Extracting and Aggregating Hierarchical Toponyms in Abstracts of Scientific Articles in Urban Studies

Tianye Ren[1,*], Nan Bai[1,*] and Ana Pereira Roders[1]

[1]*Department of Architectural Engineering and Technology, Faculty of Architecture and the Built Environment, Delft University of Technology, Delft 2628BL, the Netherlands*

Abstract

This study introduces a literature review tool for abstract-screening in urban studies. It presents two pipelines that focus on linking geoparsing outputs and hierarchical toponym aggregation. *Pipeline1* uses batched matching with GeoNames for fast but coarse aggregation. *Pipeline2* enhances the accuracy by incorporating a toponym resolution model to address geo/geo ambiguity, and semantic checks to correct potential resolution errors. Evaluated on 500 abstracts, *Pipeline2* achieves a precision of 0.96 and a recall of 0.98 for aggregated toponym output.

Keywords

hierarchical toponym aggregation, GeoNames, scientific articles, urban studies

1. Introduction

In urban studies, the research scale is a critical consideration. Bibliometric analyses show that while city-level and smaller-scale studies dominate the literature, comparative research across cities, countries, and continents remains limited [1, 2]. However, there is a need to examine urban phenomena from a broader perspective, as the scarcity of cross-scale and multi-site comparisons limits the generalizability and scalability of findings [3]. A tool that processes toponyms in abstracts is therefore needed to support the systematic review of large datasets of publications.

Geoparsing from text, which includes toponym recognition and resolution, has often focused on news and social media content [4, 5, 6]. Although some work has addressed scientific articles [7, 8], few have targeted the urban studies literature. Hierarchy aggregation is essential for screening multi-region urban studies, where toponyms may appear dispersedly that rely on readers' knowledge to infer their relationships[1]. Toponym resolution, which is applied after recognition to identify toponyms' geospatial representation, provides valuable references for aggregation. Its outputs typically consist of linking a toponym to a gazetteer entry [4, 8]. However, common gazetteers often contain multiple entries for one location, yet not all entries include complete retrieval keywords or hierarchical information[2]. Therefore, additional work is required to link geoparsing outputs and hierarchy aggregation. This study proposes two pipelines, each offering advantages in runtime or accuracy, thereby providing a benchmark for relevant applications.

2. Method

This study focuses on toponyms at the city level[3] or above. A hybrid approach integrating rule-based cleaning, gazetteer matching, and statistical learning is employed to compensate for their respective

GeoExT 2026: Fourth International Workshop on Geographic Information Extraction from Texts at ECIR 2026, April 2, 2026, Delft, The Netherlands

*Corresponding author.

T.Ren@tudelft.nl (T. Ren); N.Bai@tudelft.nl (N. Bai); A.R.Pereira-Roders@tudelft.nl (A. P. Roders)

[1]For example, 5 toponyms "*Delft*", "*The Netherlands*", "*Chengdu*", "*Sichuan*", "*China*" refers to only 2 distinct sites, i.e., "*Delft, The Netherlands*" and "*Chengdu, Sichuan, China*", in a hierarchical structure.

[2]For example, "*Medan, Indonesia*" corresponds to two entries in GeoNames: "*Kota Medan* (id=1214519)" and "*Medan* (id=1214520)". The former, as the more formal one, does not include "*Medan*" in its name variant list. And the latter is not labelled as an administrative division and thus does not appear in most of its children's hierarchy trees.

[3]Third-order administrative divisions are treated as city level for broader inclusion.

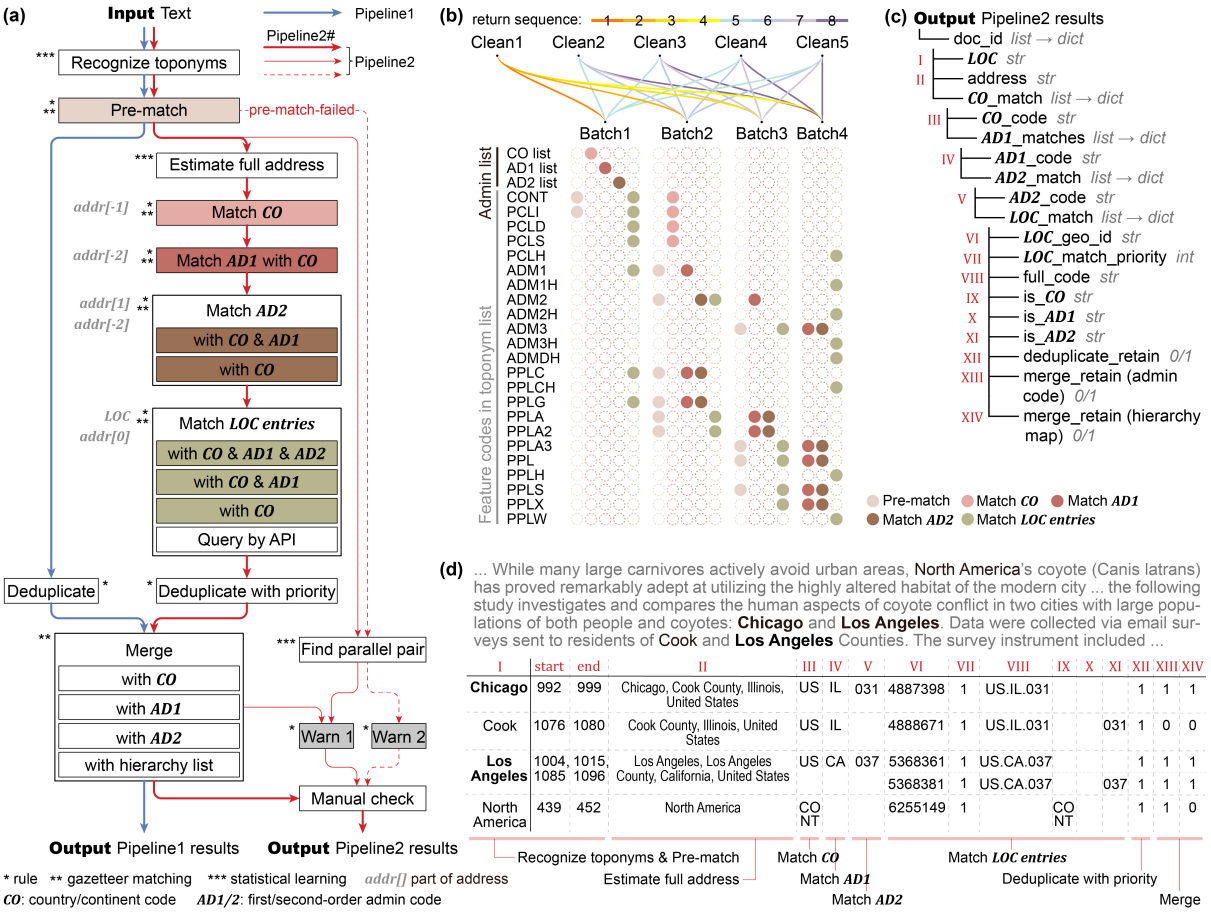

Figure 1: (a) Two pipelines of hierarchical aggregation **(b)** Priority in five different matching stages **(c)** Output dictionary structure of *Pipeline2* **(d)** Example of procesing by *Pipeline2* with final results in black and bold

shortcomings [4, 9]. GeoNames, a commonly used gazetteer [10, 11, 12], is applied. To avoid reaching the daily API limit, a downloaded copy is used[4], including 253 toponym lists grouped by country, along with three admin lists (country, first-order admin, second-order admin), and the hierarchy list.

As shown in **Figure 1(a)**, both pipelines begin with *Recognizing toponyms*. Flair NER (Ontonotes)[5] was selected due to its best performance for admin units among formal-text-oriented approaches evaluated by Hu et al. [4]. Rules and gazetteer are subsequently applied in the *Pre-matching*. This step consists of five cleaning rules[6] and three batches of feature codes[7]. As shown in **Figure 1(b)**, each color of curves between *Clean* and *Batch* represents a result-return. The batches follow a heuristic that prioritizes toponyms with higher administrative attributes. For example, *Batch1* first targets CONT (continent) and PCLI (independent political entity), as they are less likely to be ambiguous. The *Pre-matching* collects, for each toponym, all GeoNames entries that match the highest-priority batch.

Pipeline1 directly uses the *Pre-matching* results, performing simple *Deduplication* before *Merging*. For all toponyms in an abstract and their pre-matched entries, the *Merging* step identifies hierarchical relationships according to the codes of country/continent (CO), first-order admin (AD1), and second-order admin (AD2) consecutively, finally with the hierarchy list serving as a fallback.

Pipeline2 performs additional toponym resolution to estimate the full address for each toponym.

[4]https://download.geonames.org/export/dump/, using only data with feature classes A and P
[5]https://huggingface.co/flair/ner-english-ontonotes-large
[6]*Clean1*: split phrases, *Clean2*: remove articles and possessives, *Clean3*: remove prefixes/suffixes, *Clean4*: remove directional terms, *Clean5*: remove punctuation and spaces. Among them, *Clean3* and *Clean4* are informed by regular expressions and context-free grammars in existing rule-based approaches [4].
[7]http://www.geonames.org/export/codes.html

This study adopts the LoRA-fine-tuned Llama2 (7B) model[8], which demonstrated the best performance for admin units on formal text [13]. Specific parts of the full address are used to match admin codes and LOC entries in four consecutive matching stages. For example, the last part of the address is used to match the country/continent (CO) code only. Each matching stage is initially constrained by all previously matched higher-level code(s). If nothing could be matched, codes are progressively reduced to ease the constraints, until only the CO code remains. For LOC matching, the GeoNames API finally performs fuzzy search as a fallback. LOC entries matched under more admin codes' constraints are prioritized for retention during *Deduplication with priority*. Similar to *Pre-matching*, every matching stage includes cleaning address parts and matching prioritized batches. Batch usage varies by stage, as shown in **Figure 1(b)**, with *Batch4* used only in the three later stages where greater fineness is needed.

In addition to identifying hierarchical relationships based on geographic knowledge during *Merging*, *Pipeline2* incorporates a *Manual check* prompt based on semantic analysis. Parallel toponym pairs (including toponyms failed during *Pre-matching*) are extracted by identifying coordination structures in dependency parse, i.e., the head element and other element(s) attached via coordinating conjunctions (e.g., "and", "or"). If a parallel pair extracted in this step also appears in the *Merging* step, it suggests a possible error and is flagged as *Warn1*. If the pair consists of one successful and one failed toponym during *Pre-matching*, it indicates that the former may incorrectly be retained due to geo/geo ambiguity; such cases are flagged as *Warn2*. These warnings are then manually reviewed. The final output is a dictionary of toponyms with the smallest granularity (TSG). The output structure is shown in **Figure 1(c)**. An example of processing an abstract [14] is provided in **Figure 1(d)**.

Given the demonstrated effectiveness of Large Language Models in information retrieval tasks of literature review [15, 16, 17], this study also employs GPT-4.1 (prompt provided in Appendix A) for comparison. The test sample, 500 abstracts that each mentions at least two toponyms, was randomly selected from a systematic literature review of urban study from Scopus in December, 2025. After manual annotation, 787 annotated TSG serve as the gold standard (GS). The accuracy performance of the two pipelines and GPT-4.1 aggregator is evaluated by: precision $= \frac{t}{t+f}$, recall $= \frac{t}{GS}$, and F1 $= \frac{2t}{GS+t+f}$; where t and f refer to the number of correct and incorrect TSG outputs.

3. Result

The distribution of recognized toponym (RT) number and GS number cross the 500 abstracts is shown in **Figure 2(a)**. Among the RTs, 474, 314, and 283 toponyms were recognized as CO, AD1, and AD2, respectively. The stages preceding *Manual checking* in *Pipeline2* are denoted as *Pipeline2#*. The accuracy performance of *Pipeline1*, *Pipeline2#*, and GPT-4.1 are presented in **Figure 2(b)**. *Pipeline1* takes 26 minutes, and *Pipeline2#* takes 563 minutes. As shown in **Figure 2(c)**, the performance varies with the number of RT per abstract, so differences are observed between micro- and macro-averaged results. Overall, *Pipeline2#* outperforms *Pipeline1*, achieving micro-average precision and recall of 0.95, since *Estimating full address* (toponym resolution) in *Pipeline2#* avoids most geo/geo ambiguities. For example, as shown in **Table 1(a)**, while *Pipeline1* collects "*Washington County*" entries under both states of "*Idaho*" and "*Oregon*", *Pipeline2#* correctly identifies "*Washington*" as "*Washington State*".

However, *Pipeline2#* still makes mistakes. As shown in **Table 1(b)**, "*Maldonado*" was incorrectly estimated as a first-order administrative division, but *Warn1* captures that it is parallel with a second-order administrative division, thereby providing evidence for correction. In **Table 1(c)**, "*Ottawa-Gatineau*" had been excluded during *Pre-matching* but was later captured by *Warn2* as it is parallel with other eligible toponyms. *Warn1* and *Warn2* respectively flagged 13 and 11 abstracts. The warnings did not necessarily mean problems, so *Manual Checking* is needed and 17 of them required correction. With these corrections, the complete *Pipeline2* achieved micro-average precision and recall of 0.96 and 0.98.

Finally, albeit with the toponym resolution and manual checking during *Warn1&2*, some geo/geo ambiguity remain unresolved. For example, in **Table 1(d)**, "*Tohoku*" refers to a cultural region, yet it was matched to a formal administrative division with an identical name and a much lower hierarchy.

[8]https://huggingface.co/xukehu/Llama2-7B-LoRA-Toponym-Resolution

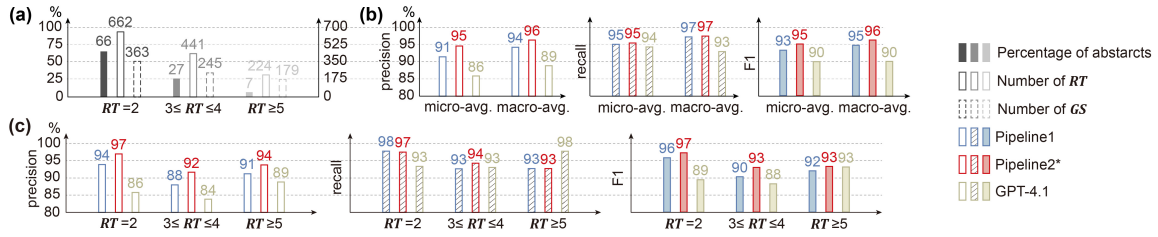

Figure 2: (a) Performance on full sample **(b)** Statistics of RT buckets **(c)** Performance on RT buckets

Table 1
Examples of three error types: merging with geo/geo ambiguity, missing toponyms, and redundant output

	Text with correct RTs in bold and errors in red	Pipeline errors: toponym[geo_id]
(a)	Rapid urbanization in high rainfall areas of western **Washington**, western **Oregon** and northern **Idaho** has increased the potential for flooding.	*Pipeline1* merges: Washington[5611570] ∈ Idaho[5596512] Washington[5759333] ∈ Oregon[5744337]
(b)	This study uses data from a sample of 420 residents from **Maldonado** and **Punta del Este**.	(*Warn1*) *Pipeline2#* merges: Punta del Este[3440939] ∈ Maldonado[3441890]
(c)	We observed spatial inequalities in population mental health across and within **Vancouver**, **Calgary**, **Edmonton**, **Toronto**, and **Ottawa-Gatineau**, which are...	(*Warn2*) *Pipeline2#* misses: Ottawa-Gatineau[6094817][5959974]
(d)	...human and material damage mainly in three Tohoku prefectures: **Iwate**, **Miyagi**, and **Fukushima**.	*Pipeline2#* gives redundant output: Tohoku[11010639]

4. Discussion

The proposed pipelines help extract toponyms at city-level and above, and aggregate them to determine the study scales. While this study uses abstracts of scientific articles in urban studies for testing, the pipelines can easily be adapted to other domains and other text forms, such as the nomination files of World Heritage or news articles about tourist sites. While both pipelines employs batched matching heuristic, the bias toward prominent places is mitigated in *Pipeline2* because the address parts themselves reflects a hierarchical order, making top-down batched matching reasonable. GPT-4.1 underperforms *Pipeline1&2#*, likely due to its lack of comprehensive geographical knowledge [18]. However, it exhibits high recall when RT≥5 in **Figure 2(c)**. This may be attributed to richer semantic context from more RTs, where GPT-4.1 excels. In *Pipeline2*, *Warn1&2* partially compensates for this capability. They flag a small subset of abstracts that are most likely to be incorrect for manual checking and correction.

Although *Pipeline2* performs best in terms of accuracy, different pipelines can be selected depend on application scenarios. *Pipeline2#* achieves F1 of 0.97 for abstracts with RT=2, as shown in **Figure 2(c)**. Such abstracts are more likely to only reference one location yet be misclassified as multi-regional before hierarchy aggregation. If the goal is solely to determine whether an abstract is multi-regional, the automatic *Pipeline2#* can be sufficient. Besides, *Pipeline2#* requires much more runtime than *Pipeline1* due to the computational cost of the resolution model. When marginal benefit of time saving is high, *Pipeline1* may be a practical choice. However, whether the difference between *Pipeline1* and *Pipeline2#* is robust across a broader range of samples remains to be validated.

Future research could advance GeoAI that jointly integrate textual data and geospatial knowledge graphs. This integration would facilitate the reconstruction of hierarchically structured entities, potentially through spatial reasoning [18, 19]. Moreover, because the annotated corpus of Flair NER (Ontonotes) only includes specific named entities [20], general phrases such as "*Australian cities*" are often ignored. Subsequent work could construct lexicons of geographic adjectives and place-type nouns, extract their modification relations, and leverage this pattern to improve hierarchy aggregation.

5. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Q. C. Doan, X. Zhang, A systematic review of urban vitality studies: Trends and research opportunities, Land Use Policy 158 (2025). doi:10.1016/j.landusepol.2025.107745.

[2] Z. Li, J. Dong, Big geospatial data and data-driven methods for urban dengue risk forecasting: A review, 2022. doi:10.3390/rs14195052.

[3] J. Luo, P. Liu, X. Kong, J. Shen, Q. Wu, D. Xu, Urban digital twins for citizen-centric planning: A systematic review of built environment perception and public participation, International Journal of Applied Earth Observation and Geoinformation 143 (2025) 104746. doi:10.1016/j.jag.2025.104746.

[4] 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 (2024) 1–37. doi:10.1145/3625819.

[5] M. Mironov, A. Marquard, D. Racek, C. Heumann, P. W. Thurner, M. Aßenmacher, A geoparsing pipeline for multilingual social media posts from ukraine, CEUR-WS, 2025, pp. 6–17.

[6] P. Smith, E. Manley, M. Gould, Assessing performance in extracting topological, direction and distance spatial relations from reddit using llms, CEUR-WS, 2025, pp. 38–45.

[7] E. Acheson, R. S. Purves, Extracting and modeling geographic information from scientific articles, PLOS ONE 16 (2021) e0244918. doi:10.1371/journal.pone.0244918.

[8] D. Weissenbacher, A. Magge, K. O'Connor, M. Scotch, G. Gonzalez-Hernandez, Semeval-2019 task 12: Toponym resolution in scientific papers, Association for Computational Linguistics, 2019, pp. 907–916. doi:10.18653/v1/S19-2155.

[9] M. J. Silva, B. Martins, M. Chaves, A. P. Afonso, N. Cardoso, Adding geographic scopes to web resources, Computers, Environment and Urban Systems 30 (2006) 378–399. doi:10.1016/j.compenvurbsys.2005.08.003.

[10] S. Malmasi, M. Dras, Location Mention Detection in Tweets and Microblogs, 2016, pp. 123–134. doi:10.1007/978-981-10-0515-2_9.

[11] R. Dutt, K. Hiware, A. Ghosh, R. Bhaskaran, Savitr: A system for real-time location extraction from microblogs during emergencies, ACM Press, 2018, pp. 1643–1649. doi:10.1145/3184558.3191623.

[12] N. J. F. Martínez, C. P. Pascual, Knowledge-based rules for the extraction of complex, fine-grained locative references from tweets, Revista Electronica de Linguistica Aplicada 19 (2020) 136–163.

[13] 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. doi:10.1080/13658816.2024.2405182.

[14] E. E. Elliot, S. Vallance, L. E. Molles, Coexisting with coyotes (canis latrans) in an urban environment, Urban Ecosystems 19 (2016) 1335–1350. doi:10.1007/s11252-016-0544-2.

[15] K. Ito, Y. Kang, Y. Zhang, F. Zhang, F. Biljecki, Understanding urban perception with visual data: A systematic review, Cities 152 (2024). doi:10.1016/j.cities.2024.105169.

[16] gkamradt, Llmtest_needleinahaystack, 2024. URL: https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

[17] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain, Structured information extraction from scientific text with large language models, Nature Communications 15 (2024) 1418. doi:10.1038/s41467-024-45563-x.

[18] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu, C. Cundy, Z. Li, R. Zhu, N. Lao, On the opportunities and challenges of foundation models for geoai (vision paper), ACM Transactions on Spatial Algorithms and Systems 10 (2024) 1–46. doi:10.1145/3653070.

[19] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, N. A. Smith, Knowledge enhanced contextual word representations, Association for Computational Linguistics, 2019, pp. 43–54. doi:10.18653/v1/D19-1005.
[20] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, M. El-Bachouti, R. Belvin, A. Houston, Ontonotes release 5.0, 2012. URL: https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf.

A. Supplementary material

Prompt for GPT-4.1 aggregator is available via

- Prompt