Disambiguating Geographic Names in Biodiversity Occurrence Data: A Retrieval-Augmented Generation Approach

Yanni Jose Ella[*◇], **Monica Ashley Laviste**[*◇], **John Michael Lastimoso**[‡], **Wilfred John Santiañez**[‡], **Riza Batista-Navarro**[†], **Roselyn Gabud**[*†]

[*] Dept. of Computer Science, College of Engineering, University of the Philippines Diliman

[‡] Gregorio T. Velasquez Phycological Herbarium and The Marine Science Institute, College of Science, University of the Philippines Diliman

[†] Dept. of Computer Science, University of Manchester

{ycella, mrlaviste, rsgabud}@up.edu.ph, riza.batista@manchester.ac.uk, {jmlastimoso, wjsantianez}@msi.up.edu.ph

Abstract

The availability of georeferenced coordinates is essential for biodiversity research, as it enables species distribution modeling and supports conservation planning. However, datasets often contain ambiguous or inconsistent geographic names that reduce spatial accuracy and underscore the need for methods that resolve geographic name ambiguity. While traditional named entity linking strategies are well established, they remain limited in low-resource domains, e.g., in biodiversity contexts, due to the scarcity of annotated training data and high lexical ambiguity of local geographic names. This study proposes a Retrieval-Augmented Generation (RAG) framework to automatically disambiguate Philippine seaweed-related geographic names in databases and literature. This approach utilizes a custom knowledge base of gazetteers to support large language models (LLMs) in the task of geospatial disambiguation. With a disambiguation accuracy of 87.8% within a 5 km distance error threshold, our evaluation shows that the RAG-enabled pipeline significantly outperforms standard LLM baselines (Accuracy@5km = 0%), demonstrating the need for external knowledge to resolve geospatial ambiguity.

Keywords: retrieval-augmented generation (RAG), entity linking, large language models (LLMs), chain-of-thought (CoT), biodiversity, seaweeds

1. Introduction

Biodiversity occurrence data consists of records that capture the presence of a particular species at a defined location and point in time. The availability of geographic coordinates in these records enables species distribution mapping and habitat modeling, providing a foundational basis for biodiversity research and environmental policymaking (Chapman and Wieczorek, 2020). Despite its importance, a substantial proportion of biodiversity occurrence databases, e.g. GBIF[1], and ALA[2], include textual locality descriptions but lack usable geographic coordinates, particularly among historical museum and herbarium specimens. Several of these occurrence data entries contain ambiguous, outdated, or inconsistently recorded geographic names due to legacy place names, overlapping administrative units, and local vernacular name variations. As a result, data integration across biodiversity repositories is hindered, limiting the overall utility of existing datasets for research and policymaking.

This issue is particularly significant for marine macroalgae (seaweeds), especially in the Philippines, which has more than 1,065 documented seaweed taxa, making it the most diverse in terms of seaweed resources in the western Pacific (Lastimoso and Santiañez, 2021). Since seaweeds have historically received less scientific attention than many other marine taxa (Arceo et al., 2024), accurate georeferencing is essential to providing marine scientists with reliable information on species occurrence and distribution. This enables a clearer understanding of seaweed ecology and its environmental interactions. It is especially critical for monitoring invasive species, detecting local extinctions, and assessing habitat changes driven by both local pressures (e.g., ocean acidification) and global threats such as climate change. Furthermore, precise georeferencing supports the documentation of ecological and phenological patterns through long-term monitoring, generating essential evidence for conservation planning, resource management, and sustainable use of marine resources.

Current practice resolves these ambiguities by manually interpreting textual descriptions in the literature and consulting maps or gazetteers to infer the most plausible geographic location. For example, the phrase "*Ceramiales* specimen collected at Magsaysay, Pangasinan", contains the location mention "Magsaysay", a common place name found across multiple provinces and municipalities in the Philippines. Using contextual cues from literature describing seaweed collection sites in Pangasinan, one can infer that it was collected

[◇] The authors contributed equally to this work.

[1] https://www.gbif.org/

[2] https://www.ala.org.au/

Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology) @ LREC 2026, pages 42–52

in “Magsaysay Island, Bolinao, Pangasinan” and retrieve its coordinates from a gazetteer. Although effective, this is time-consuming and has become increasingly unsustainable as biodiversity data continues to grow. This underscores the urgent need for automated approaches that accurately and consistently interpret geographic references.

Traditional automated georeferencing methods rely on gazetteer lookups, rule-based parsing, or supervised named entity recognition (NER) models, which struggle with ambiguous or evolving place names and require large, domain-specific annotated datasets (Peeters et al., 2024; Marcer et al., 2021). Creating such datasets is impractical, particularly in low-resource domains like biodiversity. Moreover, effective disambiguation often depends on external knowledge not present in the text (Overell, 2011). There is thus a need for scalable approaches that can integrate new information without relying on large amounts of labeled data.

Entity linking is a natural language processing (NLP) task that maps textual mentions to the corresponding entities in a knowledge base (KB). In this study, we formulate the georeferencing problem as an entity linking task, where each location mention is systematically mapped to its corresponding real-world geographic entity in a KB, e.g., GeoNames[3]. Let G denote a gazetteer KB containing a finite set of geographic entities. Each entity, $g \in G$ has associated metadata (i.e., a place name, alternative names, a feature type, administrative hierarchy, geographic coordinates). Let M denote a set of geographic mentions, where each mention $m \in M$ is a text span consisting of one or more tokens referring to a real-world geographic location. The proposed entity linking system defines a mapping function $f : M \rightarrow G$.

Recent advances in NLP, particularly large language models (LLMs) combined with retrieval-augmented generation (RAG) and in-context learning, offer a promising approach to this challenge. Recent studies such as GNEMM (Zhang et al., 2025a) show that integrating retrieval with LLMs can enhance entity linking performance without reliance on large annotated datasets. However, applying such methods to biodiversity georeferencing remains underexplored.

We propose a RAG-based approach, formulated as an entity linking task, where the system retrieves candidate entities from a gazetteer and supporting passages from literature, then uses LLM reasoning with in-context learning to disambiguate mentions and select the most plausible geographic entity. We evaluate the approach on Philippine seaweed occurrence records, where ambiguity is widespread due to shared place names across administrative levels, inconsistent spelling, and legacy toponyms.

[3] https://www.geonames.org/

2. Related Work

Various tools have been developed to semi-automate the assignment of geographic coordinates to textual location data. Tools like BioGeomancer (Guralnick et al., 2006), GeoLocate (Wieczorek et al., 2004) and BELS[4] (Marcer et al., 2021) primarily depend on gazetteer lookups, string matching, and rule-based parsing, often struggling with incomplete and ambiguous place names due to the lack of context (Gritta et al., 2017).

Georeferencing has also been framed as an entity linking task composed of (1) candidate retrieval from a dictionary and (2) context-based disambiguation using neural classifiers (Kolitsas et al., 2018). Transformer-based architectures, such as ReFinED (Ayoola et al., 2022), improve contextualized representations and demonstrate strong performance on general-purpose corpora or large annotated datasets. However, they depend heavily on supervised training, with performance degrading when transferred to specialized or low-resource domains (Soliman et al., 2022).

Applying entity linking to biodiversity-related geographic names introduces domain-specific challenges. Location names are inherently ambiguous and require contextual reasoning (Overell, 2011). Moreover, biodiversity records often contain fine-grained descriptions, historical place names, and landmarks (Chapman and Wieczorek, 2020). While heuristics such as population-based ranking can improve traditional georeferencing systems for well-known locations, their effectiveness diminishes when applied to local, regional, or highly ambiguous place names. As such, supervised entity linking systems trained on general-purpose corpora may not generalize well to biodiversity datasets.

Recent advances in LLMs offer an alternative via in-context learning, where pre-trained models generalize to new tasks from instructions or a small number of examples (few-shot prompting) without parameter updates. Peeters et al. (2024) demonstrate that prompt-based LLMs achieve competitive entity linking performance without extensive fine-tuning, making them viable in domains with limited labeled data. However, standard LLMs rely solely on parametric memory, making them prone to hallucination when resolving entities that are not included in their pre-training corpus (i.e., out-of-distribution entities).

RAG addresses this by combining a pre-trained LLM with a non-parametric external knowledge source (Lewis et al., 2020). Retrieving relevant candidates before generation grounds the LLM reasoning in domain-specific knowledge, mitigating hallucination. A study by Zhang et al. (2025a) inte-

[4] https://github.com/calacademy-research/bels_dockerized

grates RAG with a geographic named entity matching framework (GNEMM) that uses LLM reasoning to rank candidates by spatial and semantic similarity, though this was evaluated only on Chinese address records, limiting its generalizability.

In this work, we combine RAG with LLM-based disambiguation to address the challenges of georeferencing biodiversity occurrence data. Unlike existing supervised entity linking systems or heuristic geoparsing tools, this approach does not require domain-specific finetuning and uses a semantically enriched gazetteer and literature excerpts as a nonparametric knowledge source to handle ambiguous place mentions in biodiversity-related text.

3. Dataset

To support the development of methods for entity linking-based georeferencing, we utilized a dataset comprised of 6,865 occurrence data entries pertaining to various seaweed taxa located primarily in the provinces of Batangas and Pangasinan, Philippines. This data was collected by a group of researchers specialising in Philippine seaweeds. Each data row includes the herbarium code, administrative hierarchy, collection date, and taxonomic classification.

There are 281 unique location names in the seaweed occurrence dataset. The ground truth coordinates were established for 248 location names through a manual annotation process conducted independently by two (2) annotators. One annotator is a domain expert, a postgraduate researcher with field experience in marine science, including firsthand knowledge of the collection sites (senior annotator). The other is an undergraduate student of Computer Science (junior annotator). The annotators assigned coordinates based solely on the location name and its administrative context without considering taxon or collection date. They interpreted unstructured locality descriptions from literature, cross-referencing Google Maps to resolve place names. In addition, 86 of the locations were already documented in the laboratory's internal geographic database of collection sites. This was accessible to the senior annotator, who had consulted it during annotation. The annotators assigned decimal degree coordinates using the World Geodetic System (WGS84) geodetic datum. Table 1 presents the number of unique location names and their corresponding occurrence records, distinguishing between those that were double-annotated and those annotated only by the senior or junior annotators.

The dataset contains location names with highly variable spatial extents. Some refer to "barangays", the smallest administrative units in the Philippines, while others denote islands of varying sizes, or higher-level administrative divisions such as towns.

Table 1: Number of location name annotations and the corresponding number of occurrence records.

Annotator	Nbr. of Unique Location Names	Nbr. of Occur. Records
Sr Annotator	123	2654
Jr Annotator	22	506
Both Annotators	103	3040
Not Annotated	33	665
Total	281	6865

Since the dataset contains occurrences of marine (seaweed) species, several location mentions were assigned coordinates near or along coastlines by the senior annotator, drawing on prior knowledge of the collection sites. For example, the location mention "Santiago Island, Pangasinan" may be annotated by one annotator with coordinates on the central part of the island, while another may assign coordinates on the northern part. Although the coordinate pairs differ, both fall within the geographic extent of the same locality, as seen in Figure 1.

To assess the reliability of the annotations, we measured inter-annotator agreement (IAA) on the doubly annotated set composed of 103 unique location names, treating the senior annotator's coordinates as the reference standard. Given the variability in spatial extent, particularly in marine science research, coordinate agreement cannot always be determined through exact matching. Instead, it is more appropriately evaluated using kilometer-radius matching, with a threshold that may be larger than what is typically applied in georeferencing residential addresses. We applied a kilometer-radius matching criterion of 1, 3, and 5 km, where the annotators were considered to be in agreement if the distance between the coordinates they provided fell within the specified radius. The IAA, measured in terms of accuracy, is 50.98%, 79.41%, and 93.14% for the 1-, 3-, and 5-km radius, respectively, indicating a high level of consistency between annotators and demonstrating annotation reliability.

Because a single location name may correspond to multiple occurrence records, each unique location name was annotated only once, and the resulting coordinates were applied to all records sharing that location name. After removing entries without ground truth coordinates, i.e., entries whose location names could not be resolved to coordinates either by the laboratory's existing database or manual annotation, 6,186 records remained. This dataset was partitioned into training, development, and test splits using an 80:10:10 ratio. The development set was used for LLM parameter tuning and retrieval configuration experiments, while the test set was strictly reserved for final evaluation.

Figure 1: 5-km distance between the coordinates of the collection site “Station 1, Santiago Island” and of “Santiago Island” as seen in Google Maps.

4. Methods

In this section, we present our methods for linking a location mention to its corresponding real-world geographic entity. The input to the pipeline is a *query text*, q, formulated as a natural language representation of an occurrence record that includes the administrative hierarchy, collection date, and taxonomic classification. For example,

> “*Bryopsidales Caulerpaceae* (*Chlorophyta, exsiccatae*; Apr 21, 1983) collected from Long Beach, Bolinao, Pangasinan, Philippines on 21 April 1983. Specimen associated with herbarium code MSI13509, recorded as *Caulerpa racemosa*.”

The overall methodology has two phases. The first phase is the construction of the KB by embedding the gazetteer and seaweed literature corpus into a vector database. Figure 2 shows this process which is done only once. The second phase is the per-query pipeline, as shown on Figure 3, which comprises (1) the two-stage retrieval that takes q as input and retrieves gazetteer data and literature excerpts relevant to q, referred to as *candidates*, which are then reranked and passed to the LLM; and (2) LLM-based resolution that utilizes in-context Learning and Chain-of-Thought prompting. Here, the LLM evaluates the *candidates* against the provided context to resolve the location mention to specific gazetteer coordinates.

4.1. Gazetteer Knowledge Base

We constructed a Philippine gazetteer KB by obtaining country-specific data from GeoNames to enable *candidate* retrieval and LLM grounding. It uses decimal degrees and WGS84 as the coordinate system and geodetic datum, similar to the format of the coordinates in our dataset. It contains 96,643 Philippine geographic entities, each associated with the corresponding coordinates, administrative classifications, and feature codes. The administrative units were standardized to reflect country-specific terminology (e.g., `ADM1` mapped to Region, `ADM2` to Province).

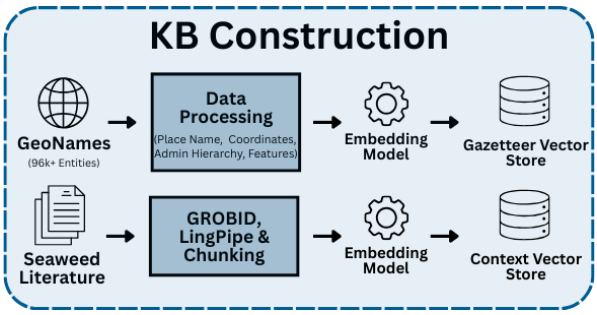

Figure 2: Knowledge Base Construction Process

Each GeoNames entry was transformed into a canonical textual description capturing the place name, feature type, administrative hierarchy, and alternate names (when available). By converting structured data into unstructured text, we are able to capture the relational context between the place name and its geographic features, allowing the vector database to represent semantic meanings. This representation constitutes a single entry in the gazetteer KB and enables the retriever to match not only on string similarity but also on the semantic properties of geographic features. For example, it facilitates disambiguation of “Pasig” as either Pasig River (a hydrographic feature) or Pasig City (an administrative unit). Figure 4 shows an example textual description formed based on GeoNames attributes.

4.2. Context Knowledge Base

We constructed a context KB based on 24 scholarly articles authored by members of a research group with expertise in Philippine seaweeds, which contain ecological descriptions, sampling site details, geographic and taxon mentions (e.g *Bryopsidales Caulerpaceae*) that may provide supporting evidence for geographic disambiguation. Text was extracted from the PDF documents using GROBID[5], which produces structured and normalized XML representations. The extracted text was segmented into sentences and grouped into overlapping chunks using a sliding window strategy. We experimented with different chunk sizes and overlap settings to determine the configuration that best balances contextual coverage and embedding quality (see Section 5.2).

4.3. Retrieval

The system generates a dense *embedding of the query text*, $e(q)$, for the retrieval stage. Querying

[5] https://github.com/grobidOrg/grobid

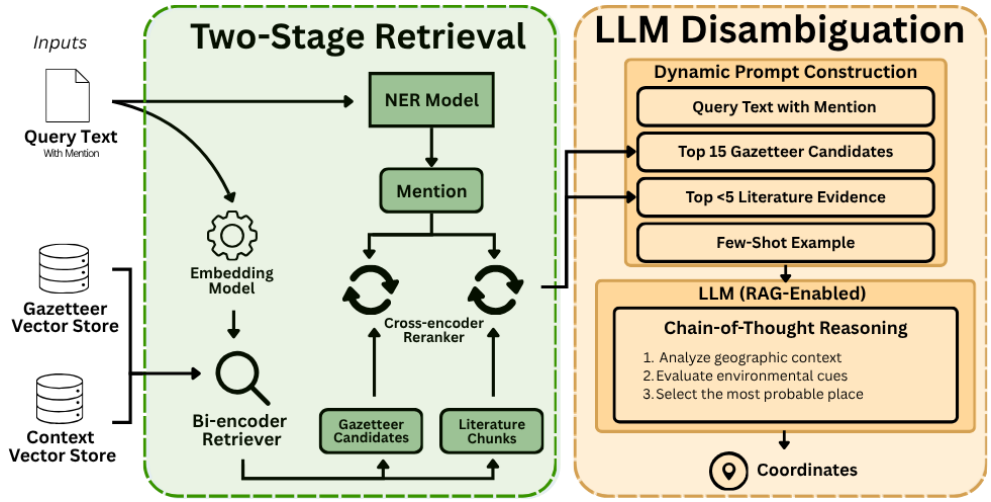

Figure 3: RAG pipeline with retrieval, few-shot prompting, and LLM generation for geographic name disambiguation.

"Trenchera Point is a tapering piece of land projecting into a body of water, less prominent than a cape (T.PT) in Bolinao (Municipality/City), Province of Pangasinan (Province), Ilocos (Region), Philippines. Alternate names: Trenchera Point, Trinchera Point."

Legend: Place Name, Feature Description, Feature Code/Class, admin3, admin2, admin1, Country, Alternate Names

Figure 4: Example concatenated text based on a GeoNames gazetteer entry.

the gazetteer KB, the retriever computes similarity scores between $e(q)$ and each gazetteer entry embedding $e(g)$ using the cosine similarity function $s(a, b)$ and returns $C_G = \text{Top}_n\{\, g \in G \mid s(e(q), e(g))\,\}$, where Top_n returns the $n = 150$ gazetteer entities with the highest similarity scores with $e(q)$. In parallel, C_L, the set of top $k = 15$ semantically relevant chunks from the context KB, is retrieved. Both retrievals use a bi-encoder approach, in which the query and each candidate are independently encoded into dense vectors. Candidate ranking is based on cosine similarity between these vectors which enables the system to identify relevant *candidates* even if the mention does not exactly match the gazetteer entries (e.g., partial names).

To determine the most effective retrieval configuration, we evaluated five (5) dense embedding models: BGE-M3 (Chen et al., 2024), MiniLM (Wang et al., 2020), Qwen3-Embedding-0.6B (Zhang et al., 2025b), Contriever (Izacard et al., 2021) and MPNet (Song et al., 2020). These models were indexed across three (3) vector databases: ChromaDB (Xie et al., 2023), Pinecone (Xie et al., 2023) and FAISS (Johnson et al., 2021).

The bi-encoder retriever encodes queries and gazetteer entries independently, which enables nearest-neighbor search over precomputed embeddings but may not capture fine-grained interactions between the *mention* and each *candidate*. To address this, we applied a cross-encoder reranking stage using *bge-reranker-base*, which jointly encodes the *mention-candidate* pair to produce more discriminative relevance scores (Wu et al., 2020). This enables the reranker to focus on token-level differences in context (e.g., distinguishing "Patar, Bolinao, Pangasinan" from "Patar, Tayug, Pangasinan") that the bi-encoder alone cannot resolve.

Reranking is also applied to the retrieved literature chunks, with a relevance threshold of 45% to filter out low-scoring literature chunks and to reduce the noise passed to the LLM. After reranking and filtering, the top 15 gazetteer *candidates* and top 5 literature chunk *candidates* are retained and passed to the LLM-based disambiguation stage.

4.4. LLM-based Disambiguation

For each geographic mention, a prompt is dynamically constructed containing the *query text*, q, the gazetteer *candidates*, and literature chunk *candidates*. This grounds the LLM's reasoning on both geographic and ecological evidence.

Chain-of-Thought (CoT) Prompting. The system employs a structured CoT strategy that explicitly instructs the LLM to follow a three-step reasoning process before answer generation. The first step (Place Analysis) interprets the mention in context and is characterized solely from the *query text* q. In the next step (Candidate Evaluation), the LLM compares each retrieved *candidate* with the context, weighing administrative-region consistency, name similarity, and environmental consistency. Lastly

(Selection & Validation), the LLM returns the best-fitting *candidate* and outputs coordinates exactly from the gazetteer, preventing data hallucination.

Few-Shot Prompting. To guide the LLM's output structure and reasoning style, the pipeline employs few-shot prompting, where the number of examples (shots) is set to 2. These examples demonstrate the resolution of the location name with its corresponding context and the expected output. This helps in anchoring the LLM's expectations, reduces the likelihood of formatting errors and ensures that the model prioritizes administrative consistency (e.g., matching "Pangasinan" with "Ilocos Region") over superficial name matches. Furthermore, this provides the LLM with a reasoning template for handling potential orthographic and typographical errors in the source literature and place name mentions.

5. Results and Analysis

We conducted experiments to identify an optimal configuration that balances accuracy with computational efficiency, such as runtime and model size, using evaluation metrics from georeferencing and NLP literature. We then report the results based on the manually labeled held-out test set. All experiments were conducted on an NVIDIA A100 Tensor Core GPU, and the reported results represent the average over three independent runs.

5.1. Evaluation Metrics

Due to variability in the extent of the locality, we evaluated the system using accuracy within a distance threshold. The Maximum Uncertainty framework proposed by Thapa and Bossler (1992), which has since been the best practice for georeferencing (Chapman and Wieczorek, 2020), explicitly accounts for uncertainty arising from the spatial extent of a locality (e.g., Santiago Island with a diameter of 7 km). This aligns with the point-radius method, which represents the best estimate point together with an uncertainty radius that bounds a plausible true location (Wieczorek et al., 2004). In addition, we used geodesic distance to measure the distance between two coordinates to account for the shape of the earth with as much precision as possible.

In this study, we report the accuracy (Acc) at multiple radii using thresholds of 1, 3 and 5 km (Acc@k km) to accommodate the varying spatial granularity of locations in our dataset, which include specific sampling sites, barangays (`ADM4`), municipalities, and entire islands.

To complement this, we also compute the Root Mean Squared Distance (RMSD) as a measure of geographic error. This captures the overall magnitude of spatial deviation and penalizes larger errors more heavily. Lower RMSD values indicate predictions that, on average, are closer to the gold standard coordinates (Leidner, 2007). Given the geoid distance $\Delta(.,.)$ between a set of N ground-truth centroids $\vec{y} = (y_1, ..., y_n)$ and chosen location centroids $\vec{d} = (d_1, ..., d_n)$ retrieved from the gazetteer:

$$\text{RMSD}(\vec{d}, \vec{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\Delta(d_i, y_i)]^2}$$

In addition, we report the median (Med) defined as the middle value of the ordered distance errors to represent the typical precision of the system that is resilient to outliers (e.g., $>$100 km distance errors).

Table 2: Impact of chunking configuration (no. of sentences - overlap) on retrieval performance.

Chunk	@1↑	@3↑	@5↑	Med↓	RMSD↓
3s-1o	27.6	54.2	72.8	2.44	116.80
5s-1o	27.8	54.6	73.0	2.23	115.40
5s-2o	26.0	52.8	68.8	2.33	136.50
7s-2o	26.2	52.5	67.5	2.44	126.70
10s-3o	24.5	51.0	68.8	2.31	103.67

5.2. Chunking Configuration Analysis

We evaluated a total of 20 retrieval configurations across 5 embedding models: BGE-M3, Contriever, MiniLM-L6-v2, MPNET-Base-v2, and Qwen3-0.6B, while remaining within their respective token limits. We tested 3, 5, 7 and 10 sentences per chunk with overlaps of 1 to 3 sentences. However, due to token size limitations, errors occurred with all-MiniLM-L6-v2 for the 5-, 7-, and 10-sentence chunks with two or more overlaps (i.e., 5s–2o, 7s–2o, and 10s–3o). Similarly, all-mpnet-base-v2 and facebook/contriever encountered errors under the 10-sentence, 3-overlap (10s–3o) configuration.

Table 2 illustrates the retrieval performance of each unique chunking configuration on the different embedding models averaged together on a fixed LLM backbone (Gemma2:9b) and vector database (FAISS). Although the 10-sentence window yielded the lowest average error at 103.67km, it was only tested on two embedding models because of token limits. The 5-sentence window with 1-sentence overlap achieved the highest accuracy and the second lowest average error at 115.4km. Consequently, we adopted the 5s-1o configuration for the remainder of our experiments to ensure consistent performance and a broader embedding model compatibility.

5.3. Retrieval Configuration Analysis

Table 3 shows the comparative performance of embedding models and vector databases under

Table 3: Retrieval performance of embedding models and vector databases using Gemma2:9b.

DB	Emb	@1↑	@3↑	@5↑	Med↓	RMSD↓
Faiss						
	Bge	24.5	52.4	69.6	2.30	71.00
	Contriever	25.8	24.0	79.7	1.54	49.00
	MiniLM	26.9	51.5	68.7	2.30	185.00
	Qwen	18.3	41.7	60.2	3.08	122.50
	MPNet	21.7	43.3	61.3	3.08	138.89
Pinecone						
	Bge	21.9	39.6	58.6	3.07	118.66
	Contriever	24.2	52.9	69.6	2.37	136.24
	MiniLM	25.2	37.8	55.0	3.49	203.33
	Qwen	18.3	41.1	59.6	3.07	124.58
	MPNet	21.8	30.8	49.0	5.57	180.20
Chroma						
	Bge	22.9	45.9	46.3	5.81	158.48
	Contriever	19.2	57.1	73.2	2.25	149.56
	MiniLM	24.5	51.0	67.5	2.41	198.00
	Qwen	17.4	34.4	52.4	4.56	206.70
	MPNet	18.5	37.8	55.8	3.07	136.79

Table 4: Embedding model with 15 candidates retrieval comparison using FAISS vector database.

Model	@1↑	@3↑	@5↑	Med↓
Bge	38.5	71.6	84.6	1.48
Contr	39.4	95.3	96.6	1.20
MiniLM	36.0	61.0	77.3	1.95
Qwen	22.6	57.5	76.0	2.14
Mpnet	25.6	58.8	66.1	2.14

a fixed Gemma2:9B backbone and the optimal 5s-1o chunking configuration (Section 5.2).

FAISS consistently achieved the lowest RMSD across all configurations, particularly when paired with Contriever (RMSD = 49km). This represents a significant improvement over Pinecone and ChromaDB, where the same model produced substantially higher RMSD ranging from 136.24 and 149.56, respectively.

In addition, Table 4 highlights a performance gap among embedding models when their retrieval performance is evaluated on the *candidate* with the lowest distance error among the 15 retrieved results. At the 5 km threshold, Contriever achieved the highest accuracy at 96.6%, while MPNet reached only 66.1%, a disparity of more than 30%. Interestingly, BGE ranks second to Contriever in terms of performance. This can be attributed to BGE's strong multilingual capabilities, which enable the LLM to deduce orthographic variations in Filipino, such as resolving the mention "Balahibong Manok" to the correct gazetteer entry "Balahibongmanoc" by relating semantic terms to geographic entities. Overall, Contriever paired with FAISS emerged as the superior configuration, reaching a peak Acc@5km of 96.6% with a median (Med) error of 1.20 km.

The superior performance of FAISS is mainly attributed to its exhaustive search capability. Using an IndexFlatL2 mechanism, FAISS performs a brute-force search that guarantees the retrieval of the mathematically closest candidates from the 96,643-point Philippine gazetteer. Meanwhile, the Contriever's success reveals that its unsupervised contrastive pre-training allows it to generalize effectively by focusing on raw semantic structures. This enables the system to match physical descriptions such as "coastal reefs" or "Sargassum beds" between the literature and the gazetteer metadata with high precision.

Table 5: LLM comparison using the optimal FAISS-Contriever retrieval configuration.

Model	@1↑	@3↑	@5↑	Med↓	RMSD↓
Gemma2-9B	25.8	62.5	87.8	1.54	49.30
Qwen2.5-7B	22.6	55.2	75.2	1.95	89.30
Randomblock	20.3	52.6	80.8	2.15	110.96
Llama3-8B	21.9	62.0	86.5	1.77	104.68

5.4. Comparison of LLMs

Using the optimal FAISS-Contriever configuration, we benchmarked different LLMs with similar parameter sizes on resolution accuracy (see Table 5). Gemma2:9b achieved the best overall performance with the highest Acc@1km (25.8%), Acc@3km (65.5%) and Acc@5km (87.8%), lowest median error (1.54km) and RMSD (49.30km). This suggests that Gemma2:9b is the most reliable in selecting the correct candidate from the retrieved set.

5.5. Ablation Study

To quantify the individual contribution of each pipeline component, we conducted an ablation study and reported their performance in Table 6.

First, we implemented a gazetteer-only baseline. This simulates the conventional approach of consulting gazetteers to resolve place names, selecting the top-ranked candidate by exact string matching of the place name, without any LLM reasoning. The configuration achieved an Acc@5 of 52.4% but had large distance errors, resulting in an RMSD of 110.40 km. The high RMSD value highlights the limitations of string matching without contextual reasoning. This is due to the system's inability to distinguish between identical or near-identical place names that refer to geographically distinct entities.

Next, we developed an LLM-only baseline configuration using the best-performing retrieval configuration (FAISS & Contriever) and LLM (Gemma2:9b). This LLM-only configuration that relies entirely on the LLM's parametric knowledge without any retrieval failed completely across all accuracy thresholds (0%). This confirms that parametric knowledge alone is insufficient to resolve place names in low-coverage areas and could lead to LLM hallucinations. The high median distance (67.25%)

Table 6: Ablation study with the best LLM and retrieval configuration. Gaz: Gazetteer; Lit: Literature.

Config	@1↑	@3↑	@5↑	Med↓	RMSD↓
Gaz only	23.5	40.1	52.4	1.94	110.40
LLM	0.0	0.0	0.0	67.25	70.61
+CoT	0.0	0.0	0.0	15.75	68.30
+Gaz	24.0	64.6	81.0	1.88	49.50
+Gaz+Lit	25.8	64.4	87.8	1.54	38.43

and RMSD (70.61km) indicate random or weak region-based predictions.

Using CoT prompting with the base LLM significantly reduced the median error from 67.25 km to 15.75 km, demonstrating that structured reasoning instructions help the model produce more coherent geographical outputs. However, the accuracy remains at 0% for all thresholds, suggesting that reasoning alone cannot compensate for the absence of external geographic knowledge.

Adding gazetteer retrieval to the LLM produces the largest performance gain, increasing Acc@5km from 0% to 81.0%, and reducing RMSD from 68.30km to 49.50km. This confirms that constraining the LLM's output to retrieved candidates is the primary reason for the improvement. The gazetteer serves as a non-parametric memory that prevents hallucinations by forcing selection from valid gazetteer entities.

Finally, incorporating the literature grounding on top of the entire pipeline yielded improvements on the system's performance. Acc@5km increased from 81.0% to 87.8%, the median error decreased from 1.88 km to 1.54 km, and RMSD dropped from 49.50 km to 38.43 km. Although the gains are smaller compared to the gazetteer component, this suggests that the literature grounding provides useful additional context in resolving the mention.

In summary, while LLMs utilize their vast general knowledge to identify and disambiguate place names with 20-50 km distance errors, they struggle with the precision required for specialized domains like country-specific seaweed research.

6. Discussion and Error Analysis

We now examine the system's success and failure cases, as summarized in Table 7.

Success Cases. The system effectively handles spelling and naming variations that may occur in biodiversity literature. For example, the mention "Trinchera", despite the spelling variation, is resolved to "Trenchera Point". The dense retriever successfully identified this candidate despite the lexical mismatch. The LLM selected "Trenchera Point" based on its classification as a coastal hydrographic feature, rather than selecting the administrative entity "Trinchera", which occurs in two municipalities within the same province. This is explicitly stated in the retrieved literature, providing additional context. A similar case is "Balahibong Manok Island", which was correctly resolved to "Balahibongmanoc Island". This shows that dense embedding retrieval captures semantic similarity beyond exact string matching.

The case of "Patar" illustrates the role of the context KB in resolving intra-provincial ambiguity. Multiple gazetteer entries share the name "Patar" within the same province. In this instance, retrieved literature describing seaweed collection activities specifically in Patar, Bolinao, enabled the LLM to distinguish the correct entity from its administratively equivalent counterparts.

Failure Cases. A primary source of error and a key contributor to the RMSD of 49.30 km, is the absence of the correct entity in the gazetteer. When the correct entity is not present in the retrieved *candidates*, the system is constrained to select the closest *candidate* that is either semantically or administratively related. For example, "Maahas Vicinity" was resolved to "Maahas" within the correct province but under a different municipality. More importantly, "Tomasa" was resolved to "Doña Tomasa", which is in an entirely different region. Despite the LLM's own reasoning and flagging this match as "geographically impossible", it still outputs the associated coordinates since it is designed to always select a candidate from the retrieved set. A confidence-based abstention mechanism that allows the system to return no prediction when the retrieval quality is low could address this issue.

Errors can also arise from misleading geographic signals in the retrieved literature chunks, particularly when taxon mentions greatly influence retrieval relevance. Because the context KB consists of papers with broad taxonomic coverage, some papers are retrieved and reranked higher primarily due to the taxon mention. These passages may contain places entirely unrelated to the query but highly relevant to the taxon. An example of this is "Botong" in Batangas. The retrieved literature contains a description of seaweed collection in Panagasinan, which the LLM treated as key geographic evidence. Since no gazetteer candidate located in Pangasinan was retrieved, the LLM attempted to find the geographically closest place to the region. In this case, it resolves it to "Botong" in Visayas, which is farther from Pangasinan. This suggests that the LLM lacks reliable spatial reasoning and is reliant on its parametric knowledge, leading to incorrect inference. Thus, for the literature grounding to be effective, the passages must contain information that is relevant to the query and supports the disambiguation task at hand.

Mention → Gazetteer Match	LLM Reasoning
Trinchera → Trenchera Point (Bolinao) Success: Spelling Variation	Literature mentions seaweed collection and selected this as a coastal landform consistent with this context.
Balahibong Manok Island → Balahibongmanoc Island (Batangas) Success: Name Variation.	Selected based on name similarity and island classification. Literature lacked geographic cues, but parametric knowledge supported the match.
Patar → Patar (Bolinao) Success: Multi-match	Literature confirms Bolinao, Pangasinan as the collection area.
Maahas Vicinity → Maahas (Los Baños) Failure: Out-of-gaz.	Literature mentions Batangas. Selected Maahas in Laguna since it is in the same region, and the correct entity is not in the gazetteer.
Tomasa → Doña Tomasa (Albay) Failure: Out-of-gaz.	Context mentions Bolinao, Pangasinan; candidate is a barangay in Albay (Bicol Region), which is geographically inconsistent.
Botong → Botong (Bohol) Failure: Misleading lit.	Literature mentions Pangasinan. Selected candidate in Bohol, believing it is closest to Bolinao in distance.

Table 7: Success and failure cases of resolution.

7. Conclusions and Future Work

This paper presented a retrieval-augmented entity linking approach for georeferencing biodiversity occurrence records that contain ambiguous place names. Our framework combines a gazetteer, a context KB derived from seaweed literature, dense retrieval with reranking and LLM-based disambiguation using structured prompting. Experiments show that retrieval quality is the main driver of performance. We also note that an LLM with no retrieval support is not viable for georeferencing due to its low accuracy, but adding gazetteer and literature retrieval yields the largest performance jump. By grounding decisions in external knowledge and restricting output to retrieved candidates, the system reduces hallucination and improves geographic accuracy.

We have identified a few potential directions for future work. First is improving candidate coverage and ensuring that the retrieved context is informative and relevant. We suggest incorporating a confidence-based abstention mechanism that can give the model the option to not return a prediction when retrieval quality is low. This approach could reduce forced match errors, especially in cases where the locality is not found in the gazetteer. Experimenting with the relevance threshold to select *candidate* literature chunks from the KB can ensure relevant literature grounding. Second is the use of a domain-specific gazetteer that can help improve the coverage of biodiversity sampling sites, particularly coastlines and marine locations. This will reduce the likelihood of resolving to coordinates that are based on geopolitical boundaries instead of the coastlines or the sea. Our third recommendation is to handle uncertainties based on locality type, wherein uncertainty thresholds vary depending on whether the mention refers to an exact place or a broader region. Finally, performance evaluation can be improved by exploring additional metrics that penalize lexically similar but geographically incompatible matches. Distance-based thresholds penalize predictions only when they fall outside the uncertainty threshold. There are instances when distance alone may not capture whether the correct entity was selected for places that are near to each other.

8. Acknowledgments

The author acknowledges the Office of the Chancellor of the University of the Philippines Diliman, through the Office of the Vice Chancellor for Research and Development, for funding support through the Outright Research Grant (262622 ORG). This work was also partially supported by an Early Career Research Fellowship 2024-25 programme grant ("Enhancing Environmental Resilience through AI-driven Analysis of Unstructured Data") from the International Science Partnerships Fund (ISPF) delivered by the British Council.

9. Bibliographical References

Hazel Arceo, Joyce Velos, Ma Nuñez, and Porfirio Aliño. 2024. *The West Philippine Sea: State of the Coasts*.

T. Ayoola, J. Fisher, and A. Pierleoni. 2022. Improving entity disambiguation by reasoning over a knowledge base. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2899–2912. Association for Computational Linguistics.

Arthur Chapman and John Wieczorek. 2020. Georeferencing Best Practices. Publisher: GBIF Secretariat.

Jianlv Chen, Shitao Xiao, Peitian Hou, Quanyue Ye, Haotian Zhang, Huaying Cao, Chao Sun, Xueying Liang, and Zhiyong Cao. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text retrieval. *arXiv preprint arXiv:2402.03216*.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. What's missing in geographical parsing? *Lang Resour Eval*, 52(2):603–623.

Robert P Guralnick, John Wieczorek, Reed Beaman, Robert J Hijmans, and the BioGeomancer Working Group. 2006. Biogeomancer: Automated georeferencing to map the world's biodiversity data. *PLOS Biology*, 4(11):1–2.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *CoRR*, abs/2112.09118.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

John Michael Lastimoso and Wilfred John Santiañez. 2021. Updated checklist of the benthic marine macroalgae of the philippines. *Philippine Journal of Science*, 150:29–92.

Jochen L. Leidner. 2007. Toponym resolution in text. *ACM SIGIR Forum*, 41:124.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Arnald Marcer, Elspeth Haston, Quentin Groom, Arturo H. Ariño, Arthur D. Chapman, Torkild Bakken, Paul Braun, Mathias Dillen, Marcus Ernst, Agustí Escobar, David Fichtmüller, Laurence Livermore, Nicky Nicolson, Kaloust Paragamian, Deborah Paul, Lars B. Pettersson, Sarah Phillips, Jack Plummer, Heimo Rainer, Isabel Rey, Tim Robertson, Dominik Röpert, Joaquim Santos, Francesc Uribe, John Waller, and John R. Wieczorek. 2021. Quality issues in georeferencing: From physical collections to digital data repositories for ecological research. *Diversity and Distributions*, 27(3):564–567.

Simon Overell. 2011. The problem of place name ambiguity. *SIGSPATIAL Special*, 3:12–15.

R. Peeters, A. Steiner, and C. Bizer. 2024. Entity matching using large language models. *arXiv preprint arXiv:2310.11244*.

H. Soliman, H. Adel, M. H. Gad-Elrab, D. Milchevski, and J. Strötgen. 2022. A study on entity linking across domains: Which data is best for fine-tuning? In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 184–190.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Khagendra Thapa and John Bossler. 1992. Accuracy of spatial data used in geographic information systems. *Photogrammetric Engineering and Remote Sensing*, 58(6):835–841.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

John Wieczorek, Qinghua Guo, and Robert Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18:745–767.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of EMNLP 2020*, pages 6397–6407.

Xingrui Xie, Han Liu, Wenzhe Hou, and Hongbin Huang. 2023. A brief survey of vector databases. In *2023 9th International Conference on Big Data and Information Analytics (BigDIA)*, pages 364–371.

W. Zhang, S. Chen, J. Li, and C. Xu. 2025a. Geographic named entity matching and evaluation recommendation using multi-objective tasks: A study integrating a large language model and retrieval-augmented generation. *ISPRS International Journal of Geo-Information*, 14(3):95.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models.

Appendix

The Chain-of-Thought Prompt structure as well as the sample Few-shot prompt can be seen in Table 8.

Table 8: Prompts for the Large Language Model.

Prompt Strategies	**Prompt Content**
Chain of Thought	ROLE: You are an expert in Philippine place name disambiguation, geography, and historical gazetteers. Use step-by-step reasoning to resolve place names to coordinates. TASK: Determine which candidate correctly identifies this place mention. Step 1: Place Analysis. Characterize the mention by its place type... Step 2: Candidate Evaluation. Multi-factor comparison... Step 3: Exact Validation. Select only from the candidates' coordinates..
Few Shot Prompting	Example A (Geographic Disambiguation): Context describes a Sargassum bed in a coastal Philippine municipality; among candidates, the Ilocos municipality best fits this description.... Example B (Geographic context over lexical match): Mention: "Panaun Island", Context explicitly states Bolinao, Pangasinan ; Candidate 1: "Panaun Island" in Southern Leyte is geographically impossible despite lexical match; Candidate 2 "Panaon" is in Pangasinan and matches context; Candidate 2 is chosen...