Evaluating the Readiness of "Ready-to-Use" Geoparsers*

A Case Study using the ACLED Data

Sebastian Gmür
Department of Geography
University of Zurich
Zurich, Switzerland
sebastian.gmuer@geo.uzh.ch

1 Introduction

Geoparsing is an essential component of Geographic Information Retrieval (GIR), involving two main tasks: identifying place names in unstructured texts (Toponym Recognition) and resolving these names to precise geographical coordinates (Toponym Resolution) [1]. This technology enables access to vast unstructured datasets, supporting applications ranging from historical research [2] to social media analysis [3]. Particularly when applied to social media data, robust geoparsing systems can facilitate real-time analyses during crises, enhancing disaster management efforts with actionable geographical insights [4,5].

Recent advancements in artificial intelligence, particularly in large language models (LLMs), have significantly influenced natural language processing. This progress has renewed interest in geoparsing and led to the rapid development of systems like the Irchel Geoparser [6] and Mordecai 3 [7], which leverage state-of-the-art technologies to address longstanding challenges in the field.

In fact, some researchers have begun to provocatively question whether the problem of geoparsing as a whole has already been solved. Titles such as "Are we there yet?" [8] and "Geoparsing: Solved or Biased?" [9] reflect this growing sentiment.

This project seeks to contribute to this ongoing discourse by evaluating the practical readiness of ready-to-use geoparsers, aiming to provide a broad assessment of how well current systems perform in real-world applications. By focusing on pre-built, out-of-the-box systems, this approach avoids deep technical comparisons of architectures or methodologies, instead exploring how a non-expert user might apply existing models.

To achieve this, two geoparser systems were selected for evaluation:

- **Edinburgh Geoparser**: a rule-based model [2].
- **Irchel Geoparser**: a novel transformer-based model [6].

*This Project was part of the Course GEO871 Geographic Information Retrieval at University of Zurich. The whole code can be accessed via: https://github.com/segmue/GIR_Project

For this study, a dataset not previously utilized in geoparsing research was employed: ACLED (Armed Conflict and Location Events Dataset) [10]. The choice of this dataset, which has yet to be explored by Geographic Information Retrieval (GIR) researchers, aims to test the practical deployment of geoparser models on real-world applications. At the time of writing, the dataset contains over 2 Million entries of armed conflict events, each one with a detailed description conaining location references, alongside structured information such as the place name and geographical coordinates [11].

This combination of structured and unstructured information provides all the necessary data to apply and evaluate geoparsing methods.

2 State of Research

2.1 Innovations and current Architectures

Recent advances in the field of machine learning have driven the development of new geoparsing models. Innovations have been particularly driven in toponym resolution.

For example, Halterman [7] presents **Mordecai 3**, an end-to-end system that integrates both toponym resolution and event geocoding. The model uses spaCy for Named Entity Recognition (NER) and a neural ranking model to match place names with entries in the GeoNames gazetteer. In these basic features, the model is very similar to the **Irchel Geoparser** by Gomes [6] also a "ready-to-use" end-to-end transformer-based model that also uses spaCy for NER to search and rank toponym candidates in the Geonames Gazetter [6]. Hu et al. [12] also apply a kind of ranking-based approach, but in which they fine-tune existing LLMs with gazetteer data to extract a disambiguous toponym reference based on the text context. Subsequently, these references are queried in established geocoders to obtain the unique coordinates [12].

In contrast to these ranking-based approaches, Cardoso et al. [13] use a localization-based approach. Their model replaces gazetteer matching with the direct prediction of geographic location. It works by using **HEALPix**, a Discrete Global Grid Systems

(DGGS), which can dispense with the use of geographic coordinates by determining the geographic location in hierarchical grid cells. In addition, the model integrates geophysical properties such as elevation and proximity to water zones, which improves the accuracy of ambiguous place names [13].

Most of the work highlighted so far has already dealt with toponym resolution, i.e., the traced process in the geoparsing pipeline. Hu et al. [14] on the other hand, developed **GazPNE2**, a model that specializes in toponym recognition in tweets. Using gazetteer data from OpenStreetMap and Geonames, a classification model was trained that distinguishes between potential place names and non-place names. The textual context of these identified candidates is then analyzed using a transformer model (BERTweet). The potential candidates from the gazetteers are then disambiguated again as a ranking-based model.

2.2 Challenges

Wang & Hu [8] investigated whether the problem of geoparsing can be declared solved. They identify several research gaps. Firstly, the performance of the tested models decreases rapidly for short and informally-written sentences. In addition, the tested models focus on well-known placenames such as cities, counties and do not consider geoparsing of fine-grained locations within cities, such as street names.

However, the situation is completely different for informal short texts (such as tweets), which can lead to drastic performance losses. The last problem highlighted is the dependency on Gazetter GeoNames as the main data source [8]

Liu et al. [9] take up this point again in their work and show that gazetteers such as GeoNames have a representation bias, both regarding high-populated areas and regarding a colonial perspective, which is related to their histories of origin. The challenges highlighted by Liu et al. [9] also make it clear that geographical biases result not only from the gazetteers themselves, but also from the way in which algorithms process this data. Representation bias is particularly evident in regions with low data coverage, such as Africa or South America.

2.3 Evaluation Frameworks

A standardized framework for the evaluation of geoparsers is indispensable for targeted research practice. Gritta et al. [15] have developed a fine-grained taxonomy for toponyms, provide standardized metrics such as Accuracy@161km and Mean Distance Error and AUC and compiled a benchmarking dataset with GeoWebNews [15]. Wang and Hu [16] introduced the EUPEG platform, which serves as a benchmarking tool for the systematic evaluation of geoparsers. EUPEG provides eight annotated corpora from different text genres and implements the same standard metrics [16]. This framework was extended by Hu et al.[17] and now includes 12 datasets, which are already used to evaluate new methods [6]. Specifically for Toponym Recognition, a comprehensive comparison was conducted by Hu et al. [1] who evaluated 27 different models on a total of 26 different datasets [1].

3 Experiment Design

The experiment aims to address the following research questions:

1. Are state-of-the-art geoparsers accurate and reliable in geocoding a formal text corpus?
2. What limitations can be identified when applying these geoparsers to a new dataset?
3. As a side question, could the ACLED dataset serve as a suitable benchmark corpus for evaluating the quality of geoparsers?

3.1 Description of ACLED

The Armed Conflict Location and Event Data Project (ACLED) provides a comprehensive database on global conflicts. For this study, two specific datasets were selected:

1. **Europe and Central Asia (2018–2024):** Initially comprising 442,000 entries.
2. **Africa (1997–2024):** Initially comprising 396,000 entries

Each entry in the dataset includes several detailed attributes:
Note: A brief textual description of the event.
Location: The name of the specific location where the conflict occurred. ACLED prioritizes recording the most specific location possible, using triangulated data from multiple sources. This includes names of populated places, natural landmarks, or neighborhoods within large cities. In such cases, locations may be recorded as "City Name – District Name" (e.g., Mosul – Old City) to provide additional granularity.
Latitude/Longitude: Geographical coordinates in the WGS84 format, provided with four decimal points to identify a central point within the specified location. These coordinates do not reflect precise positions such as street corners but rather centroids of the named areas.
Spatial Precision Code: Each event is assigned a geo-precision code (1–3) based on the specificity of the location information.

- **Code 1:** High precision, indicating coordinates for a specific town or city.
- **Code 2:** Moderate precision, representing a general area or approximate location near a town.
- **Code 3:** Low precision, used for larger regions or natural landmarks when no specific information is available.

A more detailed explanation of these attributes and their collection methodology is provided in the ACLED Codebook [11].

3.2 Preprocessing

Dataset preparation was a critical step. Entries were removed where the location mentioned in the **Location** column did not appear as a substring in the **Note** column. Manual inspection revealed that such entries often involved alternative spellings or locations not directly mentioned in the text. After preprocessing, the datasets were reduced to:

- Europe and Central Asia: 425,000 entries.
- Africa**:** 293,000 entries.

3.3 Selection of evaluated Geoparsers

Two geoparsers employing different approaches to toponym recognition and resolution were selected:

Edinburgh Geoparser: A rule-based system that identifies toponyms and resolves them using the GeoNames gazetteer [2].

Irchel Geoparser: A hybrid system using spaCy for toponym recognition and a ranking model for gazetteer resolution. For this setup, the spaCy en_core_web_sm Model was used, in combination with the GeoNames Gazetter and the dguzh/geo-all-MiniLM-L6-v2 transformer model [6].

3.4 Setup of Edinburgh Geoparser

Implementing EDG presented several challenges:

Pipeline Limitations: The standardized Edinburgh Geoparser pipeline processes individual text files, generating a separate XML file for each output. To handle over 700,000 entries, a Python script converted each entry into a .txt file and grouped them into four batches. A shell script iteratively executed the geoparser pipeline script for each file in a batch, parallelized across five subprocesses.

API Limitations: By default, Edinburgh Geoparser relies on external Gazeteer APIs for gazetteer lookups. To circumvent API rate limits, a local PostgreSQL database was set up with the complete GeoNames database.

Compatibility: The EDG only supports macOS systems, which fortunately was available.

3.5 Setup of Irchel Geoparser

The Irchel Geoparser setup was more straightforward:

Installation: The Geoparser is distributed as a Python package with preconfigured commands. These commands automatically set up a local GeoNames instance, install the required spaCy language model, and load the transformer model.

Environment: Google Colab was used for offloading computational tasks. Due to compatibility issues with PyTorch 0.2.0, an earlier version of Irchel Geoparser (0.1.8) was employed, which functioned without errors.

Batch Processing: The ACLED data was split into batches of 10,000 entries. Each batch was processed independently, and the results were appended as new columns to the original data.

3.6 Evaluation Metrics

The output of each geoparser included lists per entry containing:

- All identified toponyms as found in the text.
- Corresponding geographical coordinates.

Evaluation focused only on toponyms that matched exactly with the **Location** attribute in ACLED. Additional toponyms detected in the text but not matching the attribute were excluded. This evaluation method follows the approach used by Kenyon et al. [18], differing from the one proposed by Gritta et al. [15].

Evaluation criteria included:

1. Toponym Recognition:
    - Precision, Recall, and F1 scores were calculated.
    - Statistical significance was tested using McNemar's Test for pairwise comparisons [15].
2. Toponym Resolution:
    - Distance Error was measured using Great Circle Distance (Haversine Distances).
    - Metrics included Median/Mean Distance Errors, Accuracy@k for various k thresholds, and Area Under Curve (AUC), following the implementation by Hu et al. [17].
    - Results were statistically analyzed using a two-tailed Wilcoxon Signed-Rank Test [15].

4 Results

The study evaluated the performance of the Edinburgh and Irchel geoparsers across two datasets: Africa and Europe. Metrics were categorized into two main areas: Toponym Recognition and Toponym Resolution.

4.1 Toponym Recognition Results

In the Europe Dataset, the Edinburgh Geoparser identified toponyms in approximately 85% of all entries. Among all entries, around 54% had their toponyms correctly recognized (Recall). The precision for Edinburgh was 61%, which led to an F1 score of 57%. The Irchel Geoparser, in contrast, processed 75% of the entries and achieved a slightly higher recall of 56%. Its precision, however, was notably higher at 78%, leading to a superior F1 score of 65%. This demonstrates that while both models performed well in Europe, Irchel's higher precision gave it an edge in overall accuracy.

In Africa, both geoparsers showed varied performance. The Edinburgh Geoparser maintained consistency, recognizing toponyms in 86% of all entries. It achieved a recall of 60%, with a precision of 70%, resulting in an F1 score of 65%.The Irchel Geoparser, however, exhibited a significant decline in performance. While it processed 77% of the entries, its precision dropped sharply to 59%, and its recall fell to 46%. This combination resulted in the lowest F1 score observed in the study,

at 51% (See Table 1). The p-values of the McNemars Test were highly significant ($5e^{-141}$ and 0) for both regions, indicating that there is a significant Difference in Misclassification between both models.

Table 1: Toponym Recognition Results

Region	Geoparser	Result Returns	Precision	Recall	F1
Europe	Edinburgh	87.82%	0.61	0.54	0.57
Europe	Irchel	72.70%	0.78	0.56	0.65
Africa	Edinburgh	86.06%	0.7	0.6	0.65
Africa	Irchel	77.43%	0.59	0.46	0.51

4.1 Toponym Resolution Results

The evaluation of toponym resolution performance highlights significant differences between the geoparsers, as well as clear regional variations.

While the **mean distance error** ranged from 257 km to 740 km across all models and regions, it was clearly influenced by extreme outliers and is therefore not used further. Instead, the **median distance error** provides a more reliable measure of performance. Figure 1 shows the distance error distribution for both regions. Besides the substantial differences between the two regions, the Irchel Geopareser returned the narrowest Error Distribution in Europe (thus the "best" result) and simultaneously the widest error distribution in Africa.

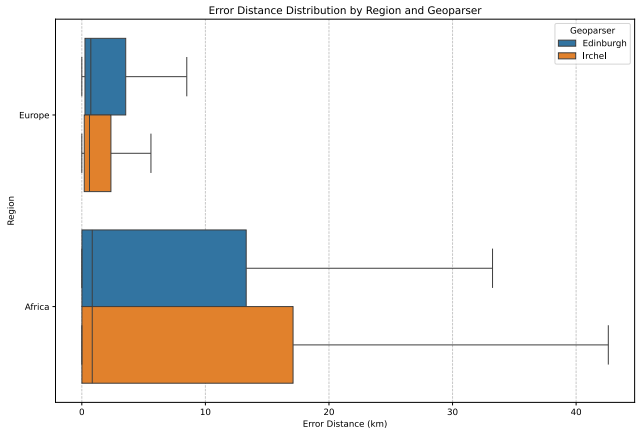


Figure 1: Distribution of meaured Error Distances

In Europe, the Irchel Geoparser outperformed the Edinburgh Geoparser in terms of median distance error, achieving 621 meters, compared to Edinburgh's 728 meters. Furthermore, Irchel's Accuracy@161km was higher at 47%, compared to Edinburgh's 36%. However, Edinburgh showed a slightly higher AUC score of 0.7327, compared to 0.7027 for Irchel.

In Africa, the Edinburgh Geoparser demonstrated greater consistency. It achieved a median distance error of 844 meters, the same as Irchel, but excelled in Accuracy@161km, reaching 52%, while Irchel only achieved 34%. Despite these differences, the Irchel Geoparser recorded a slightly better AUC score of 0.7043, compared to Edinburgh's 0.6855 (See Table 2).

Overall, the Irchel Geoparser showed strong performance in Europe, particularly in achieving lower median distance errors, while the Edinburgh Geoparser demonstrated more robust performance in Africa, especially in Accuracy@161km and error distribution stability. The p-values of the Wilcoxon Signed-Rank Tests were highly significant ($1.3e^{-36}$ and $8.3e^{-108}$), indicating that there is a significant difference in the distance errors between the models.

Table 2: Toponym Resolution Results

Region	Geoparser	Mean Distance Error (m)	Median Distance Error (m)	Accuracy@161km	AUC
Europe	Edinburgh	301084.36	728	0.36	0.7327
Europe	Irchel	257824.41	621	0.47	0.7027
Africa	Edinburgh	503306.61	844	0.52	0.6855
Africa	Irchel	742671.54	844	0.34	0.7043

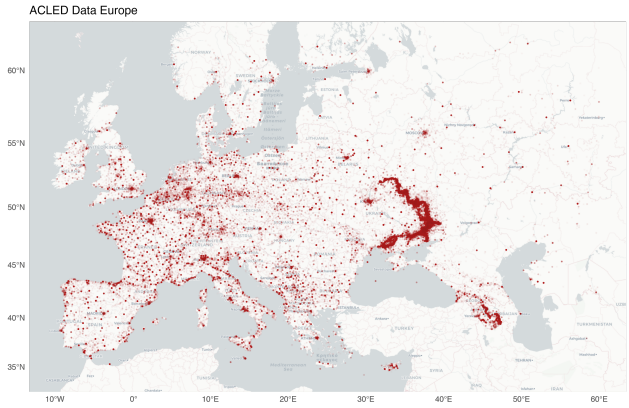
5 Discussion

5.1 Concerning ACLED

The ACLED dataset proved to be a versatile resource for evaluating ready-to-use geoparsers. It is remarkable that both the Edinburgh and Irchel Geoparsers could be applied to this new dataset with minimal preparation and still achieve reasonable results. The dataset was only lightly preprocessed, ensuring that the gold toponyms appeared as substrings within the text. Despite this, the Irchel Geoparser achieved toponym recognition performance on the European dataset comparable to the reported results of Gomes [6]. This suggests that ACLED has the potential to serve as a new evaluation benchmark for geoparsing models.

The scalability of ACLED further highlights its value. With an additional 600,000 entries available for Asia and 480,000 for the Middle East, the dataset has the potential to support global evaluations of geoparsers. However, one limitation not explored in this study is the uneven spatial distribution of ACLED entries. Conflict events are highly spatially clustered and subject to reporting biases, which may overrepresent certain areas while significantly underrepresenting others. As illustrated in **Figure 2**, the European dataset from ACLED demonstrates this spatial imbalance. Such clustering and biases should be taken into account in future studies to ensure comprehensive evaluations.

Figure 2: Map of ACLED Events in Europe



5.1 Concerning GeoNames vs ACLED

The error distance distributions reveal a striking difference between the African and European datasets, with significantly broader error distributions observed in Africa for both geoparsers. This observation should not be oversimplified as simply reflecting poorer performance in Africa by geoparsers. Instead, it underscores the influence of the quality of both the ground truth data (ACLED) and the predicted coordinates (GeoNames).

ACLED coordinates are manually determined, relying on the subjective judgment of data collectors. The extent to which these individuals possess local knowledge is unclear, especially given that ACLED is an initiative rooted in the Global North. Meanwhile, the density of place names in GeoNames is significantly lower in Africa, which likely limits the resolution process. This dual challenge—imprecision in ground truth data and limitations in GeoNames—illustrates the dependency on GeoNames as a gazetteer and highlights the lack of alternative resources. As Wang and Hu [8] noted, the reliance on GeoNames introduces a systemic bias in geoparsing evaluations, which is intensified in regions like Africa where GeoNames coverage is sparse.

5.1 Concerning the Geoparsers

The evaluation results highlight stark differences in the performance of the geoparsers, particularly on the African dataset. While the Edinburgh Geoparser maintained or slightly improved its toponym recognition metrics, the Irchel Geoparser experienced a notable performance drop. This decline may be linked to an algorithmic bias in the spaCy Named Entity Recognition (NER) model used. Liu et al. [9] previously identified NER cold spots in spaCy in Africa, which could account for the lower recognition rates.

In contrast, the toponym resolution results on the European dataset demonstrate the strengths of Irchel’s ranking-based transformer architecture. By consistently selecting the "better" toponym candidates from the GeoNames gazetteer, the Irchel Geoparser achieved higher coordinate precision. However, this advantage diminished on the African dataset. Despite both geoparsers showing the same median distance error (844 meters), the Irchel Geoparser exhibited a broader error distribution and significantly lower Accuracy@161km values. This indicates that the Irchel Geoparser produced considerably higher error distances in its incorrect predictions. At this stage, this study cannot provide a definitive explanation for this discrepancy, highlighting an area for future investigation.

6 Conclusion

This study evaluated the practical readiness of two ready-to-use geoparsers, the Edinburgh Geoparser and the Irchel Geoparser, using the ACLED dataset. By focusing on pre-built systems and avoiding in-depth architectural comparisons, the study aimed to provide a broad assessment of the current state of geoparsing models in real-world applications.

Both geoparsers demonstrated a solid initial performance on the previously unexplored ACLED dataset, showcasing its potential as a benchmark resource for future research. The Irchel Geoparser excelled in Europe, achieving higher precision and F1 scores compared to the Edinburgh Geoparser, but its performance declined significantly in Africa. In contrast, the Edinburgh Geoparser behaved differently, achieving notably better toponym recognition performance in Africa than in Europe. These findings emphasize the challenges posed by variations in dataset quality and gazetteer coverage between regions.

The ACLED dataset stands out due to its unparalleled global scope and massive size, which far exceeds previously used datasets in geoparsing research. While this study provided an initial assessment of the dataset, its full potential has yet to be explored. This initial evaluation already reveals clear regional biases and uneven spatial distributions (See Figure 2). For future research, careful consideration of these spatial biases is essential. Addressing these limitations could enable ACLED to establish itself as a foundational dataset for the field, supporting global evaluations of geoparsing systems.

The results of this study indicate that modern deep learning approaches can increase geographic bias. This is driven by their dependency on GeoNames, a gazetteer with uneven coverage, and the inherent algorithmic biases within trained language models. ACLED cannot mitigate this issue, as the dataset itself exhibits significant geographic biases. Without targeted efforts to address these challenges, the reliance on biased datasets and models risks creating a self-reinforcing cycle, where biased data is used to train new systems, perpetuating existing inequities.

6.1 Closing Remarks

The review of the current state of research highlighted the importance of a unified evaluation framework [15 – 17]. While these frameworks are partially being adopted, they have two critical shortcomings:

1. **Training on Evaluation Datasets:** Halterman [7] used have used existing datasets like GeoWebNews and TR News to train Mordecai 3. These Datasets however are also in use as Evaluation Datasets, e.g. in EUPEG [16]. While using portions of evaluation datasets for model training is not inherently problematic (due to train-test splits), it creates the risk that subsequent research may inadvertently use these models without understanding their prior exposure to the evaluation data. Without meticulous tracking of dataset usage, there is a potential danger that evaluation will occur on data previously used for training further down the research pipeline. Such scenarios compromise the validity of results and obscure true model performance.
2. **Lack of Bias Metrics:** Liu et al. [9] proposed additional metrics to assess geographic bias, which are not yet part of standard evaluation practices. The absence of these metrics in a geographically focused discipline is a significant oversight. Incorporating such metrics into standard frameworks is essential to accurately measure and mitigate geographic bias, as demonstrated by the findings of this study.

REFERENCES

[1] Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2023. Location Reference Recognition from Texts: A Survey and Comparison. *ACM Comput. Surv.* 56, 5 (November 2023), 112:1-112:37. https://doi.org/10.1145/3625819

[2] Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, 1925 (August 2010), 3875–3889. https://doi.org/10.1098/rsta.2010.0149

[3] Stuart E. Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Trans. Inf. Syst.* 36, 4 (June 2018), 40:1-40:27. https://doi.org/10.1145/3202662

[4] Yingjie Hu and Jimin Wang. 2020. How do people describe locations during a natural disaster: an analysis of tweets from Hurricane Harvey. *LIPIcs, Volume 177, GIScience 2021* 177, (2020), 6:1-6:16. https://doi.org/10.4230/LIPIcs.GIScience.2021.I.6

[5] Jyoti Prakash Singh, Yogesh K. Dwivedi, Nripendra P. Rana, Abhinav Kumar, and Kawaljeet Kaur Kapoor. 2019. Event classification and location prediction from tweets during disasters. *Ann Oper Res* 283, 1 (December 2019), 737–757. https://doi.org/10.1007/s10479-017-2522-3

[6] Diego Gomes. 2024. Geoparser: A Transformer-Based Bi-Encoder Approach for Efficient Toponym Disambiguation. University of Zurich.

[7] Andrew Halterman. 2023. Mordecai 3: A Neural Geoparser and Event Geocoder. https://doi.org/10.48550/arXiv.2303.13675

[8] Jimin Wang and Yingjie Hu. 2020. Are We There Yet? Evaluating State-of-the-Art Neural Network based Geoparsers Using EUPEG as a Benchmarking Platform. https://doi.org/10.48550/arXiv.2007.07455

[9] Zilong Liu, Krzysztof Janowicz, Ling Cai, Rui Zhu, Gengchen Mai, and Meilin Shi. 2022. Geoparsing: Solved or Biased? An Evaluation of Geographic Biases in Geoparsing. *AGILE GIScience Ser.* 3, (June 2022), 1–13. https://doi.org/10.5194/agile-giss-3-9-2022

[10] Clionadh Raleigh, Roudabeh Kishi, and Andrew Linke. 2023. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanit Soc Sci Commun* 10, 1 (February 2023), 1–17. https://doi.org/10.1057/s41599-023-01559-4

[11] ACLED. 2019. *ACLED Codebook*. Armed Conflict Location & Event Data Project (ACLED). Retrieved from www.acleddata.com

[12] Xuke Hu, Jens Kersten, Friederike Klan, and Sheikh Mastura Farzana. 2024. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *International Journal of Geographical Information Science* 0, 0 (2024), 1–28. https://doi.org/10.1080/13658816.2024.2405182

[13] Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2021. A Novel Deep Learning Approach Using Contextual Embeddings for Toponym Resolution. *IJGI* 11, 1 (December 2021), 28. https://doi.org/10.3390/ijgi11010028

[14] Xuke Hu, Zhiyong Zhou, Yeran Sun, Jens Kersten, Friederike Klan, Hongchao Fan, and Matti Wiegmann. 2022. GazPNE2: A General Place Name Extractor for Microblogs Fusing Gazetteers and Pretrained Transformer Models. *IEEE Internet Things J.* 9, 17 (September 2022), 16259–16271. https://doi.org/10.1109/JIOT.2022.3150967

[15] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A Pragmatic Guide to Geoparsing Evaluation. https://doi.org/10.48550/arXiv.1810.12368

[16] Jimin Wang and Yingjie Hu. 2019. Enhancing spatial and textual analysis with EUPEG: an extensible and unified platform for evaluating geoparsers. *Transactions in GIS* 23, 6 (December 2019), 1393–1419. https://doi.org/10.1111/tgis.12579

[17] Xuke Hu, Yeran Sun, Jens Kersten, Zhiyong Zhou, Friederike Klan, and Hongchao Fan. 2023. How can voting mechanisms improve the robustness and generalizability of toponym disambiguation? *International Journal of Applied Earth Observation and Geoinformation* 117, (March 2023), 103191. https://doi.org/10.1016/j.jag.2023.103191

[18] Jeremy Kenyon, Jason W. Karl, and Bruce Godfrey. 2023. Evaluation of Placename Geoparsers. *Journal of Map & Geography Libraries* 19, 3 (September 2023), 185–197. https://doi.org/10.1080/15420353.2024.2357115