A Large-Scale Dataset for Linking-Based Geocoding

Hibiki Nakatani[♣,*], **Yuichiro Yasui**[♠], **Ryosuke Wakamoto**[♠], **Masayuki Ishii**[♠], **Tetsuhisa Suizu**[♣], **Hiroki Ouchi**[♣,♦], **Taro Watanabe**[♣]
♣ NAIST, ♠ Nikkei Inc., ♦ RIKEN
{yuichiro.yasui, ryosuke.wakamoto, masayuki.ishii}@nex.nikkei.com
suizu.tetsuhisa.st8@naist.ac.jp, {nakatani.hibiki.ni4, hiroki.ouchi, taro}@is.naist.jp

Abstract

Linking-based geocoding is the task of linking location mentions in text to their corresponding entries in a geographic database (Geo-DB) and assigning precise coordinates. Although the task and its technology are essential for spatial information extraction, existing datasets are manually curated and lack sufficient data for training accurate models. To address this limitation, we automatically construct a large-scale dataset for linking-based geocoding by leveraging publicly available resources to generate data efficiently at scale. Specifically, we align location mentions in the first paragraphs of Japanese Wikipedia articles with their associated Wikidata entries containing geographic attributes. Wikipedia provides natural textual contexts, while Wikidata offers structured data such as coordinates, place types, and administrative divisions, which can serve as rich metadata for future extensions. Our experiments show that models trained on our dataset achieve strong performance not only on in-domain data, i.e., Wikipedia, but also on out-of-domain newspaper articles, and further confirm that hard negative mining substantially improves disambiguation among confusable candidates. Although the dataset focuses on Japanese, the construction method is language-agnostic and can be extended to other languages with sufficient Wikipedia and Wikidata coverage.

Keywords: Geocoding, Enityty Linking, Dataset and Benchmark

1. Introduction

Geocoding is the task of assigning geographic coordinates to location mentions that appear in natural language text. A ***location mention*** refers to any phrase that denotes a geographic place, such as a country, city, landmark, or facility name. Extracting and grounding such mentions in geographic space is crucial for spatial analysis from unstructured text data. Accurate geocoding supports a variety of applications, including urban planning, location-based services, disaster response, and public health monitoring (Hu et al., 2022).

Geocoding approaches can be categorized into two types: (i) *direct positioning* and (ii) *linking-based* approaches. The direct positioning approach directly estimates the geographic coordinate (or tile) of a location mention (Gritta et al., 2018; Kulkarni et al., 2021; Huang et al., 2022). In contrast, the linking-based approach searches in the geographic database (geo-DB) and identifies an entry with its coordinate corresponding to a location mention (Li et al., 2023; Halterman, 2023; Zhang et al., 2024; Gomes et al., 2024). Figure 1 shows ***linking-based geocoding***. The input is a sentence with a location mention (e.g., "Kumano Shrine"), and the output is the corresponding Wikidata entry ID (e.g., "`Q11568951`"). This task requires resolving ambiguities, as many different locations share the same name. For example, "Kumano Shrine" refers to over a thousand distinct locations across Japan, and the correct one must be identified based on the context.

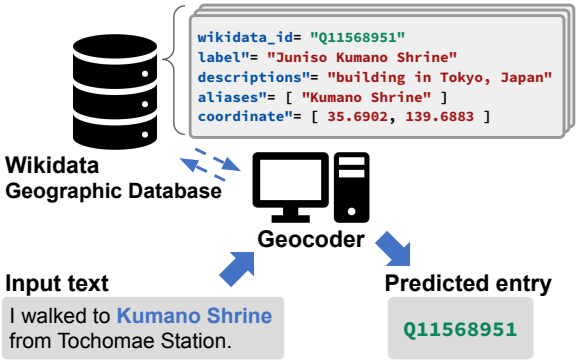

Figure 1: Overview of linking-based geocoding. Given an input sentence containing a location mention, the geocoder links it to the correct entry in a geographic database (Geo-DB). In this work, we constructed our Geo-DB from Wikidata. The database entries include attribute information such as labels, descriptions, aliases, and coordinates.

This work focuses on the linking-based approach because it has several advantages over direct positioning. Linking location mentions to entries in a geo-DB makes it possible to use attribute information such as place types, administrative divisions, aliases, and descriptions, which facilitates disambiguation between similarly named places. It also supports multilingual processing, since many geo-

* This work was done while he was affilliated with NAIST.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7644–7654
11–16 May 2026.

Dataset	Number of Mentions
GeoWebNews (Gritta et al., 2020)	2,401
TR-News (Kamalloo and Rafiei, 2018)	1,274
LGL (Lieberman et al., 2010)	4,793
ATD-MCL (Higashiyama et al., 2024b)	6,119
Ours	2,423,134

Table 1: Statistics of datasets for linking-based geocoding

DB include names in multiple languages. While linking-based geocoding has been actively studied for English, large-scale datasets for other languages, including Japanese, are still limited.

In this work, we present the first large-scale dataset for linking-based geocoding in Japanese. We automatically constructed the dataset by aligning location mentions in the first paragraph of Wikipedia articles with their corresponding Wikidata entries. The dataset covers a wide range of spatial granularities, from countries and cities to schools and stations. As shown in Table 1, our dataset comprises over two million mentions, which is several orders of magnitude larger than existing linking-based geocoding datasets. Our experiments show that models trained on our dataset achieve strong performance not only on in-domain data, i.e., Wikipedia, but also on out-of-domain newspaper articles, demonstrating robust generalization. We also find that performance scales with training data size, while a reduced subset still yields robust results, and that hard negative mining improves disambiguation among confusable entries.

Although the dataset is built for Japanese, the construction methodology is language-agnostic and can be adapted to other languages with sufficient coverage in Wikipedia and Wikidata. We position our research contribution as a step toward *more inclusive and multilingual geospatial information extraction*, addressing the current overreliance on English resources in geocoding research. Our main contributions are summarized as follows:

- **Dataset:** We present the first large-scale dataset for linking-based geocoding in Japanese, comprising over 2.4 million location mentions. The dataset is constructed through a fully automatic pipeline that leverages hyperlink structures and geographic attributes in Wikidata, enabling scalable and language-agnostic dataset generation without manual annotation. The dataset is publicly available for research.[1]
- **Evaluation:** We establish baseline results with both lexical and neural retrieval models, and demonstrate that *hard negative mining* substantially improves top-1 accuracy and cross-domain robustness, highlighting its importance for disambiguating geographically or semantically similar entities.

2. Related Work

Several datasets have been developed for the linking-based geocoding approach, mostly in English. Representative examples include GeoWebNews (Gritta et al., 2020), TR-News (Kamalloo and Rafiei, 2018), and LGL (Lieberman et al., 2010), which link place names in news articles to geographic entries in GeoNames[2]. These datasets address name ambiguity, where a single place name may refer to multiple locations. For example, “Paris” can indicate different cities depending on context.

For Japanese, the ATD-MCL dataset (Higashiyama et al., 2024b) provides fine-grained annotations by linking place mentions in travelogues to OpenStreetMap entries[3]. It covers a wide range of geographic features, from administrative areas to landmarks and facilities. Since expressions in user-generated text tend to be informal and varied, accurate linking poses unique challenges. However, the dataset is relatively small, with around 6,000 annotated mentions.

In contrast, large-scale geocoding datasets for Japanese exist only for the direct positioning approach. Ohno et al. (2024) automatically assigned coordinates to approximately 4 million Wikipedia mentions. While this offers broad geographic coverage, it does not link mentions to structured database entries, limiting its use in applications requiring attribute-rich location information.

To address this gap, we constructed a large dataset for linking-based geocoding by aligning location mentions in Wikipedia with entries in Wikidata that contain geographic attributes. While our dataset is based on Japanese text, the construction method is language-agnostic and can be extended to other languages with sufficient Wikipedia and Wikidata coverage.

[1] https://github.com/naist-nlp/wiki-geocoding-dataset

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

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

3. Dataset

3.1. Construction of the Geographic Database

We constructed a geo-DB from the Wikidata dump[4]. The Wikidata dump is a single BZIP2-compressed JSON file. It is difficult to handle because the file size is very large, around 90GB, and expands to approximately 1TB when decompressed. However, observing the dump data reveals a specific structure: the first and last lines are `[` and `]`, and every line represents a single entity. Therefore, we incrementally process the dump data by reading it line by line. Furthermore, when reading each entity line, we apply the following two filterings.

- The entry contains a Japanese label, that is, the `labels` field includes a `ja` key.
- The entry includes coordinate information, that is, the `claims` field contains the key `P625`.

The labels field contains language codes (such as ja or en) as keys and the entity labels as values. We determine if an entity has a Japanese label by checking for the presence of the ja key. Additionally, the claims field stores entity properties and values. We identify whether an entity should be included in our geographic database based on the presence of property P625, which corresponds to coordinates (latitude and longitude).

The dump used in this study, dated September 2, 2024, contains a total of 112,413,055 entries. Among these, 3,496,520 entries have Japanese labels and 435,867 entries include coordinate information. Since we focus on Japanese documents, we built the database by requiring both a Japanese label and geographic coordinates. By customizing these conditions according to the target language, it is possible to construct various types of databases.

3.2. Collection of Wikipedia Articles

Each Wikipedia page corresponds to a Wikidata entity, and each paragraph typically contains one or more hyperlinks to another page, providing naturally aligned pairs between surface mentions and entries. We leveraged this property to automatically collect text–entry pairs, storing each paragraph and its hyperlinks as individual dataset instances. This process enables efficient large-scale alignment of location mentions with their geographic counterparts without manual annotation.

Several types of Wikipedia dump data exist, and there are multiple ways to process the same information. To build a large dataset

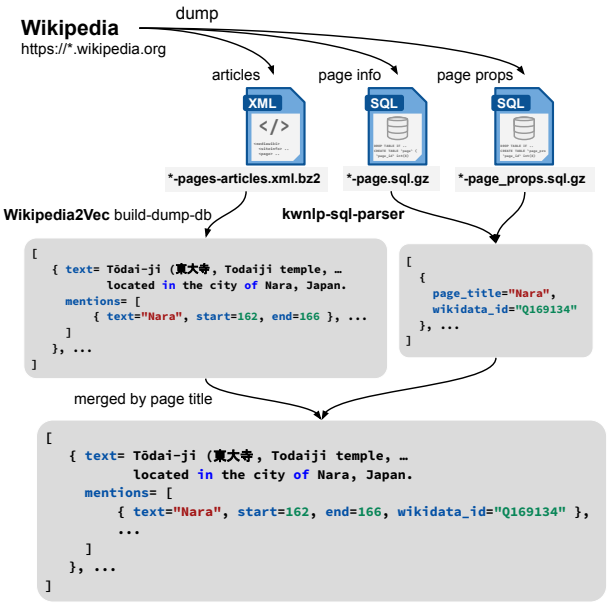

Figure 2: Overview of dataset construction. Each hyperlinked location mention (e.g., “Nara”) is extracted along with its character span and the corresponding Wikidata ID.

with minimal effort, we developed a simple data construction flow by combining existing tools. First, we use the text and link information from `*-pages-articles.xml.bz2` [5]. We process this using the `build-dump-db` command of `wikipedia2vec` (Yamada et al., 2020), which converts the dump into an LMDB-based key-value store. This allows us to retrieve the page title, text, link positions (start, end), and target page titles for each entry. In addition, we use `*-page.sql.gz` [6]. and `*-page_props.sql.gz` [7]. as a conversion table from page titles to Wikidata QIDs. According to the MediaWiki manuals: `page_title` and `pp_value` correspond to the page title and Wikidata QID on Page table [8] and Page props table [9], respectively. Since these are provided in SQL format, we converted them into tables via `kwnlp-sql-parser` [10]. To assign QIDs to linked

[4]The latest Wikidata dump is available at https://dumps.wikimedia.org/wikidatawiki/entities/latest-all.json.bz2

[5]The latest version is available at https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-pages-articles.xml.bz2.

[6]The latest version is available at https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-page.sql.gz.

[7]The latest version is available at https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-page_props.sql.gz.

[8]https://www.mediawiki.org/wiki/Manual:Page_table

[9]https://www.mediawiki.org/wiki/Manual:Page_props_table

[10]https://github.com/kensho-technologies/kwnlp-sql-parser

Split	# Pages	# Mentions
Train	718,398	1,938,507
Dev	104,556	242,313
Test	107,044	242,314
Nikkei News	28	234

Table 2: Statistics of the experimental dataset

pages, we merge them using the page titles mentioned above. When we applied the conditions that the page is in namespace 0, is not a redirect, and has a non-null title, we successfully matched 1,407,078 items. Only 38 items (less than 0.01%) failed to match. Ideally, we should not merge by page title because it is changeable and the `page_id` is available in the XML dump. However, considering the implementation cost and the high accuracy achieved, we chose this simpler method.

Figure 2 illustrates an example of the data construction process. The figure shows the first paragraph of the Wikipedia article for "Tōdai-ji (東大寺)," which contains a hyperlinked mention "Nara" from `*-pages-articles.xml.bz2`. And the page 'Nara' linked to the Wikidata entry "Q169134," from page data `*-page.sql.gz` and `*-page_props.sql.gz`. We extract pairs of hyperlinks and their corresponding Wikidata entities. These elements are stored as a triplet consisting of the full text, the mention, and its linked entry (Wikidata ID).

We constructed our dataset by extracting hyperlinked location mentions from the first paragraphs of Japanese Wikipedia articles as of September 1, 2024. It contains 929,998 documents and 151,380 unique expressions linked to 117,302 Wikidata geographic entities. The resulting data instances were divided into training, development, and test splits to support reproducible experimentation, as summarized in Table 2.

We analyze two properties of our dataset: (a) how often each mention expression appears in the dataset, and (b) how many distinct gold entities are associated with each mention expression. Figure 3 shows the distribution of mention expressions by their frequency in the dataset. Although 89.2 % of the mentions occur only once, the most frequent ones include "Japan (日本)" (180,680 occurrences), "United States (アメリカ合衆国)" (64,583 occurrences), and "Tokyo (東京都)" (43,025 occurrences). Figure 4 shows the distribution of mention expressions by the number of distinct gold entities linked to them. While 95.7% of mentions are linked to only one entity, some are highly ambiguous, including "Family Court (家庭裁判所)" (linked to 29 entities), "Subway (地下鉄)" (linked to 27 entities), and "Tokyo (東京)" (linked to 27 entities).

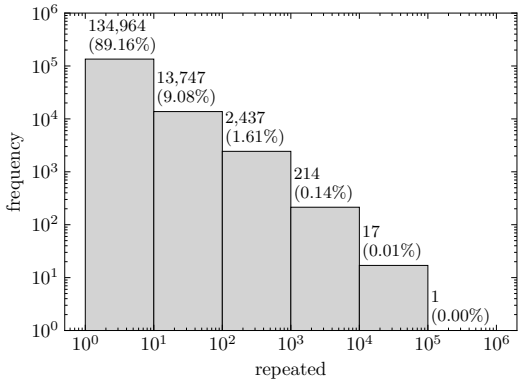

Figure 3: Distribution of mention expressions by frequency.

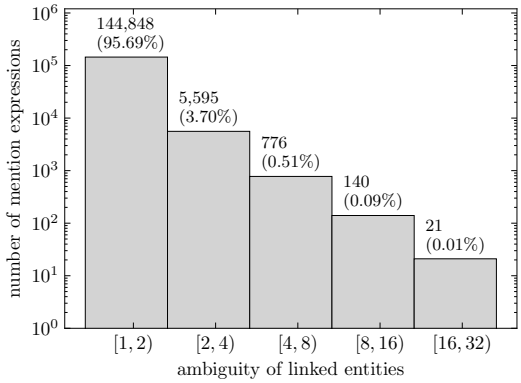

Figure 4: Distribution of mention expressions by the number of distinct linked entities.

3.3. Collection of Nikkei Newspaper Articles

We select news as a domain different from Wikipedia and use it for our experiments. We believe news is suitable for evaluation because location names appear frequently and are highly important, meaning that errors are difficult to tolerate. We randomly sampled 28 articles from the Nikkei newspaper (morning, evening, and digital editions) and Nikkei Business Daily from the years 2021–2023 and used them as evaluation data[11]. These annotated mentions are based on named entities with labels, `GPE`, `GPE_ORG`, and `FACILITY`, with the guideline as (Higashiyama et al., 2024b). The `GPE` and `GPE_ORG` labels denote place names, such as countries, prefectures, and cities, while `FACILITY` represents facility names. The distinction between `GPE` and `GPE_ORG` lies in whether the entity is a location or a governmental organization, and these entities will be linked to different database entries. There were three annotators in total, and two or more annotators agreed upon all mentions, assur-

[11]The articles and their annotations are available at the Nikkei Dataset for a fee: https://nkbb.nikkei.co.jp/en/dataset/nikkei-news-articles/.

ing the annotation in high quality.

4. Experimental Settings

In the experiments, our objective is to investigate the basic performance of baseline models. In particular, we conduct a comparative analysis between results obtained under the same training and evaluation settings (in-domain) and those obtained under different settings (out-of-domain).

4.1. Baseline Models

The linking-based geocoding task can be formulated as a *retrieval problem*. Given a location mention and its surrounding context, the goal is to retrieve the most appropriate entry from a geo-DB. Accordingly, we choose baselines that are standard in large-scale retrieval, scale to millions of geo-DB entries, and represent complementary paradigms (lexical vs. dense). Specifically, we evaluate BM25 (Robertson et al., 1995) and E5 (Nakatani et al., 2025). BM25 is an efficient lexical baseline that is commonly used in entity linking and candidate retrieval. E5 is a competitive multilingual dense retriever because prior work has shown that E5-style text embeddings are effective for geocoding, especially when combined with contrastive training and hard negative mining (Nakatani et al., 2025).

4.1.1. BM25

As a lexical baseline, we adopt BM25[12], which ranks candidate entries according to their string similarity to the input mention. Each query corresponds to a mention string, and each candidate entry is represented by its official name. We intentionally restrict the candidate representation to the official name for BM25. In preliminary experiments, concatenating additional fields such as descriptions or aliases degraded performance, likely because lexical retrieval models are sensitive to term frequency and may introduce noise when irrelevant tokens are added. This design choice is consistent with prior findings in entity linking that lexical matching benefits from concise surface-form representations, while excessive contextual text can dilute exact-match signals (Logeswaran et al., 2019). All entries are tokenized using the tokenizer of the E5 encoder for consistency. This method provides a simple yet strong baseline for assessing the effectiveness of neural representations.

[12]We used an implementation by Lù (2024), BM25-Sparse (https://github.com/xhluca/bm25s).

4.1.2. E5 and its variants

Another baseline is based on E5[13] (Wang et al., 2024a,b), a transformer-based encoder originally developed for semantic search. Both the location mention and the candidate entries are encoded into dense vector representations, and the most relevant entry is identified by computing cosine similarity in the embedding space. For each candidate entry, we concatenate its Japanese label (official name, e.g. “Juniso Kumano Shrine”), aliases (e.g. “Kumano Shrine”), and short description (e.g. “building in Tokyo, Japan”) into a single text string, and feed this as input to the E5 encoder (see Figure 1). Unlike lexical models, dense encoders benefit from richer contextual information, as additional attributes provide semantic cues that help distinguish between geographically or semantically similar entities. Prior work has shown that contextualized representations improve entity linking and retrieval performance when candidate descriptions are incorporated (Logeswaran et al., 2019).

We draw on insights from Nakatani et al. (2025), who propose a specialized text embedding model for geocoding using contrastive example mining. Their method extends standard contrastive training by effectively mining hard negative examples, which sharpens the model’s discrimination among geographically or semantically similar candidates and helps reduce confusion between locations with similar names. To investigate the impact of negative sampling strategies, we evaluate three variants of the E5 model:

- IN-BATCH: uses other samples within the same mini-batch as negative examples. This is the standard strategy in contrastive learning.
- IN-BATCH-RANDOM: supplements in-batch negatives by randomly sampling one additional negative entry from the entire database for each positive pair, increasing sample diversity.
- IN-BATCH-HARD: replaces random negatives with hard negatives that are geographically or semantically close to the positive entry. In this setting, we follow the hard negative mining procedure of Nakatani et al. (2025), in which the top-N entries are pre-retrieved from the geo-DB using BM25 before training and one of these candidates is randomly selected for each epoch. This approach allows the model to focus on more confusable entries while maintaining training efficiency.

Table 4 summarizes the hyperparameter settings used for our experiments.

[13]We used an implementation available at https://huggingface.co/intfloat/multilingual-e5-base

Method	Sampling	Recall@1	Recall@5	Recall@10	MRR
Wikipedia (In-Domain Setting)					
BM25	-	0.819	0.944	0.958	0.876
E5	IN-BATCH	0.884	0.984	0.991	0.927
	IN-BATCH-RANDOM	0.888	0.985	0.992	0.931
	IN-BATCH-HARD	**0.969**	**0.993**	**0.995**	**0.980**
Nikkei Newspaper (Out-of-Domain Setting)					
BM25	-	0.387	0.583	0.640	0.481
E5	IN-BATCH	0.605	0.890	0.920	0.715
	IN-BATCH-RANDOM	0.664	0.903	0.926	0.777
	IN-BATCH-HARD	**0.846**	**0.937**	**0.947**	**0.890**

Table 3: Main results. The best scores for each metric in each dataset are shown in bold.

Hyperparameter	Value
Training epochs	1
Batch size	16
Weight decay	0.01
Adam β_1	0.9
Adam β_2	0.98
Adam ϵ	1e-6
Learning rate	1e-5
Learning rate scheduler	linear
Warmup ratio	0.06
Optimizer	AdamW

Table 4: Hyperparameters used for model fine-tuning.

4.2. Evaluation Metrics

We evaluate model performance based on how well it ranks the correct geo-DB entry for each input mention. Following standard retrieval metrics, we report Mean Reciprocal Rank (MRR) and Recall@k (R@k). MRR is the average reciprocal rank of the correct entry across all queries:

$$\mathrm{MRR} = \frac{1}{q}\sum_{i=1}^{q}\frac{1}{\mathrm{rank}(m_i, e_i)}, \quad (1)$$

where m_i is the i-th mention, e_i is its correct geo-DB entry, and $\mathrm{rank}(m_i, e_i)$ is its position in the ranked list. Recall@k measures the proportion of cases in which the correct entry appears in the top-k results.

For BM25, which may assign the same score to multiple entries, we compute the expected value of Recall@k and MRR as proposed in (Higashiyama et al., 2024a). The expected MRR for mention m_i is calculated as:

$$\begin{aligned}\mathrm{MRR}_i &= \frac{1}{|E_i|}\sum_{j=1}^{|E_i|}\frac{1}{\mathrm{rank}(m_i, e_j)},\\ E_i &= \{e_j \mid s(m_i, e_j) = s(m_i, e_i)\}.\end{aligned}$$

5. Results and Analysis

5.1. Main Results

Table 3 presents the performance of each model on two datasets: the Wikipedia test set (in-domain) and the Nikkei newspaper dataset (out-of-domain). We report Recall@k ($k \in \{1, 5, 10\}$) and Mean Reciprocal Rank (MRR) as evaluation metrics.

5.1.1. Results on Wikipedia test set (In-Domain Setting)

Overall comparison The E5 model outperforms the BM25 baseline, achieving a 0.150 point gain in Recall@1 (from 0.819 to 0.969). This improvement indicates that the E5 model is able to effectively learn semantic associations between location mentions and entry descriptions, even when there is limited lexical overlap. In contrast, BM25 relies solely on surface-level string similarity, which can struggle when mentions are expressed in abbreviated, variant, or context-dependent forms.

Comparison among E5 variants Among the three E5 variants, the IN-BATCH-HARD model achieves the highest scores across all metrics (Recall@1 = 0.969, Recall@5 = 0.993, MRR = 0.980). The IN-BATCH-RANDOM variant yields slightly better performance than the standard IN-BATCH setup, suggesting that additional random negatives contribute to the more stable learning. However, the largest improvement is observed with the IN-BATCH-HARD strategy, which introduces challenging negative examples that are semantically or geographically similar to the correct entries. This result confirms that exposure to such hard negatives helps the model to distinguish between highly confusable locations, leading to substantial gains in top-ranked retrieval accuracy.

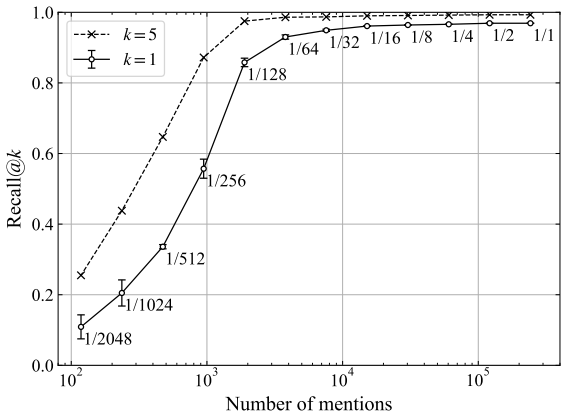

Figure 5: Recall@1 and Recall@5 on the Wikipedia test set as the training data size is reduced by $1/2^n$, $n \in \{1, 2, \ldots, 11\}$.

5.1.2. Results on the Nikkei dataset (Out-of-Domain Setting)

Overall comparison On the Nikkei dataset, the E5 model improves Recall@1 by 0.459 points compared to BM25 (from 0.387 to 0.846). This result indicates that our Wikipedia-based training dataset enables strong generalization to unseen domains, and can serve as a valuable resource in cases where training data is scarce or limited in scope.

Comparison among E5 variants Similar to the in-domain results, the three E5 variants show a consistent trend in performance. The IN-BATCH-RANDOM model performs slightly better than the standard IN-BATCH variant (Recall@1 = 0.664 vs. 0.605), suggesting that exposure to more diverse negative samples enhances robustness against domain shift. The IN-BATCH-HARD model again achieves the best performance across all metrics (Recall@1 = 0.846, Recall@5 = 0.937, MRR = 0.890), demonstrating that hard negative mining effectively improves cross-domain generalization. This result implies that distinguishing between semantically similar but contextually distinct entries is crucial for handling real-world texts, such as newspaper articles, where place mentions often appear in different styles or levels of specificity compared to Wikipedia.

5.2. Impact of Training Data Size on Performance

We examined how the training data size influences model performance. To this end, we constructed a series of reduced training sets by randomly sampling location mentions at fractional rates of $1/2^n$, where $n \in \{1, 2, \ldots, 11\}$. Figure 5 presents the results in terms of Recall@1 and Recall@5.

Two major trends are observed. First, performance decreases gradually as the training data size is reduced, confirming that the model benefits from large-scale supervision and learns meaningful associations between mentions and geographic entities. This trend underscores the importance of sufficient data coverage for achieving high top-1 accuracy, especially in ambiguous or low-frequency cases. Second, the model maintains relatively stable performance when trained with up to $1/16$ of the original data, with only minor degradation in Recall@5. This finding suggests that the dataset contains a degree of redundancy and that the model generalizes well even with a substantially smaller subset of examples.

Overall, these results indicate that the proposed dataset is both scalable and data-efficient: while larger datasets improve fine-grained disambiguation, a reduced subset is sufficient to achieve robust performance in most cases.

5.3. Qualitative Analysis

To better understand why the IN-BATCH-HARD variant achieved the best performance in Table 3, we conducted a qualitative analysis comparing its predictions with those of the other variants. By examining representative cases from the Wikipedia development set, we aimed to identify how the models differ in handling ambiguous or difficult mentions. Table 5 presents two contrasting cases that illustrate these differences.

5.3.1. Case (a): “Taipei”

For the mention “Taipei,” both the IN-BATCH and IN-BATCH-RANDOM variants incorrectly ranked “W Taipei” (a hotel) above the correct entry “Taipei City,” placing it at positions 5 and 2, respectively. The IN-BATCH-HARD, however, successfully ranked “Taipei City” at the top, which suggests that training with hard negatives helps the model handle cases where multiple entities share highly similar surface strings. The improvement may indicate that exposure to more challenging negatives enables the model to capture subtle contextual differences, rather than relying solely on surface similarity.

5.3.2. Case (b): “Hokkoku Kaidō”

For the mention “Hokkoku Kaidō,” all model variants failed to find the correct entry “National Route 18.” Instead, they ranked other historical roads, such as “Hokurikudō,” much higher. This error likely arises because the historical name in the text (“Hokkoku Kaidō”) differs greatly from the modern official name (“National Route 18”) stored in the database. In such cases, where surface forms have little overlap,

Sampling	Top Predicted Entry	Rank of Correct Entry
Case (a): mention="Taipei", correct_entry_name="Taipei City", context="The 15th AFC Futsal Championship was held in <ent> Taipei <ent2>, Taiwan from ..."		
in-batch	W Taipei	5
in-batch-random	W Taipei	2
in-batch-hard	Taipei City	1
Case (b): mention="Hokkoku Kaidō", correct_entry_name="National Route 18", context="Shinshu-Kama is a sickle made in Shinshu, along the <ent> Hokkoku Kaidō <ent2> in ..."		
in-batch	Naganokaidō 12	4,371
in-batch-random	Hokuenkaidō	4,245
in-batch-hard	Hokurikudō	4,219

Table 5: Prediction examples of the E5 model on the Wikipedia development set.

purely text-based similarity is insufficient for correct linking.

5.3.3. Discussion

The two examples above illustrate different types of challenges in linking-based geocoding.

The first case ("Taipei") represents situations where several entities share almost the same surface string. In such cases, the model must rely on contextual clues to decide which entry is correct. Our results show that hard negative mining is effective for these cases, because it exposes the model to confusing examples and helps it learn finer distinctions. This suggests that hard negative mining already improves the model's ability to handle surface-level ambiguity.

The second case ("Hokkoku Kaidō") shows a different kind of difficulty. Here, the name in the text and the name in the database are very different. Such discrepancies often arise when the name used in the text differs greatly from the official name recorded in the database, for example due to local nicknames, abbreviations, or historical naming changes, where the surface forms share little lexical overlap. Since the model mainly depends on text-based similarity, it struggles to link mentions like this correctly.

These findings suggest that contrastive learning alone cannot fully resolve all types of errors. While hard negative mining reduces confusion among mentions with similar surface forms, it remains limited when the name used in the text differs greatly from the official name in the database, such as local nicknames, abbreviations, or historical naming changes. To handle these more diverse cases, models may need to incorporate additional knowledge-based cues, for example:

- **Alias expansion:** include alternative, abbreviated, local, or historical names to improve coverage of non-standard expressions.
- **Attribute prediction:** jointly predict properties such as region or category to narrow down candidate entries.
- **Graph-aware re-ranking:** leverage structured relations in the database, such as *part of*, *follows*, or *replaced by*, to connect related entities with different names.

Combining such knowledge-based cues with contrastive learning could make the model more robust and better able to link mentions correctly, even when the wording in the text and the database entry differ substantially.

6. Conclusion

In this paper, we presented a large-scale dataset for linking-based geocoding in Japanese, automatically constructed by aligning location mentions in Wikipedia with corresponding Wikidata entries containing geographic attributes. Models trained on this dataset achieved high accuracy on both in-domain and out-of-domain evaluations, demonstrating strong generalization ability despite being trained only on Wikipedia text.

Our analysis of training data size showed that model performance improves with more training data, while a reduced subset is still sufficient to achieve robust performance in most cases. Furthermore, our qualitative analysis revealed that hard negative mining effectively reduces ambiguity among entities with similar surface forms, while challenges remain for cases where textual and official names differ greatly, such as local nicknames or historical naming variations. These findings highlight the importance of incorporating knowledge-based cues, such as alias expansion and graph-aware reasoning, to handle complex name discrepancies in real-world geocoding scenarios.

For future work, we plan to extend the dataset beyond the first paragraphs to cover full Wikipedia articles, and to include additional language editions

and geographic databases for broader multilingual and cross-domain evaluation.

Ethics Statement

License of Used Resources All resources used in this study are publicly available and comply with their respective licenses. The proposed dataset was automatically constructed from publicly accessible Japanese Wikipedia articles and their associated Wikidata entries. Both Wikipedia and Wikidata are released under the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA 3.0). The text used in our dataset consists solely of excerpts from publicly available Wikipedia pages and does not include any personal information about contributors or editors.

The evaluation set derived from Nikkei newspaper articles was used only for experimental evaluation within the scope of fair academic research. The Nikkei newspaper articles and their annotations are available at the Nikkei Dataset[14] for a fee.

The pretrained models used in this work are openly released under permissive licenses: the BM25-Sparse implementation under the MIT License, and the multilingual E5 model under the Apache License 2.0.[15] All other software dependencies follow their original open-source licenses.

Human Annotation Effort The annotation work was performed by professional annotators at a news company. The work involved three annotators, all native Japanese speakers, with one also serving as the manager overseeing the process. We informed the annotators that the data would be used for future NLP research. All had prior experience with Japanese text annotation and were in their 20s to 40s. The total annotation time amounted to approximately 3 hours.

Predicted Results for Real-World Applications Models trained on our dataset may produce incorrect link predictions, resulting in wrong geographic coordinates or mismatched place entities. Such errors could affect downstream applications that rely on precise geolocation, for example, spatial analysis or geographic information retrieval. Therefore, users should carefully validate model outputs before applying them to real-world systems or decision-making processes.

[14] https://nkbb.nikkei.co.jp/en/dataset/nikkei-news-articles/

[15] https://huggingface.co/intfloat/multilingual-e5-base

Limitations

Language Our dataset was constructed from Japanese Wikipedia and Wikidata entries, and therefore all experiments in this paper were conducted in Japanese. The data construction procedure itself, however, is language-agnostic and can, in principle, be applied to other languages with sufficient Wikipedia and Wikidata coverage. Future work will explore multilingual extensions of our dataset to evaluate how well the linking-based geocoding framework generalizes across languages.

Geographical Coverage Because the dataset relies on Japanese Wikipedia, the majority of entries correspond to locations in Japan, with limited coverage for foreign geographic entities. Moreover, the geographic database derived from Wikidata may not be uniformly detailed across regions, especially for small-scale or non-urban features. Expanding the dataset to include other language editions of Wikipedia or additional geographic databases (e.g., OpenStreetMap) would improve spatial diversity and support broader cross-regional evaluation.

Source Diversity and Generalizability Our dataset was constructed exclusively from Japanese Wikipedia, which provides a consistent but encyclopedic writing style. This reliance ensures data quality and alignment with Wikidata entries, but it limits the diversity of linguistic expressions and contexts. Although our experiments on Nikkei newspaper articles demonstrated that models trained on this dataset can generalize reasonably well to out-of-domain text, further validation on other genres, such as web documents, user-generated content, or historical texts, remains an important direction for future work. Expanding source diversity would help evaluate the robustness of linking-based geocoding models under more varied language use and context.

Dataset Size Our automatically constructed dataset contains over 2.4 million mention–entry pairs extracted from Japanese Wikipedia, making it several orders of magnitude larger than existing manually curated geocoding datasets. The large scale of the data enables robust training and evaluation of retrieval-based models for location linking. However, the evaluation data derived from Nikkei newspaper articles is relatively small, consisting of only 28 articles, and may not fully represent the linguistic and topical diversity of real-world text. In future work, we plan to expand the evaluation to include larger and more diverse out-of-domain corpora, which would allow a more comprehensive

assessment of model generalization and domain robustness.

Optimization of System Performance We performed minimum hyperparameter search for the models due to time and resource limitations. Thus, performing optimized experiments has potential for further performance improvement in these models.

Model Diversity We evaluated BM25 and E5 as representative lexical and dense retrieval baselines. While E5 is a competitive multilingual embedding model and has been shown to perform well for geocoding tasks, we did not systematically compare other embedding models or large language models (LLMs). Recent advances in LLMs suggest that strong performance may be achievable. We leave such comparison for future work.

Bibliographical References

Diego Gomes, Ross S Purves, and Michele Volpi. 2024. Fine-tuning Transformers for toponym resolution: A contextual embedding approach to candidate ranking. In *Proceedings of The GeoExT 2024: Geographic Information Extraction from Texts Workshop*, pages 43–51.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Which Melbourne? Augmenting geocoding with maps. In *Proceedings of ACL*, pages 1285–1296.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, 54(3):683–712.

Andrew Halterman. 2023. Mordecai 3: A neural geoparser and event geocoder. arXiv:2303.13675.

Shohei Higashiyama, Masao Ideuchi, and Masao Utiyama. 2024a. Construction of the administrative agency web document corpus for Japanese entity linking [in Japanese]. *IPSJ SIG Technical Report*, 2024-NL-260(10):1–15.

Shohei Higashiyama, Hiroki Ouchi, Hiroki Teranishi, Hiroyuki Otomo, Yusuke Ide, Aitaro Yamamoto, Hiroyuki Shindo, Yuki Matsuda, Shoko Wakamiya, Naoya Inoue, Ikuya Yamada, and Taro Watanabe. 2024b. Arukikata travelogue dataset with geographic entity mention, coreference, and link annotation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 513–532, St. Julian's, Malta. Association for Computational Linguistics.

Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2022. Location reference recognition from texts: A survey and comparison. arXiv:2207.01683.

Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. 2022. ERNIE-GeoL: A geography-and-language pre-trained model and its applications in Baidu maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pages 3029–3039, New York, NY, USA. Association for Computing Machinery.

Ehsan Kamalloo and Davood Rafiei. 2018. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1287–1296, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, Eugene Ie, and Li Zhang. 2021. Multi-level gazetteer-free geocoding. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 79–88, Online. Association for Computational Linguistics.

Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. 2023. GeoLM: Empowering language models for geospatially grounded language understanding. In *Proceedings of EMNLP*, pages 5227–5240.

Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering*, pages 201–212. IEEE.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of ACL*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Xing Han Lù. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. arXiv:2407.03618.

Hibiki Nakatani, Hiroki Teranishi, Shohei Higashiyama, Yuya Sawada, Hiroki Ouchi, and Taro Watanabe. 2025. A Text Embedding Model with Contrastive Example Mining for Point-of-Interest Geocoding. In *Proceedings of COLING*.

Keyaki Ohno, Hirotaka Kameko, Keisuke Shirai, Taichi Nishimura, and Shinsuke Mori. 2024. Automatic construction of a large-scale corpus for geoparsing using Wikipedia hyperlinks. In *Proceedings of LREC-COLING*, pages 1883–1888, Torino, Italia. ELRA and ICCL.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp*, 109:109.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual E5 text embeddings: A technical report. arXiv:2402.05672.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of EMNLP*, pages 23–30. Association for Computational Linguistics.

Zeyu Zhang, Egoitz Laparra, and Steven Bethard. 2024. Improving toponym resolution by predicting attributes to constrain geographical ontology entries. In *Proceedings of NAACL*, pages 35–44, Mexico City, Mexico. Association for Computational Linguistics.