Specifically, we evaluated our approach against numerous existing methods using 25 datasets in 14 languages—covering English, German, Italian, Finnish, Japanese, Spanish, Romanian, Ukrainian, Turkish, Portuguese, Farsi, Serbian, Polish, Tamil, and Arabic—comprising a total of 186,100 toponyms in 7 text types.

Our main contributions are:

- We propose a scalable and language-independent toponym resolution method that attains accuracy comparable to state-of-the-art methods while operating efficiently on standard CPU hardware.
- We conduct the most comprehensive multilingual evaluation of toponym resolution to date, comparing 21 methods on 25 datasets across 14 languages.

The remainder of the paper is divided into several parts. [Section 2](#) reviews prior studies relevant to our work. [Section 3](#) outlines the proposed approach, while [Section](#)

4 explains the experimental setup and obtained results. Section 5 discusses key observations, and Section 6 closes with conclusions and potential future directions.

2. Related works

In this section, we first examine projects and applications that employ rule-based toponym resolution methods. We then review existing resolution approaches, from early rule-based to the latest LLM-based. Since Hu *et al.* (2023a, 2026); Zhang and Bethard (2025) have already provided comprehensive overviews of toponym resolution techniques, including entity linking methods, our focus here is primarily on earlier approaches not covered in their surveys and more recent developments.

2.1. Practice in real systems: why simple methods persist

Despite the availability of advanced resolution approaches, especially those based on deep learning and LLMs, many real applications continue to use simple rule or heuristic-based methods because they are computationally efficient, fast, and easy to integrate into existing systems. The first example is the aforementioned EU *OpenWebSearch* project, which requires daily geoparsing of large-scale web data across 185 languages (Farzana and Hecking 2024). Another example is PADI-web,⁷ an epidemiological surveillance platform that automatically collects and processes news to extract diseases, locations, and dates. It resolves each detected toponym by querying a gazetteer and selecting the first returned entry (Arsevska *et al.* 2018; Syed *et al.* 2025). A further example is AMUNATCOLL⁸ (Access and Use of Natural History Collections Data), which digitizes natural history collections from Polish museums and institutions; Nowak, Słupecka, and Jackowiak (2021) geo-tag specimen-label locality descriptions using simple heuristic disambiguation.

Beyond real-world applications and projects, many studies also adopt simple rule-based resolution. For example, Stein *et al.* (2024) examine global biases in hydro-hazard research by analyzing where studies are conducted and which regions are over- or underrepresented. They geolocated nearly 300,000 abstracts using a hybrid rule- and gazetteer-based workflow; when ambiguous, they retrieved the top five candidates from Nominatim and ranked them by the OSM importance score. Similarly, Acheson and Purves (2021) built an automated system to map the geographic focus of scientific articles in ecology and cancer genetics. Their pipeline extracts location mentions using an off-the-shelf NER with simple rule-based filtering, then resolves each mention via the Google Geocoding API, retaining the top-ranked result. Many other works likewise rely on lightweight resolution to address domain-specific needs in disaster management, humanitarian reporting, and public health (Belliardo, Kalimeri, and Mejova 2023; Curtis *et al.* 2025; Vračević *et al.* 2025).

2.2. Classification of toponym resolution

Hu *et al.* (2023a); Zhang and Bethard (2025) broadly categorize toponym resolution methods into three groups: rule-based, learning-and-ranking, and learning-and-classification approaches.

Rule-based methods score gazetteer candidates using predefined features. Ardanuy and Sporleder (2017) proposed a weighted model over contextual similarity, geographic context, and inter-toponym spatial proximity, with weights tuned on a small training set. TAGGS (De Bruijn *et al.* 2018) ranks tweet toponyms using heuristics (UTC offset, home location, geotags, co-occurring place names); scores are averaged across tweets sharing the toponym, and ties go to the largest-population candidate. CIME (Scalia, Francalanci, and Pernici 2022) first ranks by single cues (e.g., other mentioned toponyms, profile location); if unresolved, it adds context from related tweets (same event/period/conversation) to rescore candidates and improve accuracy.

Learning-and-ranking methods learn ranking rules from annotated data. Mordecai 3 (Halterman 2023) uses transformer NER, retrieves GeoNames candidates, and ranks them with a neural model combining string similarity, contextual embeddings, and hierarchical cues. SemTopRes (Gomes, Purves, and Volpi 2024) generates GeoNames candidates and ranks them with a fine-tuned Siamese transformer via cosine similarity between the toponym context embedding and the candidate text embedding. GeoPLACE (Zhang, Laparra, and Bethard 2024) predicts toponym properties (e.g., country, state, feature type) with a transformer, then uses them as constraints to filter and rank candidates.

Learning-and-classification frames resolution as supervised classification over a global grid (Gritta, Pilehvar, and Collier 2018), using features from the target toponym, co-occurring toponyms, and surrounding context. Kulkarni *et al.* (2020) discretized Earth with multi-level S2 cells (level 4, ~300 km, to level 8, ~1 km) and trained a CNN on 1.76 million Wikipedia entries to predict likely cells. Yan *et al.* (2021) proposed *LGGeoCoder*, fusing local (textual/toponymic) and global (topic/spatial embedding) signals to improve classification.

With rapid LLM progress, a new paradigm has emerged. Hu *et al.* (2026) fine-tuned open-source 7B/13B models (e.g., Mistral, LLaMA2, Falcon, Baichuan2) to generate unambiguous references (formal name + higher-level units; e.g., “Paris, Lamar County, TX, US”), then geocoded them. Chen *et al.* (2026) used GPT-4 prompting to identify ancient Chinese toponyms and normalize them into unambiguous references before geocoding. RACCOON (Lin *et al.* 2025) instead prompts an LLM (GPT or Gemini) with gazetteer candidates, which selects the most likely candidate and outputs its coordinates.

This paper introduces a rule-based toponym resolution method that emphasizes cost-efficiency, computational speed, and language independence, inspired by prior rule-based methods (Ardanuy and Sporleder 2017; Kamalloo and Rafiei 2018; Karimzadeh *et al.* 2019).

3. Proposed approach

Our approach, illustrated in Figure 1, consists of two stages. First, we query gazetteers to retrieve candidate locations for all toponyms in the text. Second, we compute an overall weighted score for each candidate based on four factors and perform beam search to infer the best candidate location.

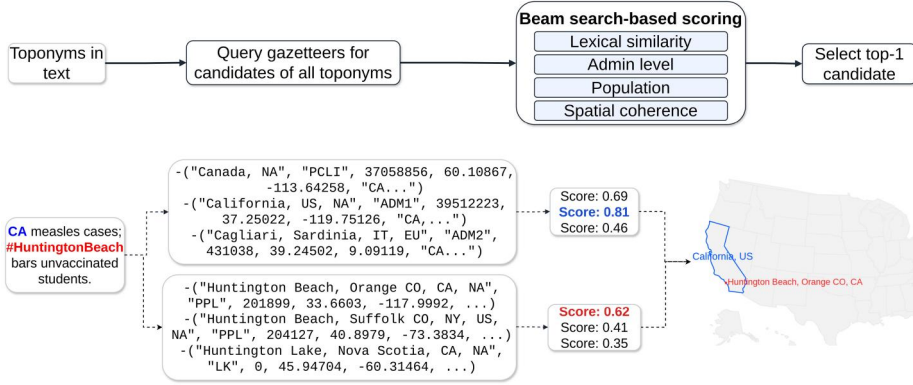


Figure 1. Workflow of the proposed method.

3.1. Stage I: candidate retrieval

Given a document, we assume all toponyms have already been detected. For each toponym, we query one or more gazetteers to obtain a candidate set. For a toponym t , the candidate set $C(t)$ contains entries with canonical names, alternative and multilingual variants, administrative attributes (including higher-level units), population statistics, and coordinates.

We run local instances of a GeoNames server⁹ and Photon (an OSM-based geocoder).¹⁰ We choose Photon over Nominatim for two reasons: (i) the public Nominatim service is rate-limited, and a local planet-scale instance requires >1 TB storage; (ii) the online Nominatim API returns at most 50 candidates and often only a few for common queries (e.g., “London”¹¹), omitting valid alternatives such as London, Kentucky, US. Photon has a much smaller footprint (about 100 GB globally) by simplifying feature representations (e.g., geometry as a single point) and can be configured to return all candidates, which better fits our method.

We enable fuzzy search in GeoNames to improve recall, but it can return hundreds of noisy candidates. To control this, we first collect candidates whose canonical or alternative names (including language variants) exactly match the query. If the number of exact matches is below a threshold P (e.g., $P = 40$), we add non-exact matches ranked by population until P is reached. Since GeoNames has limited coverage of fine-grained locations (e.g., streets, POIs), we additionally query Photon when GeoNames returns insufficient results. However, Photon lacks key metadata (e.g., population and alternative/multilingual names) used in our ranking, so GeoNames remains the primary gazetteer and Photon is only complementary. We refer to the GeoNames-only variant as *UniTopRank-GN*, and the variant that falls back to Photon when GeoNames yields no candidates as *UniTopRank-GN+Photon*.

3.2. Stage II: candidate ranking

Let $T = \{t_1, t_2, \dots, t_{|T|}\}$ denote the set of toponyms extracted from a text. For each toponym $t \in T$, we denote its candidate set obtained in Stage I as $C(t)$.

As an initial step, we check whether the text contains adjacent toponyms separated by a comma (e.g., “Paris, TX”). If such a pair is detected, we query GeoNames to verify whether a corresponding combined entry exists. Whenever a valid match is found, the candidates for both toponyms are fixed immediately, and the resolved pairs serve as anchors for the subsequent ranking steps.

For all remaining toponyms $t \in T$, each candidate $c \in C(t)$ is assigned a weighted score that integrates four complementary signals: lexical similarity, population, administrative level, and spatial coherence. We assume that ambiguous place mentions can be resolved by combining four complementary gazetteer cues. (i) *Lexical similarity* prioritizes candidates whose names or aliases closely match the mention, favoring exact or near-exact matches over loose ones. (ii) *Population* serves as a prominence prior when contextual evidence is limited (e.g., “Paris” is biased toward Paris, France over Paris, Texas, US). (iii) *Administrative level* captures the tendency for higher-level units to be mentioned more frequently (e.g., “Ontario” in news is more often used to refer to the Canadian province than to the city in California, US). (iv) *Spatial coherence* leverages that co-mentioned toponyms are often geographically consistent; for instance, “Portland” (Oregon/Maine) and “Salem” (Oregon/Massachusetts) jointly favor the coherent pair (Portland, Oregon + Salem, Oregon) over distant mixed pairs. We compute coherence using both distance-based and hierarchy-based scores and take the larger, since distance can be large even within a shared administrative context (e.g., two US states), while hierarchy can be low for geographically close cross-border cities. Formal definitions follow.

- i. **Lexical similarity** $s_{\text{name}}(t, c)$ quantifies the resemblance between the toponym mention t and the candidate c , considering both its canonical and alternative names (including multilingual variants):

$$s_{\text{name}}(t, c) = \max\left(0, 1 - \min_{n \in \text{Names}(c)} \frac{d_{\text{Lev}}(t, n)}{\max(|t|, |n|)}\right), \quad (1)$$

where $d_{\text{Lev}}(t, n)$ denotes the Levenshtein distance (Recchia and Louwerse 2013) between strings t and n , and $|t|$ and $|n|$ are their respective lengths.

- ii. **Population** $s_{\text{pop}}(c)$ favors more populous settlements while still assigning non-zero scores to smaller ones:

$$s_{\text{pop}}(c) = \left(\frac{\log(1 + \text{Pop}(c))}{\log(1 + P_{\text{max}})}\right)^{\kappa}, \quad (2)$$

where $\text{Pop}(c)$ is the population of candidate c , P_{max} is a reference maximum population (e.g., a megacity), and κ controls the curvature of the function.

- iii. **Administrative level** $s_{\text{admin}}(c)$ rewards candidates at the appropriate administrative level (e.g., state over city). Each candidate is mapped to a predefined weight based on its feature code (see Table 1). If a candidate’s code is missing or not listed, the weight is set to 0.

Table 1. Administrative-level prior weights associated with GeoNames feature codes.

Weight	Code	Description
1.0	CONT	Continent-level unit (e.g., Europe, Asia, Africa).
1.0	PCL	Generic country-level political unit.
1.0	PCLI	Fully independent sovereign entity.
0.9	PCLS	Semi-autonomous or dependent political unit.
0.8	PPLC	National capital or primary administrative center.
0.8	ADM1	First-level administrative division (e.g., province or state).
0.7	ADM1H	Former first-level administrative area.
0.7	ADM2	Second-level division (e.g., county or district).
0.7	PPLCH	Former national capital.
0.6	ADM2H	Historical second-level administrative division.
0.5	ADM3	Third-level division (below district or county).
0.4	ADM3H	Former third-level administrative division.
0.4	RGN	Region defined by natural or cultural features (non-administrative).
0.4	ADM4	Fourth-level administrative unit.
0.4	ADM5	Fifth-level administrative unit.

$$s_{\text{admin}}(c) = w_{\text{code}}(c), \quad w_{\text{code}}(c) \in [0, 1]. \quad (3)$$

iv. **Spatial coherence.** $s_{\text{spatial}}(c|T)$ quantifies the geographic compatibility of a candidate c with the potential locations of other toponyms in T . Each candidate is represented through its complete administrative hierarchy (e.g., “City of Paris, Lamar County, Texas, United States, North America” for “Paris”). The metric integrates two complementary components:

(i) a *hierarchical score* $s_{\text{hier}}(c, c')$, where $c' \in C(t')$ denotes a candidate for another toponym $t' \in T \setminus \{t\}$. The score equals 1 when c and c' correspond to sibling entities (e.g., “City of Paris, Lamar County, Texas, United States, North America” and “Blossom, Lamar County, Texas, United States, North America”) or share a direct parent–child link (e.g., “Lamar County, Texas, United States, North America” and “Texas, United States, North America”). Otherwise, it depends on the length of their common administrative suffix. Let $\mathbf{a}(c) = (a_0, \dots, a_k)$ and $\mathbf{a}(c') = (a'_0, \dots, a'_m)$ denote the administrative address tuples of c and c' , ordered from fine to coarse levels. We define

$$s_{\text{hier}}(c, c') = \begin{cases} 1, & \text{if } c \text{ and } c' \text{ are sibling entities} \\ & \text{or have a direct parent – child relation,} \\ \frac{\ell(c, c')}{\max(|\mathbf{a}(c)|, |\mathbf{a}(c')|)}, & \text{otherwise.} \end{cases} \quad (4)$$

where $\ell(c, c')$ is the length of the longest common administrative suffix:

$$\ell(c, c') = \max\{\ell \mid a_{k-i} = a'_{m-i}, \forall i \in \{0, \dots, \ell - 1\}\}. \quad (5)$$

For example, let $\mathbf{a}(c)$ be “City of Paris, Lamar County, Texas, United States, North America” and let $\mathbf{a}(c')$ be “Houston, Harris County, Texas, United States, North America”. The longest common suffix is “Texas, United States, North America”, so $\ell(c, c') = 3$ and $s_{\text{hier}}(c, c') = 3/5$.

(ii) a *distance-based score*, computed via inverse-distance weighting with a unit offset:

$$s_{\text{dist}}(c, c') = \frac{1}{1 + d(c, c')/d_{\text{max}}},$$

where $d(c, c')$ refers to the geographic separation between the two candidates, and $d_{\max}$ denotes to the normalization constant ensuring that the value decreases monotonically with distance but remains within $(0, 1]$. For every other toponym $t' \in T \setminus \{t\}$, each candidate c compares with all $c' \in C(t')$ and retains the maximum value between s_{hier} and s_{dist} . These maxima are then averaged across all remaining toponyms:

$$s_{\text{spatial}}(c|T) = \frac{1}{|T| - 1} \sum_{t' \in T \setminus \{t\}} \max_{c' \in C(t')} (s_{\text{hier}}(c, c'), s_{\text{dist}}(c, c')). \quad (6)$$

The final confidence score for a candidate $c \in C(t)$ is computed as a weighted combination of four signals:

$$S(c|t, T) = \lambda_{\text{name}} s_{\text{name}}(t, c) + \lambda_{\text{pop}} s_{\text{pop}}(c) + \lambda_{\text{admin}} s_{\text{admin}}(c) + \lambda_{\text{spatial}} s_{\text{spatial}}(c|T), \quad (7)$$

where $\lambda_{\text{name}}, \lambda_{\text{pop}}, \lambda_{\text{admin}}, \lambda_{\text{spatial}} \in [0, 1]$ control the relative weights of the respective components.

Adaptive spatial weight: Unlike the other weights, λ_{spatial} is not fixed. Instead, it is adapted dynamically according to the number of unique toponyms in T that yield non-empty candidate sets. The intuition is that spatial coherence becomes informative only when multiple toponyms are present, whereas in texts with very few toponyms, the other cues should dominate. Formally, the spatial weight is defined as

$$\lambda_{\text{spatial}} = \lambda_{\text{spatial}}^{\text{base}} \cdot \left(1 + \min \left(2, \frac{N_{\text{cand}}}{\tau} \right) \right), \quad (8)$$

where $\lambda_{\text{spatial}}^{\text{base}}$ is the base weight, N_{cand} denotes the number of unique toponyms in T with valid candidates, and τ is a threshold parameter controlling the growth rate.

Beam search with top- N pruning: Since the ranking of a candidate depends not only on its own attributes but also on its spatial relationships with candidates of other toponyms, we employ a beam search strategy with top- N pruning. After computing scores for all candidates of a toponym, only the N highest-ranked candidates are retained. These reduced candidate sets are then used when computing spatial coherence, thereby narrowing the search space, and ultimately improving efficiency.

Batch-wise efficiency optimization: In long documents, the number of unique toponyms can exceed several hundred (see Table 3). Computing spatial coherence for all pairwise toponym–candidate combinations incurs a time complexity of $O(|T|^2 \cdot \bar{C}^2)$, where $|T|$ denotes the total number of unique toponyms and $\bar{C}$ denotes the average number of candidates per toponym. This quadratic growth quickly becomes computationally prohibitive. To improve efficiency, we adopt a batch-wise processing strategy: the set of unique toponyms is partitioned into groups of size K , and each batch is processed independently through the algorithm. If the final batch contains fewer than K toponyms, it is padded with previously processed ones to maintain a stable spatial

Table 2. Parameter settings used in all experiments.

Parameter	Value / Setting
Lexical weight λ_{name}	0.3
Population weight λ_{pop}	0.2
Administrative weight λ_{admin}	0.1
Spatial weight λ_{spatial}	Adaptive: $\lambda_{\text{spatial}}^{\text{base}} = 0.3$, $\tau = 15$
Curvature κ (pop. prior)	0.35
Reference population P_{max}	20 M (megacity scale)
Distance normalization d_{max}	300 km
Beam search pruning N	5
Batch size K	30

Table 3. Statistics of 25 benchmark datasets for toponym resolution in 14 languages with a total of 186,100 toponyms in 7 text types. μ_T denotes the average number of toponyms per text. Type: Sci = scientific articles, HD = historical documents, WN = Wikinews, Wiki = Wikipedia articles. KB: GN = GeoNames, WKP = Wikipedia, WKD = Wikidata.

Name	Text	Top.	μ_T	Type	Lang	KB
TR-News (Kamalloo and Rafiei 2018)	118	1,319	11.2	News	EN	GN
GeoWebNews (Gritta, Pilehvar, and Collier 2018)	200	2,601	13	News	EN	GN
GeoCorpora (Wallgrün <i>et al.</i> 2018)	6,648	3,100	0.5	Tweet	EN	GN
WOTR (DeLozier <i>et al.</i> 2016)	1,644	11,795	7.2	HD	EN	GN
CLDW (Rayson <i>et al.</i> 2017)	62	34,713	560	HD	EN	GN
TUD (Katz and Schill 2013)	152	3,852	25.3	Web	EN	GN
NCEN (Ardanuy <i>et al.</i> 2022)	455	4,595	10.1	HD	EN	WKP
SemEval (Weissenbacher <i>et al.</i> 2019)	91	3,344	36.7	Sci	EN	GN
LGL (Lieberman, Samet, and Sankaranarayanan 2010)	588	5,088	8.7	News	EN	GN
GeoVirus (Gritta, Pilehvar, and Collier 2018)	229	2,170	9.5	News	EN	GN
WikToR (Gritta <i>et al.</i> 2018)	5,000	25,242	5	Wiki	EN	WKP
ITA-DSTR (Avvenuti <i>et al.</i> 2018)	1,807	958	0.5	Tweet	IT	N/A
FingerTweets (Leppämäki, Toivonen, and Hiippala 2024)	980	497	0.5	Tweet	FI	GN
FingerNews (Leppämäki, Toivonen, and Hiippala 2024)	42	189	4.5	News	FI	GN
Newsli-UK (Botha, Shan, and Gillick 2020)	164	258	1.6	WN	UK	WKD
Newsli-DE (Botha, Shan, and Gillick 2020)	12,954	26,575	2.1	WN	DE	WKD
Newsli-ES (Botha, Shan, and Gillick 2020)	7,372	18,115	2.5	WN	ES	WKD
Newsli-JA (Botha, Shan, and Gillick 2020)	2,533	7,967	3.1	WN	JA	WKD
Newsli-SR (Botha, Shan, and Gillick 2020)	13,780	24,305	1.8	WN	SR	WKD
Newsli-TR (Botha, Shan, and Gillick 2020)	737	2,424	3.3	WN	TR	WKD
Newsli-FA (Botha, Shan, and Gillick 2020)	71	326	4.6	WN	FA	WKD
Newsli-AR (Botha, Shan, and Gillick 2020)	675	3,612	5.4	WN	AR	WKD
Newsli-RO (Botha, Shan, and Gillick 2020)	226	385	1.7	WN	RO	WKD
Newsli-PL (Botha, Shan, and Gillick 2020)	186	196	1.1	WN	PL	WKD
Newsli-TA (Botha, Shan, and Gillick 2020)	835	2,474	3	WN	TA	WKD

context. This reduces the overall computational complexity to $O(|T| \cdot K \cdot \bar{C}^2)$, achieving near-linear scalability with the number of toponyms while preserving spatial consistency within each batch.

4. Experiments

This section presents a comprehensive evaluation of the proposed method regarding accuracy and runtime, benchmarked against 21 existing approaches across a diverse multilingual dataset collection. An ablation analysis is further performed to assess the individual influence of each scoring factor in the method. Moreover, a sensitivity study investigates how three major parameters affect the overall performance.

4.1. Parameter settings

All experiments were conducted using a single, consistent hyperparameter configuration, summarized in Table 2. We did not predefine a separate development set and did not carry out systematic hyperparameter optimization (e.g., grid search or cross-validation). Instead, the default values were determined through a small number of pilot adjustments guided by overall benchmark behaviour, after which they were *frozen* and applied uniformly across all datasets and languages (i.e., no dataset-specific or language-specific tuning). This “one-setting-for-all” design helps limit benchmark-specific overfitting and better reflects practical deployment scenarios where dataset-specific calibration is unavailable. We acknowledge that refining parameters based on benchmark feedback may introduce some optimistic bias. To address this concern, Section 4.7 reports a sensitivity analysis for the key UniTopRank parameters ($\lambda_{\text{spatial}}^{\text{base}}$, λ_{name} , λ_{pop} , λ_{admin} , P_{max} , d_{max} , and K), illustrating how performance varies when each parameter is perturbed around its default value.

4.2. Datasets

Rule-based methods are often criticized for limited generalizability; we therefore evaluate our approach on 25 datasets in 14 languages (186,100 toponyms) spanning seven text types (Table 3). For WikToR, only titles are annotated, so we extract additional toponyms from the article text using Flair NER (Akbić *et al.* 2019). For the NewsLi datasets, we use all splits and retain only entities with a GeoNames ID in Wikidata. Figures 2 and 3 show the spatial distributions.

4.3. Compared methods

In this study, we evaluate 21 methods, covering two major categories: (i) general-purpose entity linkers based on deep learning, and (ii) dedicated toponym resolution methods that employ rule-based, machine learning, deep learning, and recent LLM-based techniques. Table 4 provides an overview of all approaches. For rule-based and traditional machine learning approaches, the number of parameters is negligible compared to deep learning and LLM methods; hence, we denote their model size as 0. Most of these methods were included in the comparison by Hu *et al.* (2026, 2023a) and the description of these methods can be found in this paper.

Based on this, we include two transformer-based multilingual entity linkers: mGENRE (De Cao *et al.* 2022), the multilingual extension of GENRE, and BELA (Plekhanov *et al.* 2023), a BLINK-style multilingual entity linker that follows a bi-encoder retrieval and cross-encoder re-ranking pipeline. In addition, we include two recent transformer-based toponym ranking approaches that first retrieve candidate locations from GeoNames and then rank them: *Mordecai3* (Halterman 2023), for which we use the released pretrained model, and *SemTopRes* (Gomes, Purves, and Volpi 2024), which we fine-tune on the LGL dataset with the sentence-transformers/all-MiniLM-L6-v2. In addition, we included the FT-Mistral (7B) model (Hu *et al.* 2026), which was fine-tuned on the LGL dataset for toponym resolution and reported by the authors as the best-performing model. For the GeoNames and Photon methods, we

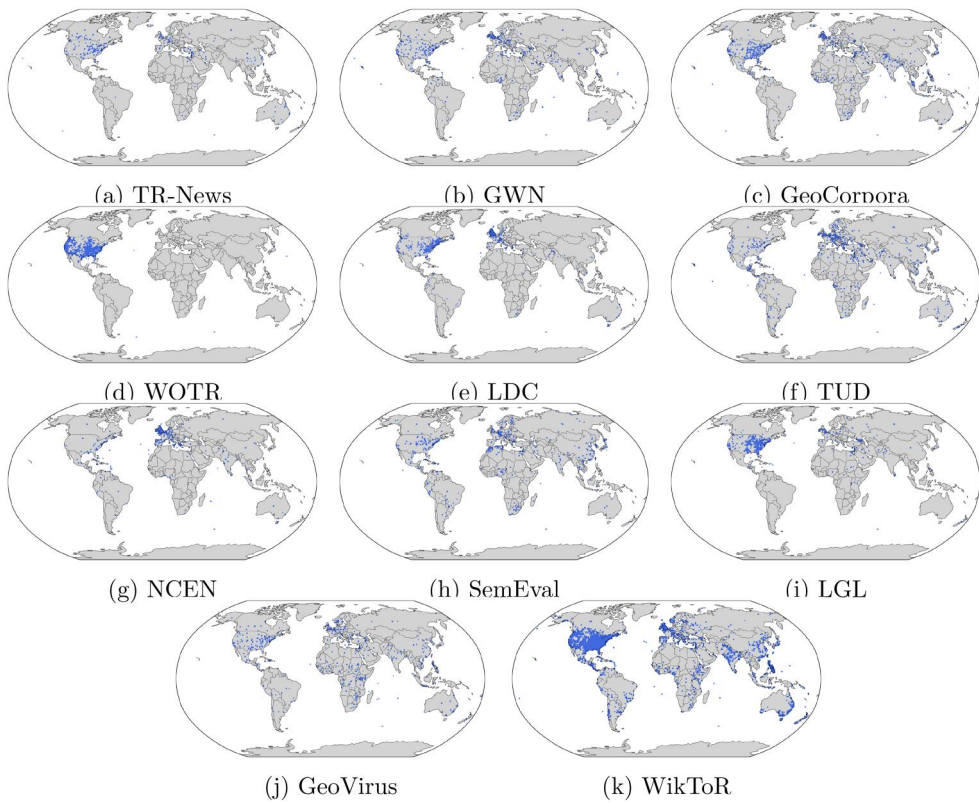


Figure 2. Spatial distributions of toponyms across English datasets.

retrieve the top-ranked result for each toponym from the respective gazetteer, which typically corresponds to the most populated or most prominent location. For non-English datasets, we compared toponym resolution methods and multilingual entity linkers, excluding English-specific linkers because they rely on English-centric resources and English-only models.

4.4. Metrics

For evaluation, we use three standard metrics, following Gritta, Taher Pilehvar, and Collier (2020):

- **Accuracy@161km**, the proportion of predicted coordinates that fall within 161 km (100 miles) of the gold location;
- **Mean Error (ME)**, the average distance error across all predictions; and
- **Area Under the Curve (AUC)**, which measures how errors are distributed across different distance scales. It is computed on a logarithmic scale, with 20,039 km (half the Earth’s circumference) as the maximum possible error. AUC can be understood as the “average error level” over all distance thresholds: a smaller AUC means that the method makes fewer large errors and is overall more reliable, while a larger AUC indicates that the method often produces very distant predictions.

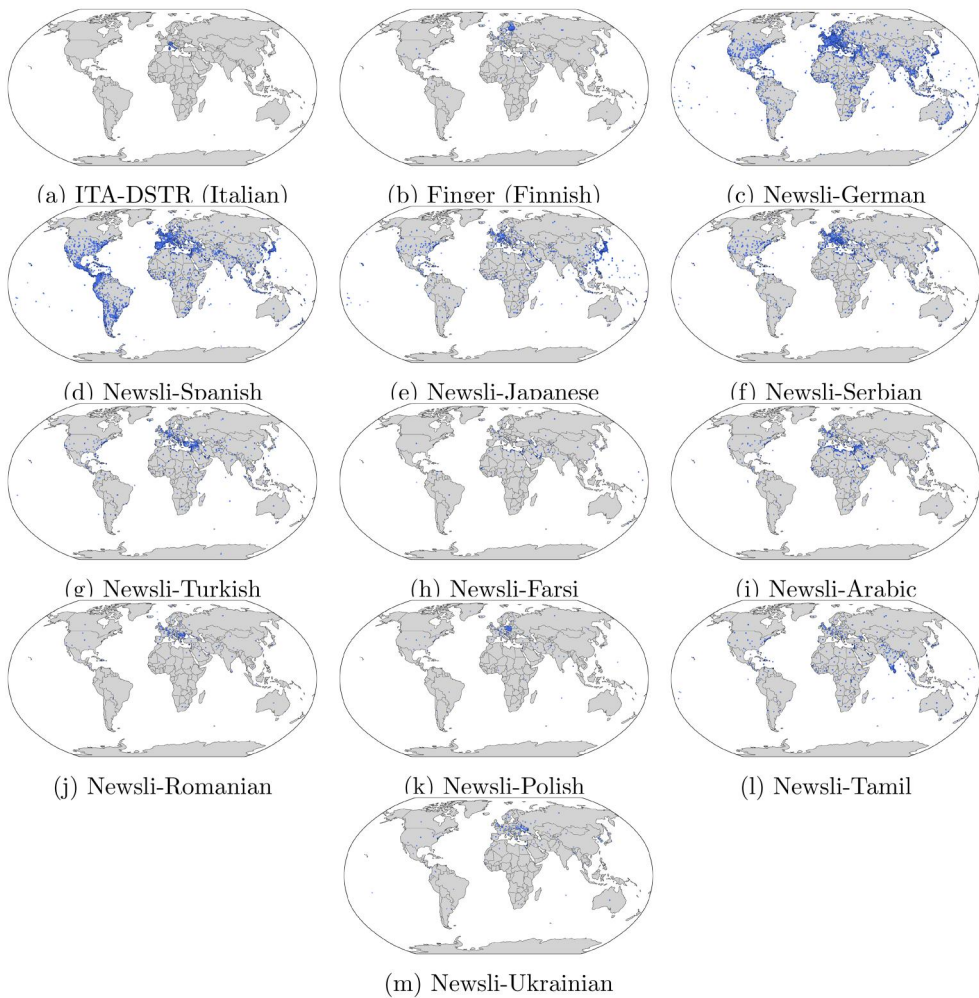


Figure 3. Spatial distributions of toponyms across non-English datasets.

Note that if a toponym cannot be resolved or linked to any location entity in the gazetteers or knowledge bases by a given method, we assign it a distance error equivalent to half of the Earth's circumference, approximately 20,039 km.

4.5. Comparison results

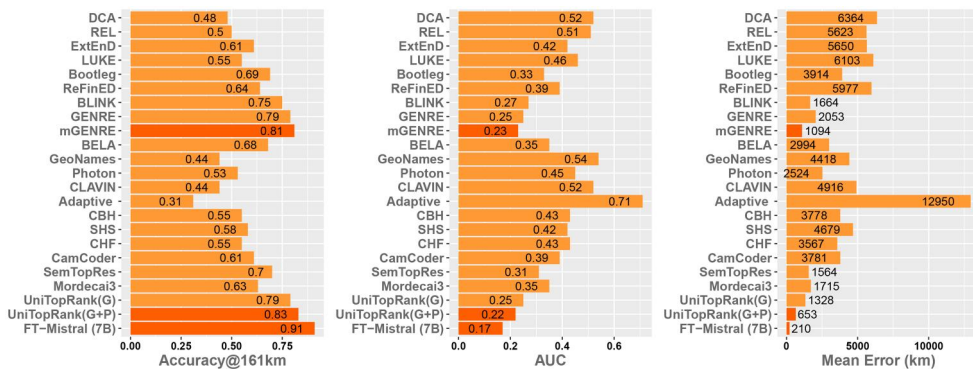
We aggregated all toponyms from the complete English datasets and computed overall scores for Accuracy@161km, ME, and AUC, which are shown in Figure 4. As expected, FT-Mistral (7B) achieved the best results, with Accuracy@161km reaching 0.91. Our proposed *UniTopRank-GN+Photon* outperformed all other competitors, while *UniTopRank-GN* matched the performance of GENRE and mGENRE, the best individual baselines outside the LLM-based method. We also observe that SemTopRes, the latest transformer-based method, outperforms all previous toponym resolution approaches but still falls behind the second-best entity linker, BLINK. Detailed per-dataset English results are provided in the (supplementary material Tables 4.5, 4.5, and 4.5).

Table 4. Representative approaches for resolving toponyms.

Name	Method Type	KB	Lang.	Model Size
MuRel (Le and Titov 2018)	EL (DL)	WP/WD	en	≈0.18M
DCA (Yang <i>et al.</i> 2019)	EL (DL)	WP	en	≈0.1M
BLINK (Wu <i>et al.</i> 2020)	EL (DL)	WD	en	680M
Bootleg (Orr <i>et al.</i> 2020)	EL (DL)	WD	en	110M
GENRE (De Cao <i>et al.</i> 2021)	EL (DL)	WP/WD	en	406M
mGENRE (De Cao <i>et al.</i> 2022)	EL (DL)	WP/WD	multi	610M
ExtEnD (Barba, Procopio, and Navigli 2022)	EL (DL)	WD	multi	435M
LUKE (Yamada <i>et al.</i> 2022)	EL (DL)	WP/WD	en	483M
ReFinED (Ayoola <i>et al.</i> 2022)	EL (DL)	WP/WD	en	154M
BELA (Plekhanov <i>et al.</i> 2023)	EL (DL)	WP/WD	multi	550M
Photon	TR (rule)	OSM	multi	0
GeoNames	TR (rule)	GN	multi	0
Adaptive (Lieberman and Samet 2012)	TR (ML+Rank)	GN	en	0
CLAVIN ^a	TR (rule)	GN	en	0
CBH (Kamalloo and Rafiei 2018)	TR (rule)	GN	en	0
SHS (Kamalloo and Rafiei 2018)	TR (rule)	GN	en	0
CHF (Kamalloo and Rafiei 2018)	TR (rule)	GN	en	0
CamCoder (Gritta, Pilehvar, and Collier 2018)	TR (DL+CLF)	WP	en	≈135M
Mordecai3 (Halterman 2023)	TR (DL+Rank)	GN	en	125M
SemTopRes (Gomes, Purves, and Volpi 2024)	TR (DL+Rank)	GN	en	22M
FT-Mistral(7B) (Hu <i>et al.</i> 2026)	LLM	GN/OSM/AG	en	7B
UniTopRank (Ours)	Rule	GN/OSM	multi	0

Abbreviations: EL = entity linker, TR = toponym resolution, ML = machine learning, DL = deep learning, WP = Wikipedia, WD = Wikidata, GN = GeoNames, OSM = OpenStreetMap, AG = ArcGIS, CLF = Classification.

^a<https://github.com/bigconnect/clavin>.

**Figure 4.** The overall Accuracy@161km, ME, and AUC on English datasets.

The WikToR, LDC, and WOTR datasets present particular challenges because many toponyms in these corpora are highly ambiguous, referring to multiple distinct locations that all appear within the datasets. For instance, “San Jose” corresponds to several provinces in the Philippines, including Nueva Ecija, Occidental Mindoro, Batangas, Camarines Sur, and Northern Samar. Similarly, “Richmond” denotes multiple places across North America, such as those in Cheshire County, New Hampshire; Ontario County, New York; McHenry County, Illinois; and Washington County, Rhode Island in the United States, as well as in Calgary, Alberta, Canada.

Such ambiguities explain the poor performance of most existing methods. Simple population- or popularity-based approaches, including GeoNames and Photon perform particularly poorly, with Accuracy@161km below 0.25 on the WikToR dataset. In

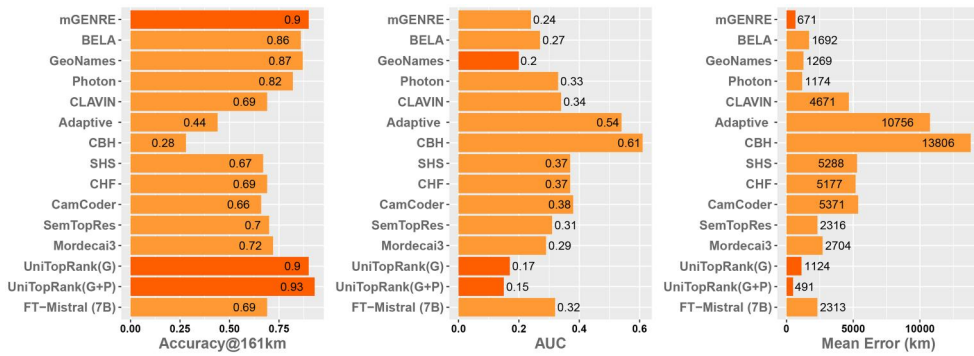


Figure 5. Overall Accuracy@161km, ME, and AUC on ITA-DSTR (Italian), FingerTweets (Finnish), and FingerNews (Finnish).

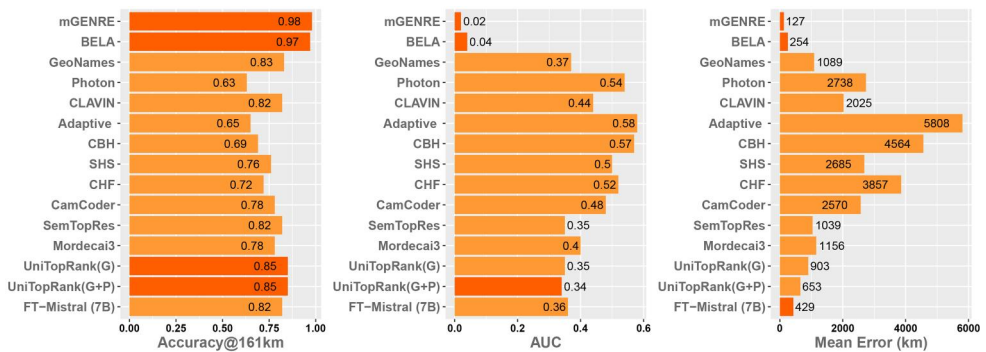


Figure 6. Overall Accuracy@161km, ME, and AUC on the Newsli datasets (RO, PL, UK, TR, FA, TA, SR, JA, ES, DE, AR).

contrast, our methods maintain competitive results, achieving scores above 0.85 on WikToR. This highlights the strong disambiguation capability of our proposed methods.

For the non-English datasets, we group them into two subsets: (i) ITA-DSTR (Italian), FingerTweets (Finnish), and FingerNews (Finnish), and (ii) the Newsli non-English datasets (RO, PL, UK, TR, FA, TA, SR, JA, ES, DE, AR). For each subset, we aggregate all toponyms and recompute the overall Accuracy@161km, ME, and AUC (Figures 5 and 6). Detailed per-dataset non-English results are provided in the (supplementary material Tables 4.5, 4.5, and 4.5). Figure 6 shows that mGENRE and BELA achieve very high Accuracy@161km (above 0.97). This can be attributed to strong domain and source alignment: NewsLi is derived from WikiNews-style articles, which closely match the Wikipedia/Wikimedia-centric text and knowledge distribution leveraged by both models. In addition, BELA explicitly performs an additional fine-tuning stage on a split of MEWSLI-9 prior to evaluation, which is expected to further strengthen in-domain performance on NewsLi-style data. Moreover, MEWSLI-9/NewsLi is a primary evaluation benchmark reported in both the mGENRE and BELA papers, which may contribute to stronger benchmark-specific alignment during model development. In contrast, the other three non-English datasets—ITA-DSTR, FingerTweets, and FingerNews (Figure 5)—as well as the 11 English datasets (Figure 4) come from more heterogeneous

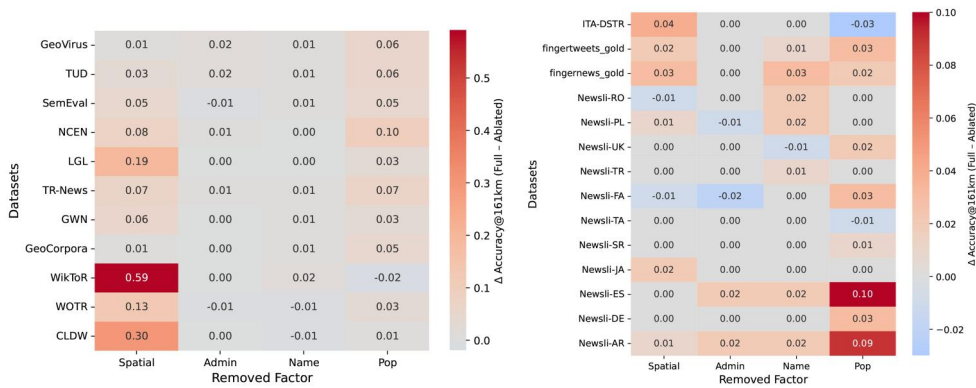
sources (e.g., social media and historical documents) and were not primary evaluation benchmarks in either model’s original study. This setting therefore better reflects out-of-distribution generalization. Under these conditions, both mGENRE and BELA drop substantially compared with their NewsLi performance, while our proposed *UniTopRank-GN+Photon* and *UniTopRank-GN* remain competitive and often achieve comparable or better results.

Our proposed methods outperform all other baselines, achieving higher accuracy than FT-Mistral (7B). This is because FT-Mistral (7B) was re-trained on the English LGL dataset, learning to predict ambiguous references for English toponyms. Consequently, when applied to non-English texts, the model often produces linguistically or contextually irrelevant outputs due to language barriers. For instance, it sometimes outputs an entire English sentence, such as “Parma man charged with murder in fatal shooting. A 20-year-old Parma man has been charged” instead of the correct geographic reference “Parma, Italy” for the toponym “Parma.” Similarly, for the Arabic phrase which refers to “Minya Governorate, Egypt”, the model incorrectly predicts “Buhayrah Governorate, Egypt.” These errors suggest that the model struggles to generalize across languages, possibly because it was fine-tuned exclusively on English-language data or because of limited robustness when processing low-resource languages. While this does not imply that LLM-based methods are inherently inferior to rule-based approaches in a multilingual context, it highlights a key limitation: current LLM-based toponym resolution systems require multilingual adaptation strategies, which, however, lie beyond the scope of this study.

Interestingly, the simple GeoNames baseline—which selects the first candidate returned by the gazetteer (typically the one with most population)—produces reasonable results. This suggests that toponyms in many non-English datasets are relatively unambiguous, and the most prominent location is often correct. Nevertheless, *UniTopRank-GN* consistently outperforms this method, demonstrating its ability to adapt to datasets with varying degrees of toponym ambiguity. We also observe that the improvement over the GeoNames baseline on non-English datasets is about 0.02, which is much smaller than the gain of 0.35 on English datasets. This can be attributed to two main factors. First, non-English datasets are generally less ambiguous. Second, they contain far fewer toponyms per document (see [Table 3](#)), averaging only 2.5 per document compared to 62 in English datasets. Since our method relies on toponym co-occurrence within documents for disambiguation, a smaller number of toponyms substantially reduces its effectiveness, as also shown in [Section 4.7](#).

4.6. Ablation analysis

This section examines how each of the four scoring factors in *UniTopRank-GN* contributes to toponym resolution performance. We conduct ablation experiments by removing either a single factor or a pair of factors, while keeping the remaining components unchanged, and quantify the impact as the drop in Accuracy@161km on English and non-English datasets.



(a) Ablation study on English datasets. (b) Ablation study on non-English datasets.

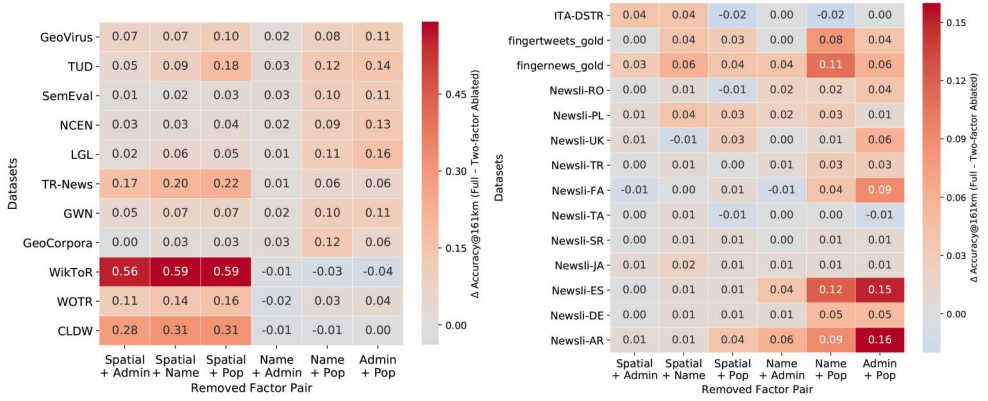
Figure 7. Ablation study of the four factors in *UniTopRank-GN*, reported as Δ Accuracy@161km (Full – Ablated) on (a) 11 English and (b) 14 non-English datasets.

4.6.1. Single-factor ablation

Figure 7a and b shows the accuracy drop for English and non-English datasets when removing one factor, respectively. The results indicate that the spatial factor contributes most significantly on difficult datasets, such as WikToR, which contain numerous ambiguous toponyms. Removing the spatial factor causes a decrease in Accuracy@161km for WikToR by 0.59 (from 0.85 to 0.26). This finding underscores that spatial relationships among toponyms play a crucial role in resolving geographic ambiguities. In contrast, for less ambiguous datasets, the influence of the spatial factor is smaller than that of population, which is particularly evident in non-English datasets. The other two factors, admin level and name similarity, contribute positively but are less important than spatial relationship and population.

4.6.2. Two-factor ablation (joint effects)

To verify joint effects (complementary vs. redundant), we remove two factors at a time (six pairs; Fig. 8) and interpret them with the single-factor drops in Fig. 7a and Fig. 7b. On highly ambiguous English data (e.g., WikToR), single-factor ablation shows that spatial is the clear bottleneck ($\Delta=0.59$), while the other factors have near-zero impact. Consistently, any pair that includes spatial (*Spatial* + *Admin*/*Name*/*Pop*) yields a similarly large drop (0.56–0.59), whereas pairs without spatial lead to almost no degradation (or slight improvements), indicating redundancy of the remaining cues once spatial evidence is removed. For several less extreme English datasets where population already contributes on its own (e.g., TUD and TR-News in Fig. 7a), removing spatial together with another cue—especially *Spatial* + *Pop*, but also *Spatial* + *Name* and *Spatial* + *Admin*—causes substantially larger drops than removing either factor alone, demonstrating complementary evidence in these cases. For the non-English datasets, single-factor results show that population is often the main driver, while spatial is typically weak. The two-factor ablation confirms this: the largest degradations consistently involve population, and *Name* + *Pop* and *Admin* + *Pop* frequently exceed the drop from removing population alone (e.g., Newsli-ES/AR), indicating that lexical matching and administrative



(a) Two-factor ablation on English datasets. (b) Two-factor ablation on non-English datasets.

Figure 8. Two-factor ablation study of the four scoring factors in *UniTopRank-GN*, reported as Δ Accuracy@161km (Full – Ablated) on (a) 11 English and (b) 14 non-English datasets.

constraints provide complementary signals when combined with the population prior. These two-factor ablations are meant as a diagnostic under the default UniTopRank setting; the observed effects may change with candidate sets, gazetteer coverage, or weight settings. Moreover, the joint effects are not uniform across datasets, depending on factors such as ambiguity and toponym density in a text. Therefore, Fig. 8 should be read as an empirical result under our experimental setting, not as a universal rule; at the same time, this dataset-level variation supports our design choice to combine all four cues, which helps maintain robust and stable ranking behaviour across different conditions.

4.7. Sensitivity analysis

We further evaluate the sensitivity of our method to key parameters and their impact on the overall Accuracy@161km across English and non-English datasets. We use Accuracy@161km rather than ME or AUC as it is the most intuitive and widely used metric in toponym resolution, effectively approximating the hit rate of correctly resolved toponyms within a reasonable geographic tolerance. For each parameter, we vary its value while keeping the others fixed to the default settings listed in Table 2, and compute the mean Accuracy@161km over all datasets within each group.

4.7.1. Weighting factors

Figure 9 suggests that the four factors behave quite differently across language groups. On the English datasets, performance improves quickly when the spatial weight increases from 0.1 to around 0.3, and then stays fairly stable; in contrast, putting more weight on the administrative and population signals steadily hurts accuracy. For the non-English datasets, we observe the opposite trend: a large spatial weight becomes harmful, while the population factor is consistently helpful and reaches its best range roughly around 0.4–0.5. The name similarity factor is comparatively insensitive in both settings, with only small changes across the tested weights.

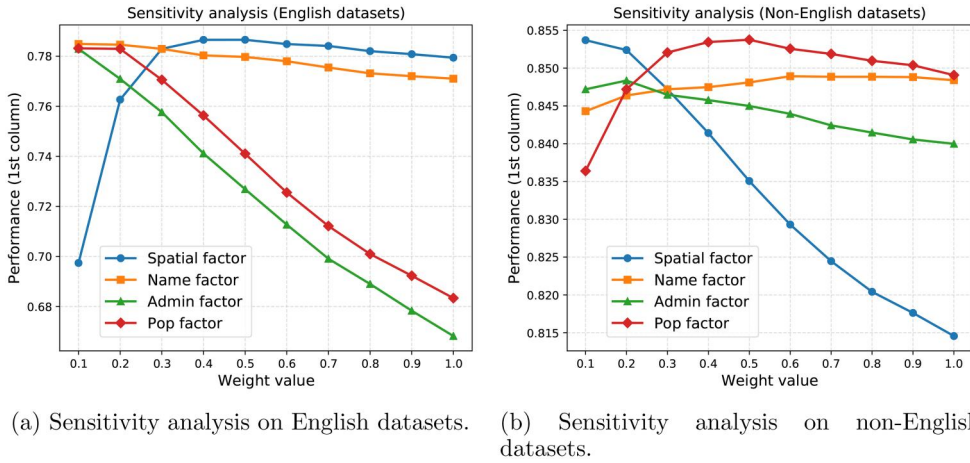


Figure 9. Sensitivity analysis of the four weighting factors in *UniTopRank-GN*. For each factor, we vary its weight from 0.1 to 1.0 while keeping the other weights at their default values, and report the resulting Accuracy@161 km on (a) 11 English and (b) 14 non-English datasets.

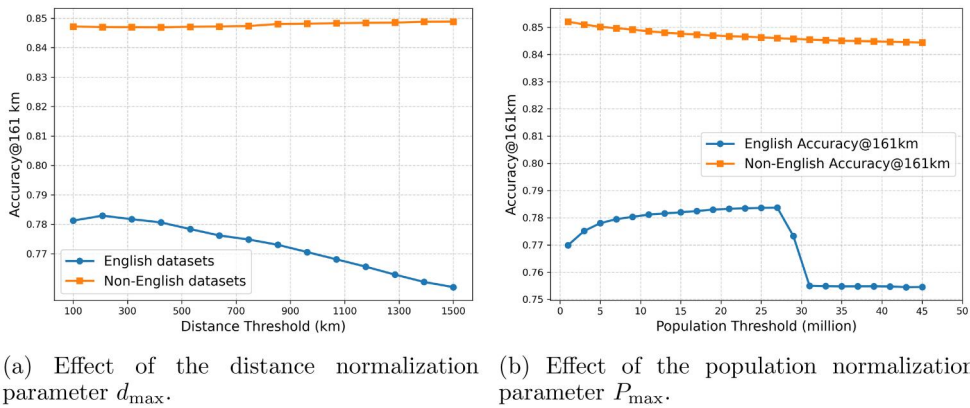


Figure 10. Sensitivity analysis of normalization parameters on overall Accuracy@161km: (a) distance normalization $d_{\max}$ and (b) population normalization $P_{\max}$, evaluated on English and non-English datasets.

4.7.2. Distance threshold

Figure 10a shows the effect of the distance normalization parameter $d_{\max}$ on accuracy. For non-English datasets, the performance remains nearly constant across the entire range, suggesting that spatial distance contributes little once administrative and textual cues are sufficient for disambiguation. In contrast, the accuracy on English datasets gradually decreases as $d_{\max}$ increases beyond 300 km. This is mainly because English datasets contain many ambiguous toponyms—such as multiple places named “Springfield” or “Cambridge”—that occur in different regions. In such cases, spatial distance acts as a key disambiguation cue: smaller thresholds help emphasize geographically coherent clusters of toponyms, whereas larger thresholds allow distant candidates to be treated as equally plausible, thus reducing resolution precision.

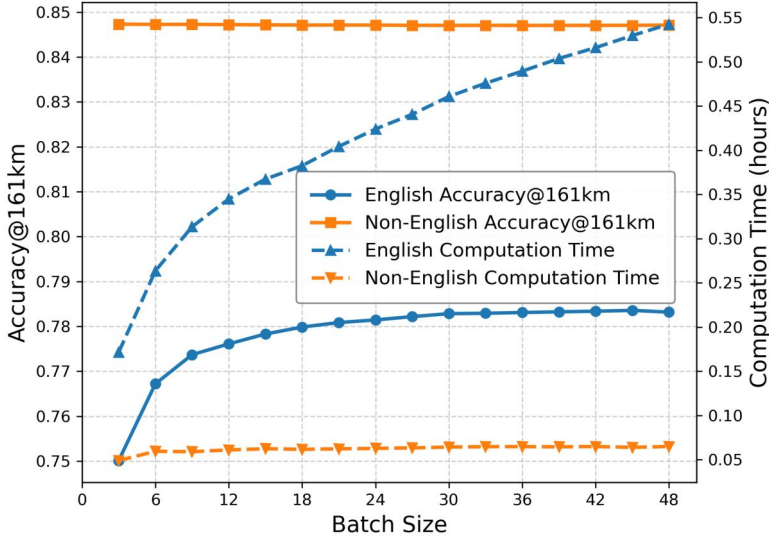


Figure 11. Sensitivity analysis of batch size K with respect to overall Accuracy@161 km and the corresponding computation time for English and non-English datasets.

4.7.3. Population threshold

Figure 10b shows the effect of the population normalization parameter $P_{\max}$ on overall accuracy. For English datasets, the accuracy increases steadily as $P_{\max}$ grows from small values to around 20 million. This reflects the role of the population prior in distinguishing between small settlements and major cities: when $P_{\max}$ is too small, large cities are excessively penalized, while moderate values provide a smoother and more realistic scaling of population influence. However, when $P_{\max}$ exceeds roughly 30 million, the relative differences between small, medium, and large settlements become too small to matter, causing the population prior to lose its ability to distinguish between them effectively. In contrast, non-English datasets show minimal sensitivity to this parameter.

4.7.4. Batch size

As illustrated in Figure 11, increasing the batch size K improves accuracy for English datasets up to around $K = 30$, after which the accuracy remains nearly unchanged. This trend indicates that spatial reasoning benefits from considering a sufficient number of toponyms within the same context, but adding more toponyms beyond this point provides little additional gain. Larger batch sizes also lead to a nearly linear increase in computation time. In contrast, the non-English datasets remain largely unaffected by this parameter, because the average number of toponyms per document is considerably smaller than in the English datasets (see Table 3), resulting in limited spatial interactions even with small K values. In practice, this setting allows users to flexibly choose K depending on their priorities—favoring either higher accuracy or faster processing. Notably, even with a small batch size of $K = 3$, our method achieves an Accuracy@161km of 0.75 on the English datasets and 0.85 on the non-English datasets, remaining competitive with other approaches while being substantially faster. It requires only about 0.17 hours to process all English datasets containing 97,819 toponyms and 0.05 hours for all non-English datasets containing 88,281 toponyms.

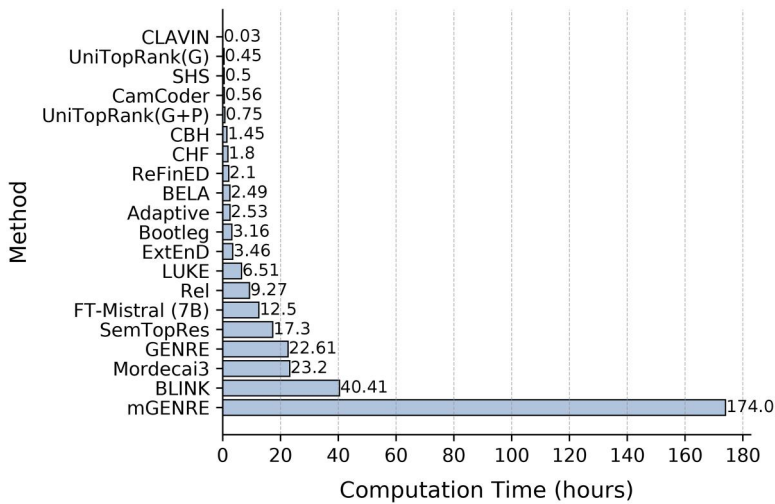


Figure 12. Runtime comparison of all evaluated methods on the complete English benchmark datasets.

4.8. Computational efficiency

While accuracy is critical for geoparsing, runtime is equally important, especially when fast inference is preferred over marginal precision gains. We therefore analyze computational efficiency.

Since all methods were evaluated on the English datasets, we base the analysis on English only. Reported runtimes cover inference only, excluding training and gazetteer/KB lookup. We also omit query time because it depends strongly on deployment (remote vs. local vs. in-memory access); accordingly, the GeoNames- and Photon-only methods are excluded. Neural entity-linking and LLM-based methods were run on an NVIDIA Tesla V100, and all other methods on CPU.

Figure 12 shows large efficiency differences. CLAVIN is fastest (about 0.03 hours). Our lightweight methods rank in the top five (0.45 and 0.75 hours) and are substantially faster than FT-Mistral (7B), GENRE, and mGENRE; notably, mGENRE requires nearly 170 hours.

5. Discussions

5.1. Limitations and generalization

UniTopRank relies on four gazetteer-based cues, and each can be less informative in certain situations. Name similarity can be weak when mentions are abbreviated, misspelled, or use nicknames not covered by the gazetteer. Population and administrative-level priors can be misleading in strongly local narratives, where small places are discussed more often than well-known ones (e.g., local reports focusing on small towns). Spatial coherence can weaken in texts that intentionally span distant regions, such as London(UK), Paris (France), Wuhan (China) and New York (US), and it is naturally less informative when only a few toponyms are available. Despite these limitations, UniTopRank is a metadata-driven ranking method that does not rely on

corpus-specific training; in practice, it can be transferred to new corpora with only lightweight adjustment of the cue weights if needed. We next discuss how gazetteer limitations affect candidate retrieval and downstream ranking, together with possible mitigation strategies.

5.2. Gazetteer coverage and metadata bias

UniTopRank is bounded by candidate generation: if the correct place is missing from $C(t)$, no ranking step can recover it. This is amplified in multilingual settings, where the same location can appear in divergent surface forms across scripts and languages, so lexical retrieval may fail even when the entry exists. Our results illustrate this bottleneck: for Tamil (ta), Ukrainian (uk), and Polish (pl), Accuracy@161km drops to 0.47, 0.65, and 0.64, and only 49%, 73%, and 72% of gold toponyms have at least one gazetteer entry within 161km of the true coordinates. A mitigation is two-stage candidate generation: conservative normalization and high-precision lexical rules first, then cross-script or missing-language recovery via multilingual dense retrieval (e.g., multilingual-E5 embeddings with a FAISS ANN index). Ranking further depends on the availability and consistency of gazetteer metadata, which is salient when fusing *GeoNames* with *OSM*-derived gazetteers: *OSM* candidates may lack population, tag-based typing is not directly comparable to *GeoNames* feature codes, and administrative fields and parent/containment relations may differ in granularity or naming (e.g., additional intermediate levels). Such heterogeneity can weaken or disable cues (population, administrative level, and the hierarchical component of spatial coherence) and can be mitigated by (i) down-weighting cues with missing fields and (ii) lightweight harmonization (e.g., tag-to-feature mappings and normalization of administrative labels). To assess transfer to a new setting, users can label a small sample and report (1) candidate recall and (2) end-to-end performance under the default configuration; optionally, users can slightly vary cue weights to verify stability.

5.3. RAG integration

The integration of RAG has recently emerged as a promising approach for enhancing LLM-based toponym resolution (Hu, Kersten, and Klan 2025; Lin *et al.* 2025). In the RAG workflow, the candidates of each toponym are initially retrieved from a gazetteer, and the language model then leverages this external geographic knowledge to identify the most plausible candidate. However, the retrieval step often yields an excessive number of candidates—sometimes over one hundred for highly ambiguous names such as Paris—which can introduce significant noise during inference and degrade overall accuracy. Therefore, an effective candidate ranking mechanism is critical to prioritize the most relevant and contextually consistent locations before they are passed to the language model. Our proposed scalable toponym ranking method addresses this need by acting as a ranking component within RAG pipelines, reducing low-ranking candidates and thereby improving reasoning efficiency. Beyond toponym resolution, this principle also applies to other geospatial tasks that retrieve geographic knowledge from external sources using toponyms in text.

6. Conclusion

This paper presents *UniTopRank*, a language-independent toponym resolution approach based on rule-driven candidate ranking. Experiments on 25 datasets in 14 languages, with 186,100 toponyms, show that *UniTopRank* attains accuracy comparable to state-of-the-art alternatives while requiring substantially lower computational cost and shorter runtime. These results indicate that *UniTopRank* is a practical option for large-scale, multilingual, and resource-constrained applications. We acknowledge that *UniTopRank* is not intended to replace high-capacity neural or LLM-based systems in settings where maximum accuracy is the primary objective and sufficient computational resources are available. One limitation is candidate coverage in low-resource languages, where missing or sparse gazetteer entries can reduce end-to-end performance. Future work will focus on improving multilingual candidate generation, integrating *UniTopRank* into *OpenWebSearch.eu* for large-scale multilingual web geo-indexing, and employing it as an efficient candidate-ranking component in RAG-based toponym resolution pipelines.

Notes

1. <https://www.dsayce.com/social-media/tweets-day/>
2. <https://en.wikipedia.org/wiki/Weibo>
3. <https://usearch.com/blog/how-many-news-articles-are-published-daily-on-the-web>
4. <https://diviflash.com/website-statistics/>
5. <https://openwebsearch.eu/>
6. <https://data.jrc.ec.europa.eu/collection/emm>
7. <https://www.padi-web-one-health.org/>
8. <https://amunatcoll.pl/>
9. <https://github.com/amagge/geonames-service>
10. <https://github.com/komoot/photon>
11. <https://nominatim.openstreetmap.org/search?q=london&format=json&addressdetails=1&limit=50>

Author contributions

CRedit: **Xuke Hu**: Conceptualization, Data curation, Methodology, Software, Writing – original draft; **Yao Sun**: Writing – review & editing; **Tobias Hecking**: Writing – review & editing; **Jens Kersten**: Writing – review & editing; **Friederike Klan**: Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Declaration of generative AI in manuscript preparation

ChatGPT-5 (instant version) was employed solely for linguistic refinement during the preparation of this manuscript. All substantive content development, revisions, and final edits were conducted by the authors, who assume full responsibility for the accuracy and integrity of the published work.

Funding

This work was supported by OpenSearch@DLR.

Notes on contributors

Xuke Hu is a senior researcher at the Institute of Data Science, German Aerospace Center (DLR). His current research focuses on geospatial NLP methods – particularly toponym recognition and resolution – that make geographic information in natural-language text explicit and reliable across languages and geographic contexts. He also investigates how such text-derived geographic information can be applied in disaster management, civil security and geo-privacy. He contributed to the methodology, software, data curation, writing, review and editing of this paper.

Yao Sun is a researcher at the department Photogrammetry and Image Analysis at the Earth Observation Center, German Aerospace Center (DLR). Her research interests include responsible artificial intelligence for Earth observation, disaster management, multi-modal data integration, and Open Science practices for equitable geospatial information access. She contributed to the review and editing of this paper.

Tobias Hecking received his Ph.D. degrees in computer science from the University of Duisburg–Essen, in 2016. Since 2020, he has been the Head of the Intelligent Software Systems Group at German Aerospace Center (DLR), Institute of Software Technology. His research focuses on the theories, methods, and tools of artificial intelligence for the development of data-driven intelligent software systems. He contributed to the review and editing of this paper.

Jens Kersten is a senior researcher at the Institute of Data Science, German Aerospace Center (DLR). His research focuses on new concepts and methods for geospatial information retrieval with the goal to utilize un- and semi-structured data as a reliable source of information for spatial applications and decision making, for example related to disaster management and situational monitoring based on web data. He contributed to the review and editing of this paper.

Friederike Klan heads a department at the Institute of Data Science, German Aerospace Center (DLR). She has a scientific background in computer science with a specialization on data acquisition and mobilization to address challenges in aerospace, energy, transport, and security. The focus of her interdisciplinary and applied research is on the development of innovative methods for collecting data, ensuring its quality, making it usable and deriving information from it. She contributed to the review and editing of this paper.

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 with the link <https://gitlab.com/dlr-dw/UniTopRank> and <https://figshare.com/articles/software/UniTopRank/30445541>.

References

- Acheson, E., and Purves, R.S., 2021. Extracting and modeling geographic information from scientific articles. *PloS One*, 16 (1), e0244918.
- Akbik, A., et al., 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

- Linguistics (Demonstrations)*, Minneapolis, Minnesota. Stroudsburg, PA: Association for Computational Linguistics, 54–59.
- Aldana-Bobadilla, E., et al., 2020. Adaptive geoparsing method for toponym recognition and resolution in unstructured text. *Remote Sensing*, 12 (18), 3041.
- Allen, T., et al., 2017. Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, 8 (1), 1124.
- Ardanuy, M.C., et al., 2022. A dataset for toponym resolution in nineteenth-century English newspapers. *Journal of Open Humanities Data*, 8, 1–7.
- Ardanuy, M.C., and Sporleder, C., 2017. Toponym disambiguation in historical documents using semantic and geographic features. In: *Proceedings of the 2nd international conference on digital access to textual cultural heritage*, Göttingen, Germany. New York, NY: Association for Computing Machinery, 175–180.
- Arsevska, E., et al., 2018. Web monitoring of emerging animal infectious diseases integrated in the French animal health epidemic intelligence system. *PloS One*, 13 (8), e0199960.
- Avvenuti, M., et al., 2018. GSP (Geo-Semantic-Parsing): geoparsing and geotagging with machine learning on top of linked data. In: *The Semantic Web: 15th International Conference, ESWC 2018*, 3–7 June 2018 Heraklion, Crete, Greece, Proceedings, Cham: Springer, 17–32.
- 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*, Hybrid: Seattle, Washington + Online. Stroudsburg, PA: Association for Computational Linguistics, 209–220.
- 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.
- Belliardo, E., Kalimeri, K., and Mejova, Y., 2023. Leave no place behind: improved geolocation in humanitarian documents. In: *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, Lisbon, Portugal. New York, NY: Association for Computing Machinery, 31–39.
- Botha, J.A., Shan, Z., and Gillick, D., 2020. Entity linking in 100 languages. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Stroudsburg, PA: Association for Computational Linguistics, 7833–7845.
- Chen, Y., et al., 2026. Geocoding the past world: unearthing coordinates of early China from texts using generative AI. *International Journal of Geographical Information Science*, 40 (3), 781–807.
- Curtis, A., et al., 2025. Providing spatial support during a major cholera outbreak in Port-au-Prince, Haiti: Creative mapping solutions in a challenging data poor environment. *Spatial and Spatio-Temporal Epidemiology*, 55, 100753.
- De Bruijn, J.A., et al., 2018. TAGGS: Grouping tweets to improve global geoparsing for disaster response. *Journal of Geovisualization and Spatial Analysis*, 2 (1), 2.
- De Cao, N., et al., 2021. Autoregressive entity retrieval. In: *International Conference on Learning Representations*. <https://openreview.net/forum?id=5k8F6UU39V>.
- De Cao, N., et al., 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10, 274–290. <https://aclanthology.org/2022.tacl-1.16>.
- 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. Stroudsburg, PA: Association for Computational Linguistics, 188–198.
- Farzana, S.M., and Hecking, T., 2024. Towards a scalable geoparsing approach for the web. In: *Proceedings of The GeoExT 2024: Geographic Information Extraction from Texts Workshop, co-located with the 46th European Conference on Information Retrieval (ECIR)*, Glasgow, Scotland. Aachen, Germany: CEUR Workshop Proceedings (CEUR-WS.org), 25–33.
- Gomes, D., Purves, R.S., and Volpi, M., 2024. Fine-tuning transformers for toponym resolution: A contextual embedding approach to candidate ranking. In: *Proceedings of The GeoExT 2024: Geographic Information Extraction from Texts Workshop, co-located with the 46th European*

- Conference on Information Retrieval (ECIR)*, Glasgow, Scotland. Aachen, Germany: CEUR Workshop Proceedings (CEUR-WS.org), 43–51.
- Gregory, I.N., and Ell, P.S., 2007. *Historical GIS: technologies, methodologies, and scholarship*. Vol. 39. Cambridge, UK: Cambridge University Press.
- 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. Stroudsburg, PA: Association for Computational Linguistics, 1285–1296.
- Gritta, M., Taher Pilehvar, M., and Collier, N., 2020. A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, 54 (3), 683–712.
- Gritta, M., et al., 2018. What's missing in geographical parsing? *Language Resources and Evaluation*, 52 (2), 603–623.
- Halterman, A., 2023. Mordecai 3: A neural geoparser and event geocoder. arXiv Preprint, arXiv: 2303.13675
- Hendriksen, G., et al., 2024. The open web index: Crawling and indexing the web for public use. In *European Conference on Information Retrieval*, 130–143. Springer.
- 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., Kersten, J., and Klan, F., 2025. Scalable toponym resolution with LLMs: Accuracy and speed optimizations. In: *Proceedings of The GeoExT 2025: Geographic Information Extraction from Texts Workshop, co-located with the 47th European Conference on Information Retrieval (ECIR)*, Lucca, Italy. Aachen, Germany: CEUR Workshop Proceedings (CEUR-WS.org), 46–51.
- Hu, X., et al., 2026. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *International Journal of Geographical Information Science*, 40 (3), 670–697.
- 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., 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, Y., and Adams, B., 2021. Harvesting big geospatial data from natural language texts. In: *Handbook of Big Geospatial Data*. Springer, 487–507.
- Hu, Y., et al., 2023b. 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.
- Kamalloo, E., and Rafiei, D., 2018. A coherent unsupervised model for toponym resolution. In: *Proceedings of The Web Conference 2018*, Lyon, France. New York, NY: Association for Computing Machinery, 1287–1296.
- Karimzadeh, M., et al., 2019. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23 (1), 118–136.
- Katz, P., and Schill, A., 2013. To learn or to rule: two approaches for extracting geographical information from unstructured text. *Data Mining and Analytics 2013 (AusDM'13)*, 117.
- Kulkarni, S., et al., 2020. Spatial language representation with multi-level geocoding. arXiv Preprint arXiv, 2008.09236
- 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. Stroudsburg, PA: Association for Computational Linguistics.
- Leidner, J.L., and Lieberman, M.D., 2011. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3 (2), 5–11.
- Leppämäki, T., Toivonen, T., and Hiippala, T., 2024. Geographical and linguistic perspectives on developing geoparsers with generic resources. *International Journal of Geographical Information Science*, 38 (10), 2039–2060.

- Li, S., et al., 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 119–133.
- 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*, Portland, Oregon, USA. New York, NY: Association for Computing Machinery, 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.
- Lin, J., et al., 2025. RACCOON: A retrieval-augmented generation approach for location coordinate capture from news articles. In: *Companion Proceedings of the ACM Web Conference 2025*, Sydney, NSW, Australia. New York, NY: Association for Computing Machinery, 1123–1127.
- McDonough, K., Moncla, L., and Van de Camp, M., 2019. Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora. *International Journal of Geographical Information Science*, 33 (12), 2498–2522.
- Moncla, L., et al., 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In: *Proceedings of the 22nd ACM sigspatial international conference on advances in geographic information systems*, Dallas/Fort Worth, TX, USA. New York, NY: Association for Computing Machinery, 183–192.
- Nowak, M.M., Stupecka, K., and Jackowiak, B., 2021. Geotagging of natural history collections for reuse in environmental research. *Ecological Indicators*, 131, 108131.
- Orr, L., et al., 2020. Bootleg: Chasing the tail with self-supervised named entity disambiguation. arXiv Preprint arXiv, 2010.10363.
- Plekhanov, M., et al., 2023. Multilingual end to end entity linking. arXiv Preprint arXiv, 2306.08896
- 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*, Redondo Beach, CA, USA. New York, NY: Association for Computing Machinery, 9–15.
- Reades, J., et al., 2025. The city as text. *Nature Cities*, 2 (9), 794–800.
- Recchia, G., and Louwerse, M., 2013. A comparison of string similarity measures for toponym matching. In: *Proceedings of the First ACM SIGSPATIAL International Workshop on Computational Models of Place (COMP '13)*, Orlando, FL, USA. New York, NY: Association for Computing Machinery, 54–61.
- Scalia, G., Francalanci, C., and Pernici, B., 2022. CIME: Context-aware geolocation of emergency-related posts. *Geoinformatica*, 26 (1), 125–157.
- Stein, L., et al., 2024. Wealth over Woe: Global biases in hydro-hazard research. *Earth's Future*, 12 (10), e2024EF004590.
- Syed, M.A., et al., 2025. GeospatRE: extraction and geocoding of spatial relation entities in textual documents. *Cartography and Geographic Information Science*, 52 (3), 221–236.
- Vračević, N., et al., 2025. More than just Tweets: the potential of alternative geo-social media data for disaster management. *Social Network Analysis and Mining*, 15 (1), 74.
- 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.
- Weissenbacher, D., et al., 2019. Semeval-2019 task 12: Toponym resolution in scientific papers. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA. Stroudsburg, PA: Association for Computational Linguistics, 907–916.
- Wu, L., et al., 2020. Zero-shot entity linking with dense entity retrieval. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Stroudsburg, PA: Association for Computational Linguistics, 6397–6407.
- 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. Stroudsburg, PA: Association for Computational Linguistics, 3264–3271.

- Yan, Z., et al., 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, 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. Stroudsburg, PA: Association for Computational Linguistics, 271–281.
- Zhang, Z., and Bethard, S., 2025. A survey on geocoding: algorithms and datasets for toponym resolution. *Language Resources and Evaluation*, 59 (2), 1775–1796.
- Zhang, Z., Laparra, E., and Bethard, S., 2024. Improving toponym resolution by predicting attributes to constrain geographical ontology entries. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 35–44.
- Zheng, Q., Juanle, W., and Jie, Z., 2025. Mapping public responses to natural disasters in China across text mining. *Geomatics World*, 32 (3), 307–318.
- Zou, L., et al., 2023. Social media for emergency rescue: An analysis of rescue requests on Twitter during Hurricane Harvey. *International Journal of Disaster Risk Reduction*, 85, 103513.