Arukikata Travelogue Dataset with Geographic Entity Mention, Coreference, and Link Annotation

Anonymous ACL submission

Abstract

Geoparsing or geo-entity linking is a fundamental technique for analyzing geo-entity information in text, which is useful for geographic applications, e.g., tourist spot recommendation. We focus on *document-level* geoparsing that considers geographic relatedness among geo-entity mentions and present a Japanese travelogue dataset designed for evaluating document-level geoparsing systems. Our dataset comprises 200 travelogue documents with rich geo-entity information: 12,171 mentions, 6,339 coreference clusters, and 2,551 geo-entities linked to geo-database entries.

1 Introduction

Natural language expressions of locations or geographic entities (*geo-entities*) are often written in text to describe real-world events and human mobility. Thus, technologies for extracting and grounding geo-entity expressions are important for realizing various geographic applications, e.g., recommendation of tourist spots and tour routes to travelers.

Geoparsing (Leidner, 2006; Gritta et al., 2020) is a fundamental technique involving two subtasks: *geotagging*, which identifies geo-entity mentions, and *geocoding*, which identifies corresponding database (DB) entries for (or the coordinates of) geo-entities. Notably, geoparsing, geotagging, and geocoding can be regarded as special cases of entity linking (EL), named entity recognition (NER) or mention recognition (MR), and entity disambiguation (ED), respectively.

This study focuses on geoparsing from the perspective of *document-level* analysis. Geo-entity mentions that co-occur in a document tend to be geographically close or related to each other; thus, information about some geo-entity mentions can be useful in specifying information about other mentions. For example, by considering the context that describes a trip to Nara Prefecture, Japan, the mention of 興福寺 *kofukuji* ‘Kofukuji Temple’ in Figure 1 ⟨3⟩ can be disambiguated to refer to the temple in Nara rather than temples with the same name at different locations.

近鉄奈良駅$^{\text{FAC-NAME}}_{\langle 1\rangle}$ に到着。そこ$^{\text{DEICTIC}}_{\langle 1\rangle}$ から
奈良公園$^{\text{FAC-NAME}}_{\langle 2\rangle}$ までは歩いてすぐです。
お寺$^{\text{FAC-NOM}}_{\langle\text{GENERIC}\rangle}$ が好きなので最初に興福寺$^{\text{FAC-NAME}}_{\langle 3\rangle}$
に行きました。境内$^{\text{FAC-NOM}}_{\langle 3\rangle}$ で鹿と遭遇し、
奈良$^{\text{LOC-NAME}}_{\langle 4\rangle}$ に来たことを実感しました。

I arrived at Kintetsu Nara Station$^{\text{FAC\_NAME}}_{\langle 1\rangle}$.
From there$^{\text{DEICTIC}}_{\langle 1\rangle}$ it’s a short walk to
Nara Park$^{\text{FAC\_NAME}}_{\langle 2\rangle}$. I like temples$^{\text{FAC\_NOM}}_{\langle\text{GENERIC}\rangle}$
so I first went to Kofukuji Temple$^{\text{FAC\_NAME}}_{\langle 3\rangle}$.
I encountered a deer in the precincts$^{\text{FAC\_NOM}}_{\langle 3\rangle}$ and
felt that I had come to Nara$^{\text{LOC\_NAME}}_{\langle 4\rangle}$.

⟨1⟩ https://www.openstreetmap.org/relation/11532920
⟨2⟩ https://www.openstreetmap.org/way/456314269
⟨3⟩ https://www.openstreetmap.org/way/1134439456
⟨4⟩ https://www.openstreetmap.org/relation/3227707

Figure 1: Example illustration of an annotated document with English translation. Expressions underlined in blue indicate geo-entity mentions, superscript strings (e.g., FAC-NAME) indicate entity types of mentions, and subscript numbers (e.g., ⟨1⟩) indicate coreference cluster IDs of mentions. URLs indicate OpenStreetMap entries that correspond to coreference clusters.

This paper presents a dataset suitable for document-level geoparsing: the Arukikata Travelogue Dataset with geographic entity Mention, Coreference, and Link annotation (ATD-MCL). Our dataset includes the three types of geo-entity information illustrated in Figure 1: (1) spans and entity types of geo-entity mentions, (2) coreference relations among mentions, and (3) links from coreference clusters to corresponding entries in a geographic DB (geo-DB).

Our dataset has two desirable characteristics for document-level geoparsing. The first characteristic is that travelogues in our dataset have a sufficient amount of *geography-related content*, that is, a series of geo-entity mentions that are geographically related to each other, e.g., coreference relations. This is in contrast to short documents, e.g., tweets (Matsuda et al., 2017; Wallgrün et al., 2018). The second characteristic is their *geographic continuity* among co-occurring mentions; that is, mentions that refer to nearby locations in the real world tend to appear near to one another within a document. Because travel records reflect the actual trajectories of travelers, this characteristic is more notable in travelogues than other text genres, e.g., news articles (Lieberman et al., 2010; Kamalloo and Rafiei, 2018; Gritta et al., 2018a, 2020).

As a result of manual annotation, our dataset comprises 12,273 sentences from the full text of 200 travelogue documents with 12,171 geo-entity mentions, 6,339 coreference clusters (geo-entities), and 2,551 linked geo-entities. Furthermore, we have conducted two types of evaluation using our dataset. First, we have measured inter-annotator agreement (IAA) for three types of information; the results indicate the practical quality of our dataset in terms of consistency. Second, we have evaluated current entity analysis systems on our dataset for benchmarking baseline performance; the results demonstrate that reasonable performance can be achieved for MR and coreference resolution (CR), but performance has room for improvement in ED.[1]

2 Dataset Annotation

Design Strategy For building geoparsing datasets, it has been challenging to achieve a high coverage for facility entity mentions mainly because of the limited coverage of public geo-DBs, e.g., GeoNames. To address this DB coverage problem, we adopt OpenStreetMap (OSM),[2] a free, editable, and large-scale geo-DB of the world. The usefulness of OSM has been continually increasing, as evidenced by the increase in node entries from over 1.5B in 2013 to over 80B in 2023.[3] Furthermore, we define entity types to cover broad types of location and facility mentions, including districts, buildings, landmarks, roads, and public transport lines and vehicles, as described in §2.2.

Annotation Flow Following the data preparation by the authors, annotation work was performed by native Japanese annotators at a professional data annotation company according to the three-step annotation flow: (1) mention annotation, (2) coreference annotation, and (3) link annotation.

2.1 Data Preparation

As raw text data, we adopted the ATD[4] (Arukikata. Co., Ltd., 2022; Ouchi et al., 2023), which was constructed from user-posted travelogues written in Japanese. We first sampled documents about Japanese domestic travel with a reasonable document length (500–3000 characters, that is, approximately 300–1800 words) from the ATD. We then applied the GiNZA NLP Library[5] (Matsuda et al., 2019) to the raw text for sentence segmentation and automatic annotation of named entity (NE) mention candidates.

2.2 Mention Annotation

In the mention annotation step, we required the annotators to identify spans of geo-entity mentions in the documents, which may or may not refer to real-world locations, and assign entity type tags to the identified mentions by modifying the auto-annotated NE mentions. We adopted the brat annotation tool[6] (Stenetorp et al., 2012) for mention annotation (and succeeding coreference annotation).

The criteria for mention annotation define the *entity types* of geo-entity mentions, along with *mention spans* explained in Appendix B. Specifically, we define the following eight main entity types, which roughly correspond to `Location`, `Facility`, and `Vehicle` in Sekine's Extended Named Entity (ENE) taxonomy (version 9.0)[7] (Sekine et al., 2002). (1) `LOC`, (2) `FAC`, and (3) `TRANS` respectively represent locations, facilities, and public transport vehicles; (4) `LINE` represents roads, waterways/rivers, or public transport lines. The above four types are further divided into `NAME` and `NOM` subtypes, corresponding to whether a mention is named or nominal, as described in Table 1. (5) `LOC_ORG` and (6) `FAC_ORG` indicate location and facility mentions, respectively, that metonymically refer to organizations, e.g., ホテル *hoteru* in a sentence such as "The hotel serves its lunch menu." (7) `LOC_OR_FAC` indicates nominal mentions that

[1] We will release our annotated data and experimental codes at ANONYMIZED_URL.

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

[3] https://wiki.openstreetmap.org/wiki/Stats

[4] https://www.nii.ac.jp/dsc/idr/arukikata/

[5] https://github.com/megagonlabs/ginza

[6] https://github.com/nlplab/brat

[7] http://ene-project.info/ene9/

Type and subtype	Example mentions
LOC-NAME LOC-NOM	奈良 'Nara'; 生駒山 'Mt. Ikoma' 町 'town'; 島 'island'
FAC-NAME FAC-NOM	大神神社 'Ōmiwa Shrine' 駅 'station'; 公園 'park'
LINE-NAME LINE-NOM	近鉄奈良線 'Kintetsu Nara Line' 国道 'national route'; 川 'river'
TRANS-NAME TRANS-NOM	特急ひのとり 'Ltd. Exp. Hinotori' バス 'bus'; フェリー 'ferry'

Table 1: Examples of NAME and NOM entity mentions.

can refer to both location and facility, e.g., 観光地 *kankōchi* 'sightseeing spot.' Finally, (8) DEICTIC indicates deictic expressions that refer to other geo-entity mentions or real-world locations, e.g., そこ *soko* 'there' in Figure 1.

2.3 Coreference Annotation

In the coreference annotation step, we required the annotators to assign mention-level *specificity tags* or mention-pair-level *relations* to mentions identified in the previous step (except for those labeled with TRANS tags) using brat.

The criteria for coreference annotation define two types of specificity tags and two types of relations. As the representative cases, we introduce here the GENERIC specificity tag and the COREF coreference relation, and explain the remaining tags/relations in Appendix B. GENERIC is assigned to a generic mention, e.g., お寺 *otera* 'temples' in Figure 1, to distinguish singleton mentions that refer to real-world location, but are not coreferenced with other mentions. COREF is assigned to two mentions that both refer to the same real-world location, e.g., 近鉄奈良駅 *kintetsu nara eki* 'Kintetsu Nara Station' and そこ *soko* 'there' in Figure 1 ⟨1⟩. After relation annotation, a set of mentions that is sequentially connected through binary relations is regarded as one coreference cluster. A mention without any relations or specificity tags is regarded as a singleton, e.g., Figure 1 ⟨2⟩ and ⟨4⟩.[8]

2.4 Link Annotation

In the link annotation step, we required the annotators to link each coreference cluster to the URL of the corresponding OSM entry (e.g., ⟨1⟩–⟨4⟩ in Figure 1) on the basis of OSM and web search results. For URL assignment, the annotators added URLs to the cells representing coreference clusters in TSV files, which were converted from the brat output files.

The criteria for link annotation define the annotation flow as follows. For each coreference cluster, an annotator determines one or more normalized names of the referent location, e.g., formal or common name. The annotator then searches and assigns a URL of an appropriate OSM entry to the coreference cluster using search engines.[9]

The specific assignment process of entries is as follows. (a) If one or more candidate entries for a coreference cluster are found, assign the most probable candidate as BEST_URL and (up to two) other possible candidates as OTHER_URLS. (b) If the only candidate entry geographically includes but does not exactly match with the real-world referent, assign the found entry with the PART_OF tag. (c) If no candidate entries are found in OSM, search and assign an appropriate entry from alternative DBs: Wikidata, Wikipedia, and general web pages describing the real-world referent.[10] (d) If no candidate entries are found in any DBs, assign the NOT_FOUND tag instead of an entry URL. The annotators can skip the search steps and assign the NOT_FOUND tag when all member mentions and surrounding context do not provide any specific information that identifies the referent.

3 Dataset Statistics

The annotators first annotated 200 documents with mention information, then annotated the same 200 documents with coreference information, and finally annotated 100 of those documents with link information.[11] We call the latter 100 documents that contain link annotation Set-B and refer to the remaining 100 documents without link annotation as Set-A. The numbers of documents (#Doc), sentences (#Sent), words (#Word), mentions (#Men), and entities (coreference clusters) (#Ent) in the ATD-MCL are listed in Table 2. We used Mode B (the middle unit) of the SudachiPy tokenizer (ver-

[8] Although singleton mentions are marked with coreference cluster IDs in Figure 1 for clarity, singletons were not annotated with any coreference information in the actual work.

[9] Because it was sometimes difficult to find the desired entries using the Nominatim search engine available on the official OSM site, we asked the annotators to use additional search engines: web search engines and an original search engine that we developed.

[10] These auxiliary DBs enable referent information to be preserved in cases where the expected entries are not present in OSM (at the time of annotation).

[11] Different annotators could be assigned for each step.

	#Doc	#Sent	#Word	#Men	#Ent
Set-A	100	5,949	85,741	6,052	3,131
Set-B	100	6,324	87,074	6,119	3,208
Total	200	12,273	172,815	12,171	6,339

Table 2: Statistics of the ATD-MCL.

sion 0.6.7)[12] (Takaoka et al., 2018) for counting the number of words in the Japanese text.

Detailed statistics of our dataset are described in Appendix C. The notable characteristics are summarized below. (1) Facility mentions account for 50.3% (6,090/12,114) and nominal or demonstrative expressions account for 48.4% (5,867/12,114) of geo-entity mentions. (2) Multi-member clusters account for 35.6% (2,256/6,339) of coreference clusters, and the average number of member mention text types (distinct strings) for the multi-member clusters is 1.85, suggesting that the same geo-entity is often repeatedly referred to by different expressions in a document. (3-i) Geo-entities assigned with some URLs account for 97.1% (1,942/2,001) of entities with NAME mentions ("HasName" entities) and 50.5% (609/1,207) of the remaining entities (in the PART_OF-inclusive setting), suggesting that identifying the referents that are not clearly written in text is difficult even for humans. (3-ii) Geo-entities assigned with OSM entry URLs account for 75.7% (1,514/2,001) of all "HasName" entities and 74.0% (811/1,096) of "HasName" facility entities (in the PART_OF-exclusive setting), indicating that OSM has reasonable coverage of various types of locations in Japan.

4 Inter-Annotator Agreement

For mention, coreference, and link annotation, we requested two annotators to independently annotate the same 10, 10, and 5 documents out of the 200, 200, and 100 documents, respectively.[13] We measured the inter-annotator agreement (IAA) for the three annotation tasks.

4.1 Mention Annotation

As the IAA measure for mention annotation, we calculated the F1 scores between the results of two annotators (W1 and W2), based on exact match of both spans and tags. Table 3 shows the F1 score for each tag set and the numbers of annotated mentions by W1, W2, and both (M).

Tag set	Token				Type	
	F1	#W1	#W2	#M	#W1	#W2
NAME	0.835	229	243	197	162	174
NOM	0.867	195	197	170	97	106
L_O_F	0.552	19	10	8	8	5
DEICT	0.621	19	10	9	6	3
L_ORG	–	0	0	0	0	0
F_ORG	0	1	0	0	1	0
All	0.832	463	460	384	274	283

Table 3: IAA for mention annotation. NAME, NOM, L_O_F, DEICT, L_ORG, and F_ORG indicate all NAME mentions, all NOM mentions, LOC_OR_FAC, DEICTIC, LOC_ORG, and FAC_ORG, respectively. The token and type columns indicate the scores and numbers based on token and type frequencies of mention text, respectively.

The F1 score for all mentions was 0.832. Higher F1 score for NOM mentions (0.867) than that for NAME mentions (0.835) is probably because the less variety of NOM mention text types eased the annotation work for those mentions, as suggested by the mention token/type frequencies in Table 3.

4.2 Coreference Annotation

To assess IAA for COREF relation annotation, we used the metrics commonly used in coreference resolution studies: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF$_e$ (Luo, 2005), and the average of the three metrics (a.k.a the CoNLL score) (Pradhan et al., 2012).

Table 4 shows the F1 scores between two annotators' (W1 and W2) results for each IAA measure and the numbers of clusters constructed from two annotators' results for 2×2 settings: (a) original coreference clusters with all mentions or (b) clusters where only NAME mentions are retained, and (i) clusters with size ≥ 1 or (ii) clusters with size ≥ 2. In the basic setting (a)-(i), the average F1 score was 0.858. In addition, we observed two intuitive results. One is the lower scores for (a) than for (b), indicating that it was difficult to identify which mentions coreferred with non-NAME mentions. The other is the higher scores for (i) than for (ii); this is because leaving mentions as singletons is more likely to agree, since each mention is a singleton by default.

4.3 Link Annotation

As the IAA measure for link annotation, we calculated the F1 score of OSM (or other DB) en-

[12] https://github.com/WorksApplications/SudachiPy

[13] For coreference annotation, 10 documents annotated by two annotators did not include any mentions with specificity tags or mention pairs with attributive coreference relations.

	#W1/#W2	MUC	B^3	$CEAF_e$	Avg.
(a) Original clusters with all mentions					
(i)	237/297	0.913	0.878	0.782	0.858
(ii)	91/79	0.797	0.768	0.811	0.792
(b) Clusters only with NAME mentions					
(i)	237/297	0.959	0.935	0.893	0.929
(ii)	91/79	0.912	0.868	0.844	0.874

Table 4: IAA between two annotators for coreference clusters in coreference annotation. The top two rows (a) and the bottom two rows (b) show the results in the described settings. (i) and (ii) show the results in the settings where singletons are included or not, respectively.

	#W1/#W2	(a) Original		(b) Grouped	
		#M	F1	#M	F1
URL	81/75	56	0.718	64	0.821
NF	16/22	14	0.737	14	0.737
All	97/97	70	0.722	78	0.804

Table 5: IAA between two annotators for link annotation in (a) the original URL and (b) the grouped URL settings. The "URL" and "NF" rows show the results for the assigned URLs and NOT_FOUND tag, respectively.

try URL assignment for the same entities between two annotators (W1 and W2), which is similar to cluster-level hard F1 score (Zaporojets et al., 2022).

Table 5 shows the F1 scores along with the numbers of entities to which URLs or the NOT_FOUND tags were assigned by W1, W2, and both (M).[14] We used two settings about the equivalence for assigned URLs. (a) The original URL setting compares raw URL strings assigned by the annotators. (b) The grouped URL setting treats OSM entries or web pages representing practically the same locations as the same and compares the grouped URL sets instead of original URLs.[15]

The F1 scores for URLs and NOT_FOUND were over 0.7 in both settings, indicating that the annotator could assign the same URL (or the NOT_FOUND tag) to the majority of geo-entities in spite of the huge number of candidate URLs. The lower F1 scores in (a) the original setting than those in (b) the grouped setting is because the annotators assigned different but practically equivalent entry URLs to eight entities.

5 Experiments

We conducted experiments on the ATD-MCL for three tasks: MR, CR, and ED. The purpose of the experiments is to clarify the performance level of current entity analysis systems, including off-the-shelf and finetuned models, on our dataset.

5.1 Data Split

We regarded all Set-A documents as train-a and split the Set-B documents into train-b, development, and test sets at a ratio of 1:1:8. The union of train-a and train-b (110 documents) was used as the training set for both MR and CR. The development set (10 documents) and test set (80 documents) were commonly used for the three tasks.

5.2 Database Reorganization

The original OSM data contains a huge number of entries, and multiple entries can refer to almost the same real-world locations; e.g., we found 72 entries named 東京 'Tokyo,' including multiple railway station platforms and train stop positions, some of which can be equated with each other. For practical evaluation of ED systems, different entries that can be treated as equivalent should be grouped together, and such groups should be considered as linking units rather than individual entries.

Therefore, we reorganized the raw OSM data as follows. (1) We downloaded an OSM data file consisting of Japanese domestic location entries.[16] (2) We extracted 2.8M entries with "name" attributes from the total of 2.6B entries. (3) We added 14 out of 16 entries without name attributes that were assigned to domestic geo-entities in the Set-B data, but were not contained in the extracted entries (the remaining two entries had been deleted from OSM). This resulted in DB coverage of 99.86% for the Set-B entities annotated with OSM URLs. (4) We then generated an *extended name* from the original name attribute for each entry by concatenating part of the address and notable OSM tags, such as the branch name and amenity type.[17] (5) Finally, we grouped entries with the same extended name into the same entry group. This series of processes resulted in 1.8M entry groups.

[14] We regarded an entity as a matched URL instance when both annotators assigned the same URL and as a matched NOT_FOUND instance when both annotators assigned NOT_FOUND.

[15] The first author manually judged the practical equivalence of different OSM entries and web pages for 34 entities unmatched between two annotators.

[16] japan-230601.osm.bz2 (http://download.geofabrik.de/asia/)

[17] An example extended name: "name=スターバックス|branch=None|prefecture=奈良県|city= 奈良市|quarter=樽井町|road=猿沢遊歩道|amenity=cafe" (Starbucks Coffee at Sarusawa pathway, Tarui-cho, Nara City, Nara Prefecture).

5.3 Mention Recognition

Task Setting We treat MR as the task of identifying spans and entity types of geo-entity mentions in given documents. As the evaluation measure, we use the F1 score between the gold and predicted mentions based on exact match of both spans and entity types.

Systems We evaluated two systems that we finetuned models on our training set (spaCy-MR and mLUKE-MR) and two off-the-shelf systems without model finetuning (KWJA and GiNZA). spaCy-MR indicates a transition-based parsing model on the spaCy NLP library[18] that we built using a pretrained Japanese ELECTRA (Clark et al., 2020) model.[19] This corresponds to the finetuned version of the GiNZA model. mLUKE-MR is our implementation of a span-based MR system using a pretrained multilingual LUKE (mLUKE) (Ri et al., 2022) model.[20] As the off-the-shelf systems, we used KWJA "base" (version 2.1.1)[21,22] (Ueda et al., 2023) and GiNZA "ja_ginza_electra" (version 5.1.2). GiNZA and KWJA follow the ENE and IREX (Sekine and Isahara, 2000) tag sets, which are different from ours. Thus, we applied tag conversion rules to their outputs. Because the LOCATION tag in IREX semantically includes LOC_NAME, FAC_NAME, and LINE_NAME tags, we converted each KWJA output mention with the LOCATION tag into three mention instances with the same span and with one of the three tags, which prioritizes recall over precision. More detailed settings are described in Appendix D.

Results Table 6 shows the performance of the MR systems for the test set. GiNZA and KWJA achieved the recall of 0.55–0.70 for NAME mentions, indicating moderate coverage for named geo-entity mentions. However, the two systems failed to extract non-NAME mentions (the F1 scores were 0), which is natural because these systems had been trained on only NE annotations (not nominal phrases). Owing to our finetuning, spaCy-MR and mLUKE-MR improved the performance: the overall F1 scores of 0.74–0.82. Both finetuned models achieved better performance for NOM mentions than for NAME mentions, indicating the difficulty of recognizing the NAME mentions with more diverse surfaces. For the fine-grained results for each tag, see Appendix E.

System	Tag	P	R	F
KWJA	Overall	.279	.352	.311
	NAME	.279	.695	.398
GiNZA	Overall	.574	.277	.374
	NAME	.574	.548	.560
spaCy-MR	Overall	.752	.732	.742
	NAME	733.	.719	.726
	NOM	798.	.763	.780
mLUKE-MR	Overall	**.813**	**.817**	**.815**
	NAME	.828	.813	.821
	NOM	.832	.826	.829

Table 6: System performance for mention recognition. NAME and NOM indicate the micro-averaged scores for the entity types with NAME and NOM subtypes, respectively.

5.4 Coreference Resolution

Task Setting We define CR as the task of clustering the given gold mentions that corefer the same real-world locations. We use the same evaluation metrics as the IAA measures.

Systems We evaluated one finetuned system (mLUKE-CR), one off-the-shelf system (KWJA), and two rule-based systems (Rule-CR-1 and 2). mLUKE-CR is our implementation of an end-to-end CR model based on a pretrained mLUKE model,[23] which identifies the antecedent for a given mention following Lee et al. (2017). We used the KWJA 'base' model and applied a modification rule to the KWJA's output clusters so that the union of all output clusters matched the set of all gold mentions.[24] Simple rule-based systems are as follows. Rule-CR-1 treats all given mentions as singletons. Rule-CR-2 groups together sets of mentions with the same surface form in a document into clusters and treats the remaining mentions as singletons.

Results Table 7 shows the performance of the CR systems for the test set. The simplest rule-based system, Rule-CR-1, appears to have achieved the

[18] https://spacy.io/api/architectures#parser

[19] https://huggingface.co/megagonlabs/transformers-ud-japanese-electra-base-discriminator

[20] https://huggingface.co/studio-ousia/mluke-large-lite

[21] https://github.com/ku-nlp/kwja

[22] There was no KWJA documentation describing how to train a custom model, and we attempted but failed to perform training/finetuning.

[23] https://huggingface.co/studio-ousia/mluke-large

[24] The modification rule removes predicted mentions that do not match any gold mentions from the output clusters and adds gold mentions that do not match any predicted mentions as singletons on the basis of mention span overlapping.

System	Size	MUC	B^3	CEAF$_e$	Avg.
Rule-CR-1	≥ 1	0	.755	.639	.465
	≥ 2	0	0	0	0
Rule-CR-2	≥ 1	.622	.840	.790	.750
	≥ 2	.622	.613	.629	.621
KWJA	≥ 1	.694	.839	.793	.775
	≥ 2	.694	.661	.658	.671
mLUKE-CR	≥ 1	**.753**	**.875**	**.839**	**.822**
	≥ 2	**.753**	**.733**	**.737**	**.741**

Table 7: System performance for coreference resolution.

System	R@1	R@5	R@10	R@100
Rule-ED	.221	.323	.345	.362
BERT-ED	.219	.366	.399	.482

Table 8: System performance for entity disambiguation.

moderate B^3 and CEAF$_e$ scores for clusters with size≥ 1 (although resulted in the zero score for the link-based MUC metric), due to the dataset distribution biased toward a high population of singletons. Thus, it is necessary to pay attention to the improvement from these baseline scores as meaningful performance evaluation measures. Another rule-based system, Rule-CR-2, achieved the scores of 0.61–0.84 for the three metrics, indicating that the simple heuristic regarding surface forms was a strong clue for finding coreferent mentions. The superior performance of KWJA and mLUKE-CR over Rule-CR-2 indicates that these two systems identified (part of) coreferent mentions with different surface forms, although mLUKE-CR expectedly performed better owing to finetuning.

5.5 Entity Disambiguation

Task Setting We define ED as the task of selecting appropriate extended names, i.e., entry group IDs, from all entry groups for each given geo-entity. As the evaluation measure, we use recall@k (R@k) for the given entities; the prediction is regarded as correct if one of the predicted k entity groups contains the gold OSM entry URL for each geo-entity.

Systems We evaluated an unsupervised system (BERT-ED) and a rule-based system (Rule-ED). For an input entity, both systems regard the longest mention surface among its member mentions with NAME entity subtype tags as the entity name and predict DB entry groups based on the entity name. The systems return no entry groups if the entity contains no NAME mentions. BERT-ED is our implementation of an ED system based on a pretrained Japanese BERT (Devlin et al., 2019) model.[25] BERT-ED calculates the similarity between each entity's name and "name" attribute value of each candidate entry group, and then ranks the candidates. For the similarity score, we used the cosine similarity score between vector representations, that is, the average of hidden states at the last layer for input words within the name string.[26] Rule-ED extracts entry groups whose "name" attribute values exactly match the entity's name for each given entity, and then ranks them in lexicographic order of full extended names.

Results Table 8 presents the performance of the ED systems for the test set. Overall, BERT-ED achieved better scores than Rule-ED owing to soft matching and ranking using vector representations. In particular, BERT-ED outperformed Rule-ED by a larger margin on R@k with larger k. Although this result suggests the effectiveness of vector representations, the performance for R@1 can be improved by introducing more sophisticated disambiguation strategies that consider the geography-related content in a document, including location and facility types identified from the surrounding context, and geographic areas mentioned within the document.

5.6 Discussion

For MR and CR, the finetuned systems achieved the reasonable performance in our experiments. For ED, in contrast, the simple unsupervised systems did not achieve practical performance. A possible solution is training supervised ED systems on in-domain training data. However, we suppose that predicting appropriate DB entries for unknown instances would remain a main challenge due to limits to improving coverage by increasing training instances.

Another challenge in geographic ED is that natural language descriptions of geo-DB entries are unavailable, different from general DBs represented by Wikipedia. This also makes it difficult to directly apply state-of-the-art general ED systems using entry description text (Wu et al., 2020; Yamada et al., 2022) to geographic ED. Instead, OSM

[25] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[26] We also tried an entity representation calculated from the full sentence where its representative mention occurred, but confirmed its poor performance on the development set.

Dataset Name		Lang	Text Genre	Geo-database	Facility	Nominal
LGL Corpus	(Lieberman et al., 2010)	en	News	GeoNames	✗	✗
TR-News	(Kamalloo and Rafiei, 2018)	en	News	GeoNames	✗	✗
GeoVirus	(Gritta et al., 2018a)	en	News	Wikipedia	✗	✗
WikToR	(Gritta et al., 2018b)	en	Wikipedia	Wikipedia	✗	✗
GeoCorpora	(Wallgrün et al., 2018)	en	Microblog	GeoNames	△	✗
GeoWebNews	(Gritta et al., 2020)	en	News	GeoNames	✗	✓
CLDW	(Rayson et al., 2017)	en	Historical	Unlock	✗	✗
LRE Corpus	(Matsuda et al., 2017)	ja	Microblog	ISJ & Original	△	✓
ATD-MCL	(Ours)	ja	Travelogue	OpenStreetMap	✓	✓

Table 9: Characteristics of representative geoparsing datasets and ours. The facility and nominal columns show the availability of geoparsed facility mentions and nominal mentions, respectively: ✓ (available), ✗ (not available), and △ (available to a limited extent).

entries have rich information of semantic attributes and geographic relations, such as distance and hierarchy. A prospective direction is learning mention/entry representations that leverage or encode such geographic information, as well as entity type and population information (Zhang and Bethard, 2023). For example, if some geographic relations between two mentions are indicated by calculation based on their representations, geo-entities referred to by them may also have similar relations, which would be useful for CR and ED.

6 Related Work

Entity Analysis Datasets For over two decades, efforts have been devoted to developing annotated corpora for English entity analysis tasks, including NER (Tjong Kim Sang, 2002; Ling and Weld, 2012; Baldwin et al., 2015), anaphora/coreference resolution (Grishman and Sundheim, 1996; Doddington et al., 2004; Pradhan et al., 2011; Ghaddar and Langlais, 2016), and ED/EL (McNamee et al., 2010; Hoffart et al., 2011; Ratinov et al., 2011; Rizzo et al., 2016). For Japanese text, annotated corpora have been developed for general NER (Sekine et al., 2002; Hashimoto and Nakamura, 2010; Iwakura et al., 2016), coreference resolution (Kawahara et al., 2002; Hashimoto et al., 2011; Hangyo et al., 2014), and EL (Jargalsaikhan et al., 2016; Murawaki and Mori, 2016).

Geoparsing Datasets Table 9 summarizes the characteristics of representative geoparsing datasets and the ATD-MCL. For English geoparsing, annotated corpora have been developed and used as benchmarks for system evaluation. The Local Global Corpus (Lieberman et al., 2010), TR-News (Kamalloo and Rafiei, 2018), and GeoWebNews (Gritta et al., 2020) contain approximately 100–600 news articles from global and local news sources. GeoVirus (Gritta et al., 2018a) comprises 229 WikiNews articles focusing on viral infections. The SemEval-2019 Task 12 dataset (Weissenbacher et al., 2019) comprises 150 biomedical journal articles on the epidemiology of viruses. GeoCorpora (Wallgrün et al., 2018) comprises 1,639 tweets with the very limited coverage of facility mentions. The Corpus of Lake District Writing (CLDW) (Rayson et al., 2017) consists of 80 historical texts, including travelogues, with auto-annotated coordinates of location mentions. For Japanese geoparsing, Matsuda et al. (2017) constructed the LRE corpus, comprising 10,000 Japanese tweets, 793 of which have geo-entity-related tags. They used Ichi Sansho Joho (ISJ) ‘City-block-level location reference information’ and their original gazetteer of facilities, but the latter gazetteer has not been available due to licensing reasons.

7 Conclusion

This paper has described the ATD-MCL dataset, which is designed for document-level geoparsing, along with the annotation criteria, IAA assessment, and performance evaluation of the baseline systems. Our dataset enables other researchers to conduct reproducible experiments through the public release of our annotated data. We expect that our dataset contributes to fostering future research and advancing geoparsing techniques.

In future work, we plan to (1) develop a document-level geoparser that leverages both characteristics of geo-entity mentions in text and geo-DB entries, (2) enhance our dataset with additional semantic information, such as the movement trajectories of travelogue writers, for more advanced analytics, and (3) construct annotated travelogue datasets in other languages.

Limitations

Bias of Referent Locations of Mentions in the Dataset Since our dataset only comprises Japanese domestic travelogues, most of the mentions refer to locations in Japan. This is because we prioritized increasing the coverage of locations in a specific region, i.e., Japan. However, a possible extension is annotating overseas travelogues in the ATD, which include many mentions referring to locations around the world. It would supplement our current dataset.

Optimization of Database Reorganization As the reorganized DB for ED, we used 2.8M OSM entries of Japanese domestic locations with "name" attributes. While checking a portion of the generated entry groups, we performed rule engineering to make the original DB more desirable for our ED task, which means entries that can be regarded as practically equivalent to each other belong to the same groups. Over- and under-aggregated groups in the final DB could produce the evaluation results with underestimated or overestimated system performance. This would have a greater influence on the recall@k scores with smaller k for evaluating disambiguation accuracy, but a lesser influence on the scores with larger k for evaluating extraction coverage.

Optimization of System Performance We performed not systematic but minimum hyperparameter search for mLUKE-based models due to time and resource limitations. Similarly, we used the fixed hyperparameters for spaCy-MR, which correspond to those used for GiNZA. Thus, performing optimized experiments has potential for further performance improvement in these systems.

Independent Experiments on Geoparsing Subtasks As a first step toward comprehensive evaluation of geoparsing techniques, we independently evaluated the baseline systems on each subtask in the gold input setting; that is, gold mention spans were given in the CR experiments and gold entities were given in the ED experiments. However, it is also necessary to explore developing and evaluating more practical systems in the full geoparsing setting, which requires systems to predict mentions, coreference clusters, and links from raw documents.

Ethics Statement

As a potential risk associated with our dataset, a model trained on the dataset has the ability, to some extent, to identify locations mentioned in input texts and could be applied to link the content of individual posts containing private information with the mentioned locations. In addition, regardless of the purpose of use, the predicted locations may be inaccurate due to the limitations of the model's performance or the discrepancy of domains, writing styles, and mentioned regions between our dataset and input texts.

Consistently with their intended use, we used existing language resources and tools to develop or evaluate NLP datasets or models under the specified license or terms of use. As for the dataset that we constructed, its intended use is for academic research purposes related to information science, similarly to that of the ATD. The text in our dataset is a subset of the original ATD data, and the original data does not contain any information about the travelogue authors. Before commencing the annotation work to construct our dataset, we explained to the annotators that we or other researchers would use the annotated data for future research related to NLP.

References

Arukikata. Co., Ltd. 2022. Arukikata travelogue dataset. Informatics Research Data Repository, National Institute of Informatics. https://doi.org/10.32130/idr.18.1.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In The first international conference on language resources and evaluation workshop on linguistics coreference, volume 1, pages 563–566.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In Proceedings of the Workshop on Noisy User-generated Text, pages 126–135, Beijing, China. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In 8th International Conference on Learning Representations, Addis Ababa, Ethiopia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA).

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018a. Which Melbourne? augmenting geocoding with maps. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1285–1296, Melbourne, Australia. Association for Computational Linguistics.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A pragmatic guide to geoparsing evaluation: Toponyms, named entity recognition and pragmatics. Language resources and evaluation, 54:683–712.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018b. What's missing in geographical parsing? Language Resources and Evaluation, 52:603–623.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2014. Building and analyzing a diverse document leads corpus annotated with semantic relations. Journal of Natural Language Processing, 21(2):213–247.

Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato, and Masaaki Nagata. 2011. Construction of a blog corpus with syntactic, anaphoric, and sentiment annotations. Journal of Natural Language Processing, 18(2):175–201.

Taiichi Hashimoto and Shun'ichi Nakamura. 2010. Kakuchō koyū hyōgen tag tsuki corpus-no kōchiku—hakusho, shoseki, Yahoo! chiebukuro core data—(Construction of an extended named entity-annotated corpus—white papers, books, Yahoo! chiebukuro core data). In Proceedings of the 16th Annual Meeting of the Association for Natural Language Processing.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tomoya Iwakura, Kanako Komiya, and Ryuichi Tachibana. 2016. Constructing a Japanese basic named entity corpus of various genres. In Proceedings of the Sixth Named Entity Workshop, pages 41–46, Berlin, Germany. Association for Computational Linguistics.

Davaajav Jargalsaikhan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. 2016. Building a corpus for Japanese wikification with fine-grained entity classes. In Proceedings of the ACL 2016 Student Research Workshop, pages 138–144, Berlin, Germany. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

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.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Jochen L Leidner. 2006. An evaluation dataset for the toponym resolution task. Computers, Environment and Urban Systems, 30(4):400–417.

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.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In Proceedings of the 26th AAAI Conference on Artificial Intelligence.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Hiroshi Matsuda, Mai Omura, and Masayuki Asahara. 2019. Tantan’i hinshi-no yōhō aimaisē kaiketsu-to ison kankē labeling-no dōji gakushū (Simultaneous learning of usage disambiguation of parts-of-speech for short unit words and dependency relation labeling.). Proceedings of the 25th Annual Meeting of the Association for Natural Language Processing.

Koji Matsuda, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. 2017. Geographical entity annotated corpus of japanese microblogs. Journal of Information Processing, 25:121–130.

Paul McNamee, Hoa Trang Dang, Heather Simpson, Patrick Schone, and Stephanie M. Strassel. 2010. An evaluation of technologies for knowledge base population. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta, Malta. European Language Resources Association (ELRA).

Yugo Murawaki and Shinsuke Mori. 2016. Wikification for scriptio continua. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 1346–1351, Portorož, Slovenia. European Language Resources Association (ELRA).

Hiroki Ouchi, Hiroyuki Shindo, Shoko Wakamiya, Yuki Matsuda, Naoya Inoue, Shohei Higashiyama, Satoshi Nakamura, and Taro Watanabe. 2023. Arukikata travelogue dataset. arXiv:2305.11444.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In Joint Conference on EMNLP and CoNLL - Shared Task, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Paul Rayson, Alex Reinhold, James Butler, Chris Donaldson, Ian Gregory, and Joanna Taylor. 2017. A deeply annotated testbed for geographical text analysis: The corpus of lake district writing. In Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities, GeoHumanities’17, page 9–15, New York, NY, USA. Association for Computing Machinery.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Giuseppe Rizzo, Marieke van Erp, Julien Plu, and Raphaël Troncy. 2016. Making sense of microposts (#Microposts2015) named entity recognition and linking (NEEL) challenge. In Proceedings of the 6th Workshop on ’Making Sense of Microposts’, pages 50–59.

Satoshi Sekine and Hitoshi Isahara. 2000. IREX: IR & IE evaluation project in Japanese. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00), Athens, Greece. European Language Resources Association (ELRA).

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02), Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107, Avignon, France. Association for Computational Linguistics.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: a Japanese tokenizer for business. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002).

Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2023. KWJA: A unified japanese analyzer based on foundation models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Toronto, Canada. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.

Jan Oliver Wallgrün, Morteza Karimzadeh, Alan M MacEachren, and Scott Pezanowski. 2018. Geocorpora: building a corpus to test and train microblog geoparsers. International Journal of Geographical Information Science, 32(1):1–29.

Davy Weissenbacher, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2019. SemEval-2019 task 12: Toponym resolution in scientific papers. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 907–916, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6397–6407, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6442–6454, Online. Association for Computational Linguistics.

Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. Global entity disambiguation with BERT. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3264–3271, Seattle, United States. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Yiwei Jiang, Thomas Demeester, and Chris Develder. 2022. Towards consistent document-level entity linking: Joint models for entity linking and coreference resolution. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 778–784, Dublin, Ireland. Association for Computational Linguistics.

Zeyu Zhang and Steven Bethard. 2023. Improving toponym resolution with better candidate generation, transformer-based reranking, and two-stage resolution. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), pages 48–60, Toronto, Canada. Association for Computational Linguistics.

A Licenses of Used Resources

We used some existing NLP software and language resources as described in the main sections. The licenses of the used resources are as follows. The Arukikata Travelogue Dataset is available via the Informatics Research Data Repository, National Institute of Informatics under specific terms of use.[27] brat, spaCy, GiNZA, KWJA, the pretrained Japanese ELECTRA model are available under the MIT License. SudachiPy and the pretrained mLUKE models are available under the Apache License 2.0. The pretrained Japanese BERT model is available under CC BY-SA 4.0. Although the OpenStreetMap data files are available via Geofabrik[28] under the Open Database License 1.0, the data file that we used in our experiments `japan-230601.osm.bz2` is currently no longer available. We will rerelease the same data file on our website under the same license.

B Detailed Annotation Criteria

B.1 Mention Span Annotation

The spans of geo-entity mentions are determined as follows. Generally, a noun phrase (NP) in which a head h is modified by a nominal modifier m is treated as a single mention (Table 10-a). An appositive compound of two nouns n_1 and n_2 is treated as a single mention (Table 10-b) unless there is some expression (e.g., *no*-particle "の") or separator symbol (e.g., *tōten* "、") inserted between them. A common name is treated as a single mention even if it is not a simple NP (Table 10-c). For an NP with an affix or affix-like noun a representing directions or relative positions, a cardinal direction prefix preceding a location name is included in the span (Table 10-d-1), but other affixes are excluded from the span (Table 10-d-2). There may be instances in which a modifier m represents a geo-entity, but its NP head h does not. In such cases, the modifier is treated as a single mention if the head is a verbal noun that means move, stay, or habitation (Table 10-e-1), but the NP is not treated as a mention if not (Table 10-e-2). In the case that a geo-entity name g is embedded in a non-geo-entity mention n, the inner geo-entity name is treated as a geo-entity mention if the external entity corresponds to an event held in the real world (Table 10-f). If the external entity corresponds to other types of entities, such as an organization or the title of a work, the inner geo-entity name is not treated as a geo-entity mention.

(a)	[山頂]$_m$ [駐車場]$_h$ [parking area]$_h$ [on top of the mountain]$_m$
(b)	[駅ビル]$_{n_1}$ [「ビエラ奈良」]$_{n_2}$ [station building]$_{n_1}$ [Vierra Nara]$_{n_2}$
(c)	天国への階段 Stairway to Heaven
(d-1)	[東]$_a$ [東京] [East]$_a$ [Tokyo]
(d-2)	[北海道] [全域]$_a$ [the whole area of]$_a$ [Hokkaido]
(e-1)	[京都]$_m$ [旅行]$_h$ [Kyoto]$_m$ [Travel]$_h$
(e-2)	[三輪]$_m$ [そうめん]$_h$ [Miwa]$_m$ [somen noodles]$_h$
(f)	[[保津川]$_g$ 下り]$_n$ [[Hozugawa river]$_g$ boat tour]$_n$

Table 10: Examples of mention spans.

B.2 Coreference Annotation

Following (or concurrently with) specificity tag annotation, relations are assigned to pairs of mentions that have not been labeled with either specificity tag.

Specificity Tags Specificity tags can be either `GENERIC` or `SPEC_AMB`. `GENERIC` is assigned to a generic mention, as explained in §2.3. `SPEC_AMB` (which means "specific but ambiguous") is assigned to a mention that refers to a specific real-world location, but there is some ambiguity about the detailed area to which it refers, e.g., 海 *umi* in a sentence such as "You can see a beautiful sea from this spot."

Coreference Relations Coreference relations can be either the identical coreference relation `COREF` or the attributive coreference relation `COREF_ATTR`. The coreference relation `COREF` is assigned to two mentions that both refer to the same real-world location, as explained in §2.3. The directed relation `COREF_ATTR` is assigned to mention pairs in which one expresses the attribute of the other, either in appositive phrases or copular sentences. For example, a sentence in Figure 2 is annotated with `COREF_ATTR` relations from mention 2 to mention 1 and from mention 2 to mention 3. This

[27] https://www.nii.ac.jp/dsc/idr/arukikata/documents/arukikata-policy.html (in Japanese)

[28] http://www.geofabrik.de/data/download.html

^{1}世界遺産・^{2}白川郷は素敵な^{3}ところでした。

A ^{1}world heritage site, ^{2}Shirakawago was a nice ^{3}place.

Figure 2: Examples of attributive mentions.

	LOC	FAC	LINE	TRANS	GeoOther
NAME	2,289	3,239	462	257	–
NOM	861	2,851	582	666	–
Other	–	–	–	–	907
Total	3,150	6,090	1,044	923	907

Table 11: Tag distribution of geo-entity mentions in the whole dataset. "GeoOther" mentions consist of 372 LOC_OR_FAC and 535 DEICTIC mentions.

Size	1	2	3	4	5	6	$\geq$7
#Cls	4,083	1,278	507	240	103	58	70
#Typ	1.0	1.5	2.0	2.3	2.6	2.8	3.3

Table 12: Number of geo-entity coreference clusters (#Cls) and the average number of member mention text types (#Typ) for each size.

	LOC	FAC	LINE	MIX	UNK
Set-A	819	1,823	327	29	133
Set-B	852	1,819	370	22	145
Total	1,671	3,642	697	51	278

Table 13: Tag distribution of geo-entities.

schema is similar to that in WikiCoref (Ghaddar and Langlais, 2016).

Notably, no coreference relations are assigned to mentions whose referents geographically overlap but are not identical; e.g., 首都高速道路 *shuto kōsoku dōro* 'Metropolitan Expressway' and 湾岸線 *wangansen* 'Bayshore Route,' which have a whole–part relation.

C Detailed Dataset Statistics

C.1 Mention Annotation

In the mention annotation step, 12,171 mentions were identified; they consist of 12,114 geo-entity and 57 non-geo-entity mentions (23 LOC_ORG and 34 FAC_ORG mentions). Table 11 shows the distribution of geo-entity mentions for entity type tags. The tag distribution represents some characteristics of travelogue documents of our dataset. First, the documents contain the largest number of facility mentions, which is even more than the number of location mentions. Second, the documents also contain the similar number of non-NAME (5,867)[29] to NAME mentions (6,247).

C.2 Coreference Annotation

As a result of the coreference annotation step, 289 GENERIC mentions and 322 SPEC_AMB mentions along with 923 TRANS mentions were excluded from the coreference relation annotation. Out of the remaining 10,580 mentions, 6,497 mentions were annotated with one or more COREF and/or COREF_ATTR relations among other mentions, of which 350 mention pairs were annotated with COREF_ATTR relations. These mentions comprise coreference clusters with size ≥ 2, and the remaining 4,083 mentions correspond to singletons. Table 12 shows the number of clusters and the average number of mention text types (distinct strings) among members[30] for each cluster size. This indicates that 35.6% (2,256/6,339) of coreference clusters have more than one member; that is, multiple mentions in a document often refer to the same referent.

In addition, we automatically assign an entity type tag to each coreference cluster, i.e., entity, from the tags of its member mentions.[31] Table 13 shows the tag distribution of entities, which is similar to the tag distribution of mentions shown in Table 11.

C.3 Link Annotation

As shown in Table 14, in the link annotation step for Set-B, 79.5% (2,551) and 64.2% (2,059) of 3,208 entities have been annotated with any URLs and OSM entry URLs, respectively, including entities annotated with PART_OF tags. For "HasName" entities in which at least one member mention is labeled as NAME, any URLs and OSM entry URLs are assigned to 97.1% (1,942/2,001) and 78.7% (1,574/2,001) of them, respectively. This indicates that the real-world referents can be easily identified for most of the entities explicitly written with their names. For the remaining "HasNoName" entities,

[29] Non-NAME mentions include LOC_OR_FAC, and DEICTIC mentions, in addition to all NOM mentions.

[30] For example, for clusters $C_1 = \{$"Nara Station", "Nara Sta.", "Nara"$\}$ and $C_2 = \{$"Kyoto Pref.", "Kyoto", "Kyoto"$\}$, the numbers of distinct member mention strings are three and two, respectively, and their average is 2.5.

[31] (a) LOC, FAC, or LINE is assigned to an entity that the members' tags include only one of the three types and optionally include LOC_OR_FAC or DEICTIC. (b) UNK is assigned to an entity that all members' tags are LOC_OR_FAC or DEICTIC. (c) MIX is assigned to an entity that the members' tags include two or three of LOC, FAC, and LINE.

	All	HasRef	HasOSMRef
HasName	2,001	1,942	1,574
HasNoName	1,207	609	485
Total	3,208	2,551	2,059

Table 14: Numbers of Set-B entities that have names and/or references in the PART_OF-inclusive setting where entities assigned with PART_OF (along with URLs) are counted as instances of “Has(OSM)Ref.”

	All	HasRef	HasOSMRef
HasName	2,001	1,861	1,514
HasNoName	1,207	298	221
Total	3,208	2,159	1,735

Table 15: Numbers of Set-B entities that have names and/or referents in the PART_OF-exclusive setting where entities assigned with PART_OF (along with URLs) are NOT counted as instances of “Has(OSM)Ref.”

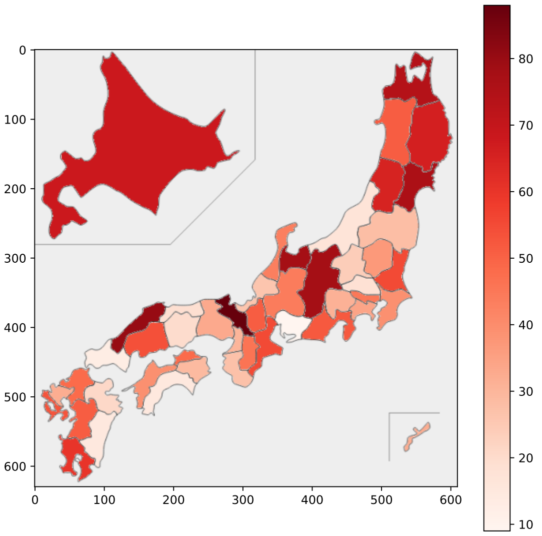


Figure 3: Numbers of linked entities located in each prefecture. Deeper red indicates the larger number. The units of the numerical values on the vertical and horizontal axes of the map are kilo-miles.

any URLs and OSM entry URLs are assigned to 50.5% (609/1,207) and 40.2% (485/1,207) of them, respectively. This suggests that identifying the referents from unclearly written mentions and context is difficult even for humans.

As shown in Table 15, the percentages of referent-identified entities decrease in the setting where entities assigned with PART_OF are excluded. The result indicates the reasonable coverage of OSM for various types of locations in Japan. Overall, entities assigned with OSM entries account for 75.7% (1,514/2,001) of “HasName” entities. For details on each entity type tag of LOC, FAC, LINE, and the others, entities assigned with OSM entries account for 79.3% (811/1,096), 74.0% (544/686), 72.7% (144/198), and 71.4% (15/21) of “HasName” entities with the specified tag, respectively.

C.4 Geographical Distribution of Linked Entities

Figure 3 shows the geographical distribution of linked entities in our dataset, namely, the number of entities located in each prefecture among entities annotated with OSM entry URLs. For example, there are 45 linked entities to which the coordinates of OSM entries are linked within the area of Tokyo Prefecture in all annotated travelogue documents, and thus the count of Tokyo Prefecture is 45. The minimum, maximum, and average numbers of entity counts in all 47 prefectures are 9 (Aichi), 88 (Kyoto), and 42.8, respectively.

Figure 4 shows actual examples of mentions with *geographic continuity*; that is, mentions that refer to nearby locations in the real world tend to appear near to one another within a document (§1). The example text in a travelogue document, whose ID is 00019, describes five geo-entities located nearby in the real world.

D Details on Experimental Settings

D.1 Evaluation Scripts

We used our code that calculates general precision, recall, and F1 score in the mention recognition and entity disambiguation experiments. We used our code that calculates the MUC, B^3, and $CEAF_e$ scores in the manner equivalent to an existing evaluation tool[32] in the coreference resolution experiments.

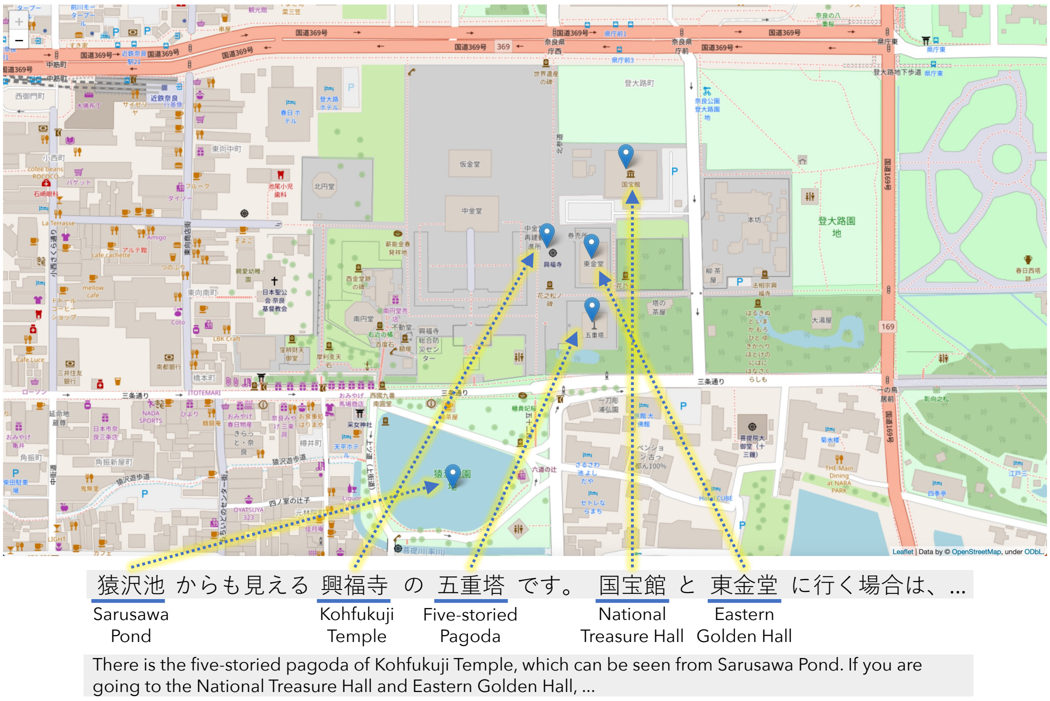
D.2 Entity Type Conversion Rules

IREX We used the following rules to convert the IREX tags to our entity type tags. (1) Each output mention with the LOCATION tag was converted into three mention instances with the same span and with one of LOC_NAME, FAC_NAME, and LINE_NAME tags. (2) ARTIFACT was converted into TRANS_NAME.

ENE We used the following rules to convert the ENE tags (version 7.1.0),[33] which GiNZA adopted,

[32] https://github.com/ns-moosavi/coval/blob/master/coval/eval/evaluator.py

[33] https://nlp.cs.nyu.edu/ene/version7_1_0Beng.html



猿沢池 からも見える 興福寺 の 五重塔 です。 国宝館 と 東金堂 に行く場合は、…

There is the five-storied pagoda of Kohfukuji Temple, which can be seen from Sarusawa Pond. If you are going to the National Treasure Hall and Eastern Golden Hall, …

Figure 4: Example mentions with *geographic continuity* in a travelogue document. The map depicts part of the Nara Park area, a popular sightseeing area in Nara City, Japan.

to our entity type tags. (1) The `Location` subtype tags except for the `Astral_Body` subtype tags, the `Address` subtype tags and `River` were converted to `LOC_NAME`. (2) The `Facility` subtype tags except for the `Line` subtype tags were converted to `FAC_NAME`. (3) `River` and the `Line` subtype tags were converted to `LINE_NAME`. (4) `Service` and the `Vehicle` subtype tags were converted to `TRANS_NAME`.

D.3 Settings of spaCy-MR

For building our custom MR model with spaCy, namely, spaCy-MR, we used almost the same settings as GiNZA,[34] including model architecture and hyperparameters, tokenizer, and training settings except that we disabled unnecessary pipelines other than "transformer" and "ner." We reported the result of a single run of spaCy-MR in §5.3 and Appendix E.

D.4 Implementation and Settings of mLUKE-MR/CR

We reported the results of single runs of mLUKE-MR and mLUKE-CR in §5.3 and Appendix E.

Mention Recognition Following Yamada et al. (2020), we tackle the task by enumerating and classifying all possible spans in each sentence. The representation of each candidate span is a concatenation of the word representations of the first and last tokens of the span, and the entity representation corresponding to the span, all of which are computed by the LUKE Transformer model. We employ a linear classifier to classify spans into the target entity types or *non-entity* type. We restrict candidate spans to the positions where their first and last tokens correspond to word boundaries (obtained using Sudachi Mode B), and exclude spans longer than 16 tokens.[35] Following Devlin et al. (2019) and Yamada et al. (2020), we prepend/append the surrounding tokens to a target sentence (up to 512 tokens in total) to give sufficient contextual information to the model.

Coreference Resolution Following Lee et al. (2017), we solve the task as antecedent identification for each mention. We follow the architecture proposed by Joshi et al. (2019) except that we do not use a unary score for each mention or coarse-to-fine inference because gold mentions are given in our setting.[36] The representation of each mention

[34] `https://github.com/megagonlabs/ginza/blob/develop/config/ja_ginza_electra.cfg`

[35] We also enforce word boundaries on the mLUKE tokenizer because (word-level) mention annotation in the ATD-MCL does not align with unigram segmentation used in the tokenizer.

[36] We also omit discrete features based on the metadata available only in some datasets.

Task	Name	Value
MR	Learning rate	1e-5
	Batch size	8
	Training epochs	10
CR	Learning rate	5e-5
	Batch size	4
	Training epochs	20
(Common)	Learning rate decay	linear
	Warmup ratio	0.06
	Dropout	0.1
	Weight decay	0.01
	Gradient clipping	none
	Adam β_1	0.9
	Adam β_2	0.98
	Adam ϵ	1e-6

Table 16: Hyperparameter values used in the mLUKE-MR/CR experiments.

is computed in the same way as the MR model. The model is trained by optimizing the marginal log-likelihood of the possibly correct antecedents including a dummy antecedent, which indicates no antecedents associated with a target mention. Because CR in the ATD-MCL is a document-level task and documents in the dataset are too long to be processed by a Transformer-based model for computational reasons, we independently feed each sentence in a document to the LUKE model, but optimization/prediction is made in each document.

Hyperparameters The hyperparameter values used in the experiments using mLUKE-MR/CR are listed in Table 16. Because our computational resources were limited, we did not conduct hyperparameter tuning except learning rate. We chose the best setting of learning rate and the number of training epochs from the search space of {1e-5, 2e-5, 3e-5, 4e-5, 5e-5} and {5, 10, 20}, respectively. We specifically selected batch size for each task, but we followed Yamada et al. (2020) for the other hyperparameters.

D.5 Size of Used Models

Table 17 shows the numbers of model parameters in the systems that we used in the experiments. For KWJA, we report the number of parameters (112M) in the pretrained model[37] used in the KWJA base model (while the actual number of parameters in the whole model would be larger).

Tasks	System	#Params
MR	mLUKE-MR	561M
MR	spaCy-MR	109M
MR	GiNZA (ja_ginza_electra)	110M
MR, CR	KWJA (base)	112M+
CR	mLUKE-CR	877M
ED	BERT-ED	111M

Table 17: Numbers of model parameters in evaluated systems.

D.6 Computational Budget for Finetuning

In our experiments, mLUKE-MR was finetuned for 130 minutes (10 epochs) using four NVIDIA Tesla V100 GPUs with 16GB memory. mLUKE-CR was finetuned for 15 minutes (20 epochs) using four NVIDIA A100 Tensor Core GPUs with 40GB memory. spaCy-MR was finetuned for 17.4 hours (20000 steps) using a four-core Intel Xeon Gold 6150 CPU (32 cores total).

E Detailed Experimental Results on Mention Recognition

Table 18 shows detailed performance of mention recognition systems. The finetuned systems spaCy-MR and mLUKE-MR achieved F1 scores higher than 0.6 and 0.7, respectively, for all tags except for TRANS_NAME and FAC_ORG.

[37] https://huggingface.co/ku-nlp/deberta-v2-base-japanese

Tag	#	KWJA			GiNZA			spaCy-MR$^{\circ}$			mLUKE-MR$^{\circ}$		
		P	R	F	P	R	F	P	R	F	P	R	F
Overall	4,958	.279	.352	.311	.574	.277	.374	.752	.732	.742	**.813**	**.817**	**.815**
NAME	2,509	.279	.695	.398	.574	.548	.560	.733	.719	.726	.828	.813	.821
NOM	2,054	0	0	0	0	0	0	.798	.763	.780	.832	.826	.829
LOC_NAME	881	.378	.857	.525	.617	.717	.664	.727	.822	.771	.830	.863	.846
FAC_NAME	1,285	.409	.635	.497	.589	.504	.543	.770	.689	.727	.843	.807	.825
LINE_NAME	195	.061	.621	.110	.425	.405	.415	.673	.677	.675	.804	.800	.802
TRANS_NAME	148	.193	.358	.251	.176	.101	.129	.525	.432	.474	.707	.588	.642
LOC_NOM	349	0	0	0	0	0	0	.739	.691	.714	.748	.808	.777
FAC_NOM	1,135	0	0	0	0	0	0	.816	.757	.785	.855	.819	.837
LINE_NOM	236	0	0	0	0	0	0	.749	.822	.784	.865	.818	.841
TRANS_NOM	334	0	0	0	0	0	0	.840	.817	.829	.830	.877	.853
LOC_OR_FAC	149	0	0	0	0	0	0	.676	.617	.646	.731	.711	.721
DEICTIC	222	0	0	0	0	0	0	.645	.721	.681	.616	.896	.730
LOC_ORG	11	0	0	0	0	0	0	.750	.545	.632	.900	.818	.857
FAC_ORG	13	0	0	0	0	0	0	0	0	0	.500	.077	.133

Table 18: System performance for mention recognition. “◦” indicates the models finetuned on the ATD-MCL training set. “#” indicates the number of mentions for each tag in the test set.