Enhancing Toponym Resolution with Fine-Tuned LLMs (Llama2)

Xuke Hu[1,*], Jens Kersten[1]

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

Abstract

In this study, we investigate the use of mid-sized and open-source large language models to enhance the extraction of geographic information from texts, focusing on toponym resolution. Our approach involves fine-tuning Llama2 (7B) to accurately derive the unambiguous references of toponyms within textual contexts and subsequently assign geo-coordinates using geocoders. The method is evaluated on two challenging datasets featuring 28,342 global toponyms. The findings demonstrate notable performance improvements compared to existing state-of-the-art methods while maintaining computational efficiency.

Keywords

geoparsing, toponym resolution, large language model, Llama2

1. Introduction

Unstructured texts such as news articles, historical documents, and social media posts are rich sources of geographic information. The extraction of this information, known as geoparsing, is essential in areas like spatial humanities [1], geographic search [2], and disaster management [3]. Geoparsing involves two key steps: toponym recognition (identifying toponyms in texts) and toponym resolution (inferring the geo-coordinates of these toponyms). While toponym recognition has advanced notably [4][5][6], toponym resolution still faces challenges in disambiguation accuracy [7].

In the rapidly evolving field of natural language processing, large language models (LLMs) such as GPT4 have brought significant changes, also impacting research in geoparsing [8][9]. Yet, existing studies using LLMs for geoparsing focus primarily on toponym recognition. Our research, in contrast, targets the more complex sub-task of geoparsing: toponym resolution. Specifically, we fine-tuned Llama2 (7B) [10], an open-source and powerful model in language comprehension and inference, to estimate toponyms' unambiguous references, followed by their conversion to geographical coordinates using free geocoders. Our approach demonstrates greater efficacy than seveal leading methods across two challenging datasets. Besides, the approach is computationally efficient, requiring about 14 GB of memory for operation on a standard GPU.

GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, 2024, Glasgow, Scotland

*Corresponding author.

xuke.hu@dlr.de (X. Hu); Jens.Kersten@dlr.de (J. Kersten)

0000-0002-5649-0243 (X. Hu); 0000-0002-4735-7360 (J. Kersten)

2. Proposed approach

Our approach, depicted in Figure 1, involves two phases: training (fine-tuning) and geocoding. Initially, we fine-tune Llama2 using Low-Rank Adaptation (LoRA) [11], a technique that optimizes GPU resource usage, to predict the unambiguous references (e.g., city, state, county) of toponyms based on their context. Our training dataset is the LGL[1] (Local-Global Lexicon) corpus, developed by Lieberman et al. [12], comprising 588 human-annotated news articles with 5088 toponyms from 78 local newspapers. In the geocoding phase, the fine-tuned model first deduces the unambiguous reference of toponyms from their contextual cues. It is then fed into a sequence of free geocoders—primarily GeoNames[2], followed by Nominatim[3] and ArcGIS[4]. This sequential querying strategy is designed to consult the next geocoder if one fails, enhancing the reliability and precision of the geocoding process.

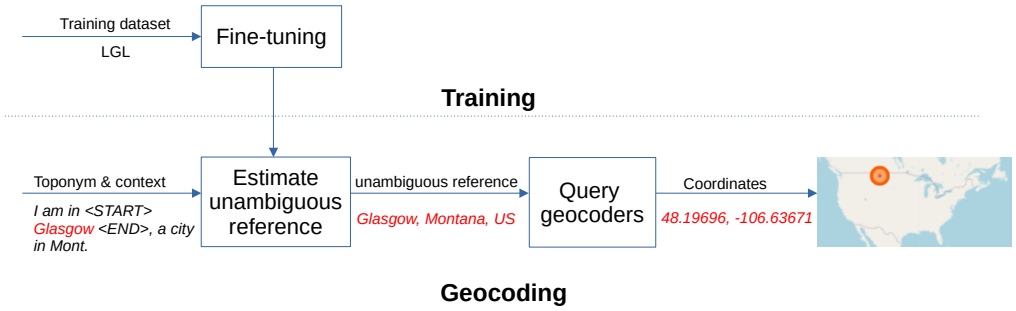

Figure 1: Workflow of the proposed approach.

3. Experiments and evaluation

3.1. Experimental setting

For LoRA, the attention dimension, the scaling parameter (*alpha*), and the dropout rate are set to 8, 16, and 0.1, respectively. We employed the AdamW optimizer for fine-tuning with a learning rate of 0.003, over 300 epochs, and a batch size of 128. This fine-tuning process was executed on an NVIDIA Tesla V100 GPU, utilizing about 14 GB of GPU memory.

For testing, we used two public datasets, detailed in Table 1. The geographical distribution of the toponyms in the test dataset is shown in Figure 2. Our evaluation employed two metrics [13]: *Accuracy@161km* for geocoding precision within 161 km (100 miles), and *Mean Error (ME)* for average distance error.

[1]https://github.com/milangritta/Pragmatic-Guide-to-Geoparsing-Evaluation/blob/master/data/Corpora/lgl.xml
[2]https://www.geonames.org/
[3]https://nominatim.org/
[4]https://developers.arcgis.com/documentation/mapping-apis-and-services/geocoding/

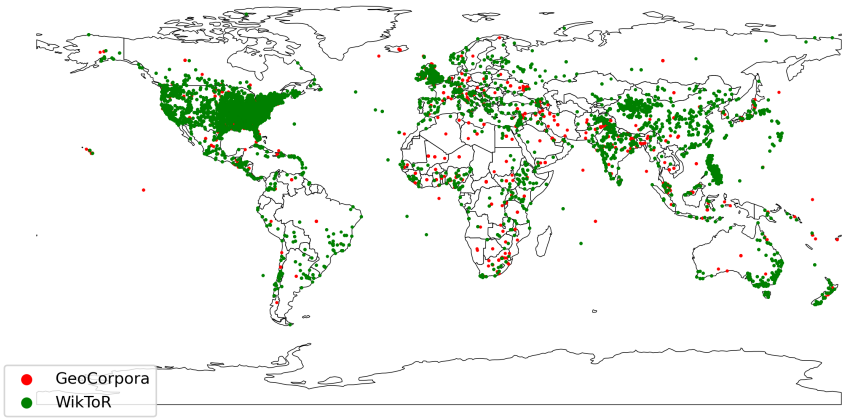

Figure 2: Geographical spread of 28,342 toponyms from the two datasets.

We compared our approach with 10 representative methods. These include a Voting system [7], CamCoder [14], CHF [15], Clavin[5], Blink [16], GENRE [17], Bootleg [18], and the three standard geocoders: Nominatim, GeoNames, and ArcGIS. Among these, CamCoder is a deep learning-based geoparser; CHF and Clavin are rule-based; and Blink, GENRE, and Bootleg are deep learning-based entity linkers. The Voting system integrates seven methods, such as GENRE, Blink, and CamCoder.

Table 1
Summary of the two test datasets. KB is the abbreviation of Knowledge Base.

Name	Text/Tweet Count	Toponym Count	Type	KB/Gazetteer
GeoCorpora[19]	6,648	3,100	Tweet	GeoNames
WikToR[20]	5,000	25,242	Wiki article	Wikipedia

3.2. Experimental results

The outcomes of our evaluation are presented in Table 2. The results show that our approach outperforms others. On average, it exceeds the performance of the previously best method, the voting system, by 7% in *Accuracy@161km* and 61% in *ME*. Compared to the top individual method, GENRE, our approach demonstrates more substantial improvements of 13% in *Accuracy@161km* and 83% in *ME*. These findings underscore the effectiveness of our proposed approach.

4. Conclusion

This research presents an innovative method for toponym resolution utilizing mid-sized, open-source large language models, specifically Llama2 (7B). Its efficiency is validated through testing

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

Table 2
Evaluation results on GeoCorpora and WikToR. Bold numbers indicate the best scores and the second best scores are underlined.

	GeoCorpora		WikToR	
	Accuracy@161km	ME (km)	Accuracy@161km	ME (km)
CamCoder	0.72	3506	0.67	501
CHF	0.75	2985	0.44	1264
Nominatim	0.74	1731	0.21	3894
GeoNames	0.71	3683	0.22	4179
ArcGIS	0.77	1224	0.24	3884
Clavin	0.77	2777	0.22	4171
Blink	0.75	1577	0.68	1217
GENRE	0.79	684	0.88	1006
Bootleg	0.69	4425	0.7	1483
Voting	<u>0.84</u>	<u>460</u>	<u>0.91</u>	<u>273</u>
Llama2 (7B)	**0.9**	**247**	**0.98**	**37**

on two public datasets, establishing a new standard in the field. Furthermore, it maintains significant computational efficiency with a reasonable GPU memory requirement of 14 GB. Future research will aim to investigate a broader range of open-source LLMs for this task and conduct extensive comparative analyses with existing methods across a more diverse array of test datasets. Furthermore, efforts will be directed towards augmenting the models' geographical knowledge during the inference process by incorporating a toponym' candidates retrieved from gazetteers, aiming to enhance accuracy and performance further.

Declaration of generative AI in manuscript preparation

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

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] X. Hu, H. Al-Olimat, J. Kersten, M. Wiegmann, F. Klan, Y. Sun, H. Fan, Gazpne: Annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and

synthetic data by rules, International Journal of Geographical Information Science (2021) 1–28. doi:10.1080/13658816.2021.1947507.

[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 (2022) 1–1. doi:10.1109/JIOT.2022.3150967.

[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] 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.

[8] G. Mai, C. Cundy, K. Choi, Y. Hu, N. Lao, S. Ermon, Towards a foundation model for geospatial artificial intelligence (vision paper), in: Proceedings of the 30th International Conference on Advances in Geographic Information Systems, 2022, pp. 1–4.

[9] 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.

[10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[11] 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).

[12] 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.

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

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

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

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

[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] L. Orr, M. Leszczynski, S. Arora, S. Wu, N. Guha, X. Ling, C. Re, Bootleg: Chasing the tail with self-supervised named entity disambiguation, arXiv preprint arXiv:2010.10363 (2020).

[19] 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.

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