Scalable Toponym Resolution with LLMs: Accuracy and Speed Optimizations

Xuke Hu[1,*], Jens Kersten[1] and Friederike Klan[1]

[1]*Institute of Data Science, German Aerospace Center, Jena, 07745, Germany*

Abstract

Toponym resolution plays a crucial role in geoparsing. While recent approaches leveraging lightweight LLMs, such as Mistral 7B, have shown promise, they still suffer from inefficiency and suboptimal accuracy. In this work, we propose an improved method that enhances both inference speed and resolution accuracy. Instead of processing toponyms individually, our approach resolves multiple toponyms simultaneously within the same text, leveraging contextual relationships among toponyms to refine predictions while reducing inference time. Furthermore, we integrate Retrieval-Augmented Generation (RAG) to incorporate candidate locations retrieved from GeoNames during inference, providing additional geographic context and improving disambiguation. To further accelerate processing, we adopt vLLM as an optimized inference engine. Experimental results on seven public datasets with 83,365 toponyms demonstrate that our solution increases accuracy from 0.90 to 0.93 and is seven times faster than previous LLM-based methods using the same base model.

Keywords

geoparsing, toponym resolution, large language model

1. Introduction

Unstructured texts, such as news articles, historical documents, and social media posts, contain valuable geographic information. Geoparsing, the process of extracting this information, is crucial for applications like spatial humanities [1], geographic search [2], and disaster management [3]. It consists of two main tasks: toponym recognition (identifying toponyms) and toponym resolution (determining their geographic coordinates or spatial footprints). While toponym recognition has made significant strides in accuracy[4, 5, 6], toponym resolution remains challenging. The resolution of toponyms has been approached through two distinct methodologies. The first, traditional toponym resolution, focuses specifically on toponyms or location entities, which can be further classified into several categories. These include rule-based ranking methods [7, 8, 9], which search gazetteers to identify potential candidates for a toponym and then rank or score these candidates through manually defined rules; statistical learning-based ranking approaches [10, 11, 12], which resemble rule-based methods but differ in that the rules are not explicitly defined but instead learned from annotated examples; and statistical-based classification methods [13, 14, 15], which partition the Earth's surface into discrete cells and subsequently assign toponyms to specific cells. The second are entity linkers [16, 17], which extend beyond the resolution of toponyms to link broader entities—such as persons, organizations, and locations—with corresponding entries in knowledge bases (KBs).

To address the limitations of individual methods, we proposed an ensemble approach [18] that integrates seven approaches using a voting mechanism. Inspired by recent advances in LLMs, we investigated their effectiveness for toponym resolution. Specifically, we fine-tuned lightweight models, such as Mistral-7B [19], to generate unambiguous toponym references, which are then geocoded using GeoNames and Nominatim[20]. This approach improved accuracy by 8% over the previously best-performing voting-based method.

GeoExT 2025: Third International Workshop on Geographic Information Extraction from Texts at ECIR 2025, April 10, 2025, Lucca, Italy

[*] Corresponding author.

Xuke.Hu@dlr.de (X. Hu); Jens.Kersten@dlr.de (J. Kersten); Friederike.Klan@dlr.de (F. Klan)

0000-0002-5649-0243 (X. Hu); 0000-0002-4735-7360 (J. Kersten); 0000-0002-1856-7334 (F. Klan)

However, the LLM-based approach still has two key limitations. First, it processes each toponym individually, even when multiple toponyms appear in the same text, which may cause the LLM to overlook important contextual relationships between the toponyms. Additionally, geographic knowledge is utilized only after inference, leaving the LLM without crucial disambiguation cues in inference. Second, resolving toponyms individually requires multiple inference passes, which increases computational cost.

To address these issues, we propose a parallel inference approach that resolves multiple toponyms simultaneously, enhancing both accuracy and computational efficiency. We further integrate RAG to incorporate candidate locations for toponyms from GeoNames during inference, and utilize vLLM [21] to accelerate processing.

2. Proposed Approach

Our approach, illustrated in Figure 1, consists of two phases: training (fine-tuning) and inference. For training, we use the LGL dataset [22], the same as in our previous work [20], which is the Local-Global Lexicon (LGL) corpus. This corpus contains 588 human-annotated news articles with 5,088 toponyms from 78 local newspapers. We employ Mistral-7B-v0.2 as the base model and apply Low-Rank Adaptation (LoRA) [23] for model fine-tuning.

Figure 3 in the Appendix presents a training example. In **Instruction**, we specify the toponyms to be resolved. The *char_index* in «START:*char_index*» indicates the starting character index of the toponym in the text, which helps distinguish between multiple occurrences of the same name. Candidate locations for each toponym are retrieved from GeoNames and ranked in two steps: first grouped into exact (matching the primary or alternative names) and non-exact matches, then ranked by population within each group. The top candidates from the exact group are selected first; if fewer than 17 are found, the highest-ranked non-exact matches are added to complete the Top-17. They are then included in **Instruction**. In **Input**, the target toponyms are marked within the original text. The **Output** consists of unambiguous references for all the toponyms.

During inference, candidate locations are queried and ranked using the same strategy, and the Top-17 candidates are retained. Given the context that includes the target toponyms and their candidate locations, the model generates an unambiguous reference for each toponym. This reference consists of the toponym's full name along with its higher-level administrative divisions (e.g., country, state, or province) necessary to uniquely identify the location. Each reference is sequentially queried in GeoNames and, if necessary, in Nominatim to obtain geographic coordinates. GeoNames is deployed locally, while Nominatim is queried online as a fallback, mainly for fine-grained locations such as points of interest. Most references are resolved by GeoNames, with only a small fraction requiring Nominatim. A caching mechanism is also employed to ensure fast query performance.

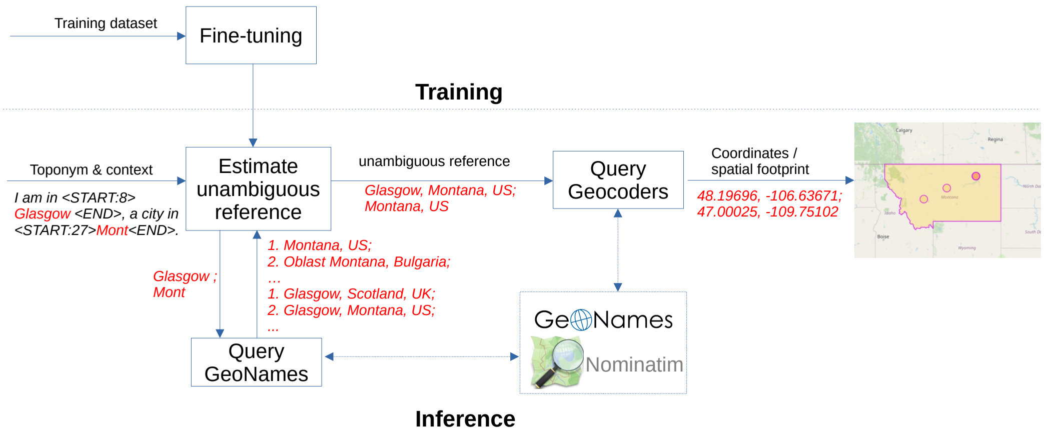

Figure 1: Workflow of the proposed approach.

3. Experiments and Evaluation

3.1. Experimental Settings

For LoRA, we set the attention dimension, scaling parameter, and dropout rate to 16, 16, and 0.1, respectively. We used the AdamW optimizer for fine-tuning with a learning rate of 0.003, over 300 epochs, and a batch size of 16. This fine-tuning process was performed on an NVIDIA Tesla V100 GPU, utilizing approximately 14 GB of GPU memory.

For testing, we employed the seven public datasets: **TR-News** [24], **GeoWebNews** [13], **GeoCorpora** [25], **WikToR** [26], **WOTR** [27], **CLDW** [28], and **NCEN** [29], which together contain a total of 83,365 toponyms. The geographical distribution of the toponyms in the test datasets is shown in Figure 4 in the Appendix. We evaluated the accuracy using *Accuracy@161km* [30], which measures geocoding precision within 161 km (100 miles).

We compared our approach with 10 representative methods, including transformer-based entity linkers (BLINK [31], GENRE [17]), rule-based toponym resolution approaches (CLAVIN[1], CHF [24]), deep learning-based classification (CamCoder [13]), and the voting-based ensemble method [18] that integrates seven individual approaches. We also included our previous LLM-based solution [20], which infers each toponym independently and does not incorporate candidate locations during inference. This solution was applied across four models, including Mistral (7B), referred to as FT-Mistral (7B).

3.2. Experimental Results

Figure 2 demonstrates that the new solution based on the Mistral-7B-v0.2 model improves the accuracy of the previous solution [20] using the same base model from 0.9 to 0.93. This performance matches that of the previous solution using the Llama2-70B model. Compared with non–LLM-based approaches, the new method improves the voting ensemble by 11% and GENRE, the best individual approach, by 15%. Moreover, the new solution is seven times faster than our previous implementation using the same base model, achieving performance that is comparable to—or even exceeds—that of traditional deep learning– and rule-based methods such as CHF and CamCoder.

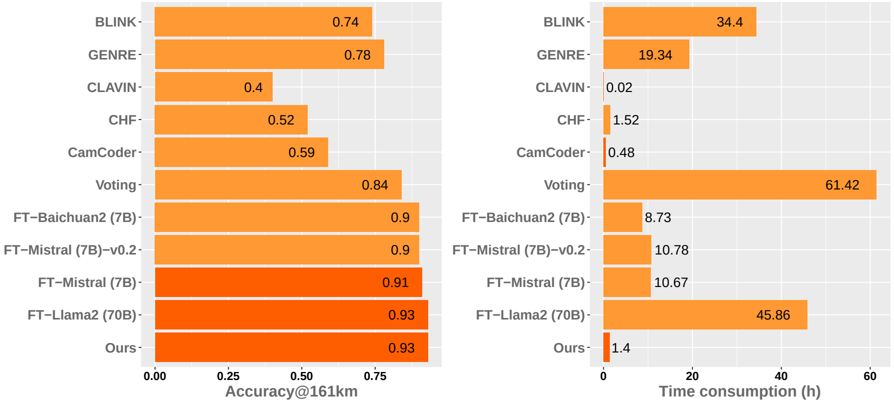

Figure 2: Evaluation of accuracy and speed across various approaches.

4. Conclusion

In this work, we presented an improved approach for toponym resolution based on a light-weight LLM, leveraging parallel inference, RAG, and faster inference engines to enhance both accuracy and efficiency. Our experiments on seven public datasets demonstrate that the proposed solution outperforms previous

[1] https://github.com/Novetta/CLAVIN

LLM-based methods, achieving higher accuracy and significantly faster processing times. One limitation of the approach lies in the candidate ranking algorithm, which currently considers only string similarity and population. This may result in the correct candidate being excluded from the Top-17 list. In future work, we will propose a more robust ranking strategy that considers the spatial relationships among toponyms within the same text.

Declaration on Generative AI

The authors employed ChatGPT and Mistral to polish the text. Following this, the manuscript underwent a thorough review and necessary modifications by the authors, who assume complete responsibility for the final content.

References

[1] I. Gregory, C. Donaldson, P. Murrieta-Flores, P. Rayson, Geoparsing, gis, and textual analysis: current developments in spatial humanities research, International Journal of Humanities and Arts Computing 9 (2015) 1–14.

[2] R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, V. Murdock, Geographic information retrieval: Progress and challenges in spatial search of text, Foundations and Trends in Information Retrieval 12 (2018) 164–318.

[3] Y. Zhang, Z. Chen, X. Zheng, N. Chen, Y. Wang, Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data, Journal of Hydrology 603 (2021) 127053.

[4] Y. Hu, G. Mai, C. Cundy, K. Choi, N. Lao, W. Liu, G. Lakhanpal, R. Z. Zhou, K. Joseph, Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages, International Journal of Geographical Information Science 37 (2023) 2289–2318.

[5] X. Hu, Z. Zhou, Y. Sun, J. Kersten, F. Klan, H. Fan, M. Wiegmann, GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models, IEEE Internet of Things Journal 9 (2022) 16259–16271.

[6] X. Hu, Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, F. Klan, Location reference recognition from texts: A survey and comparison, ACM Computing Surveys 56 (2023) 1–37.

[7] E. Aldana-Bobadilla, A. Molina-Villegas, I. Lopez-Arevalo, S. Reyes-Palacios, V. Muñiz-Sanchez, J. Arreola-Trapala, Adaptive geoparsing method for toponym recognition and resolution in unstructured text, Remote Sensing 12 (2020) 3041.

[8] A. Molina-Villegas, V. Muñiz-Sanchez, J. Arreola-Trapala, F. Alcántara, Geographic named entity recognition and disambiguation in mexican news using word embeddings, Expert Systems with Applications 176 (2021) 114855.

[9] D. Weissenbacher, T. Tahsin, R. Beard, M. Figaro, R. Rivera, M. Scotch, G. Gonzalez, Knowledge-driven geospatial location resolution for phylogeographic models of virus migration, Bioinformatics 31 (2015) i348–i356.

[10] X. Wang, C. Ma, H. Zheng, C. Liu, P. Xie, L. Li, L. Si, Dm_nlp at semeval-2018 task 12: A pipeline system for toponym resolution, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 917–923.

[11] D. Gomes, R. S. Purves, M. Volpi, Fine-tuning transformers for toponym resolution: A contextual embedding approach to candidate ranking., in: GeoExT@ ECIR, 2024, pp. 43–51.

[12] Z. Zhang, S. Bethard, Improving toponym resolution with better candidate generation, transformer-based reranking, and two-stage resolution, arXiv preprint arXiv:2305.11315 (2023).

[13] M. Gritta, M. Pilehvar, N. Collier, Which melbourne? augmenting geocoding with maps (2018).

[14] S. Kulkarni, S. Jain, M. J. Hosseini, J. Baldridge, E. Ie, L. Zhang, Spatial language representation with multi-level geocoding, arXiv preprint arXiv:2008.09236 (2020).

[15] Z. Yan, C. Yang, L. Hu, J. Zhao, L. Jiang, J. Gong, The integration of linguistic and geospatial features using global context embedding for automated text geocoding, ISPRS International Journal of Geo-Information 10 (2021) 572.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[17] N. De Cao, G. Izacard, S. Riedel, F. Petroni, Autoregressive entity retrieval, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=5k8F6UU39V.

[18] X. Hu, Y. Sun, J. Kersten, Z. Zhou, F. Klan, H. Fan, How can voting mechanisms improve the robustness and generalizability of toponym disambiguation?, International Journal of Applied Earth Observation and Geoinformation 117 (2023) 103191.

[19] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[20] X. Hu, J. Kersten, F. Klan, S. M. Farzana, Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge, International Journal of Geographical Information Science (2024) 1–28.

[21] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, in: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.

[22] M. D. Lieberman, H. Samet, J. Sankaranarayanan, Geotagging with local lexicons to build indexes for textually-specified spatial data, in: 2010 IEEE 26th international conference on data engineering (ICDE 2010), IEEE, 2010, pp. 201–212.

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[24] E. Kamalloo, D. Rafiei, A coherent unsupervised model for toponym resolution, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1287–1296.

[25] J. O. Wallgrün, M. Karimzadeh, A. M. MacEachren, S. Pezanowski, Geocorpora: building a corpus to test and train microblog geoparsers, International Journal of Geographical Information Science 32 (2018) 1–29.

[26] M. Gritta, M. T. Pilehvar, N. Limsopatham, N. Collier, What's missing in geographical parsing?, Language Resources and Evaluation 52 (2018) 603–623.

[27] G. DeLozier, B. Wing, J. Baldridge, S. Nesbit, 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), 2016, pp. 188–198.

[28] P. Rayson, A. Reinhold, J. Butler, C. Donaldson, I. Gregory, J. Taylor, 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, 2017, pp. 9–15.

[29] M. C. Ardanuy, D. Beavan, K. Beelen, K. Hosseini, J. Lawrence, K. McDonough, F. Nanni, D. van Strien, D. C. Wilson, A dataset for toponym resolution in nineteenth-century english newspapers, Journal of Open Humanities Data 8 (2022).

[30] M. Gritta, M. T. Pilehvar, N. Collier, A pragmatic guide to geoparsing evaluation, Language resources and evaluation 54 (2020) 683–712.

[31] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Zero-shot entity linking with dense entity retrieval, in: EMNLP, 2020.

A. Additional Figures

Instruction: Given the following toponyms, marked with unique identifiers in the text, estimate their full addresses or unambiguous references. The full address should include the toponym's formal name followed by its higher-level administrative units, such as city, state/province, and country.
Input Toponyms: 'Hurricane' marked with «START:23»«END»; 'Hurricane' marked with «START:372»«END»; 'Putnam County' marked with «START:398»«END»
Reference Information: Below are candidate addresses for each toponym. While these are valid options, the correct address might not be included among them. Use your geographic knowledge, along with these candidates, to infer the most accurate address.
Partial candidate locations of 'Hurricane':
1. Hurricane, Washington County, Utah, United States;
2. Hurricane, Putnam County, West Virginia, United States;
...
17. Lac Hurricane, Mauricie, Quebec, Canada;

Partial candidate locations of 'Putnam County':
1. Putnam County, New York, United States;
2. Putnam County, West Virginia, United States;
...
17. Putnam County, Indiana, United States;

Input: Sheriff's deputies and «START:23» **Hurricane** «END» police found the two.‘The guys were down in the area for about 10 minutes when they found them near the tracks,' the dispatcher said. Neither of the teenagers were hurt and both are safe, the dispatcher said. Keya Phillips and Seth Pettry, both 14, were found Monday after a tip came in that the two were hiding under a railroad bridge in «START:372» **Hurricane** «END» , according to a «START:398» **Putnam County** «END» dispatcher.

Output: («START:23»Hurricane, Putnam County, West Virginia, United States «END»)
(«START:372»Hurricane, Putnam County, West Virginia, United States «END»)
(«START:398»Putnam County, West Virginia, United States «END»)

Figure 3: Training example illustrating unambiguous reference estimation for multiple toponyms at one time.

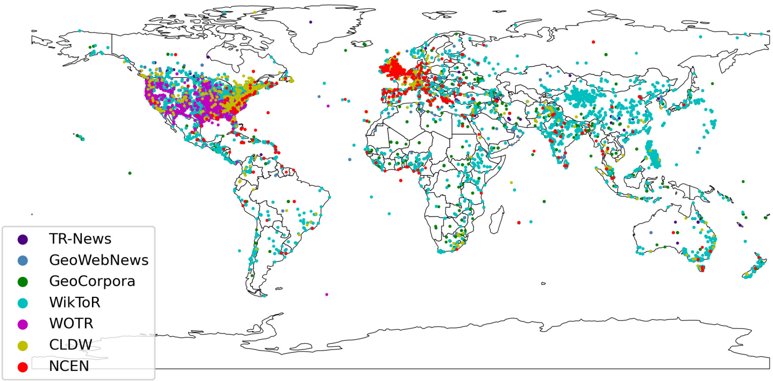

Figure 4: Geographical distribution of 83,365 toponyms from the seven datasets.