Electronic Preprint for Journal of Information Processing Vol.25

Regular Paper

Geographical Entity Annotated Corpus of Japanese Microblogs*

Koji Matsuda[1,a)] Akira Sasaki[1] Naoaki Okazaki[1,2] Kentaro Inui[1]

Received: June 10, 2016, Accepted: October 8, 2016

Abstract: This paper addresses the issues in the task of annotating geographical entities on microblogs and reports the preliminary results of our efforts to annotate Japanese microblog texts. Unlike prior work, we aim at annotating not only geographical location entities but also facility entities, such as stations, restaurants and schools. We discuss (i) how to build a gazetteer of geographical entities with a sufficiently broad coverage, (ii) what types ambiguities that need to be considered, (iii) why the annotator tends to disagree, and (iv) what technical problems should be addressed to automate the task of annotating the geographical entities. All the annotation data and the annotation guidelines are publicly available for research purposes from our web site.

Keywords: corpus annotation, microblogs, natural language processing, location reference expressions

1. Introduction

The ability to analyze microblog texts according to a spatial or temporal axis has become increasingly important in recent years. For example, with Twitter, users can share knowledge of situations and sightings of events at a low cost, with much of the information being integrated in the form of natural language. If it were possible to anchor these posts (known as "tweets") to specific locations in the real world, this would benefit a wide variety of applications such as marketing, social surveys [1], disease monitoring [2], [3], and disaster response [4], [5], [6].

For example, with respect to natural disasters, such as the 2011 Tohoku earthquake, large amounts of information were posted on social networking services (SNS), and some of these posts offered information that could aid rescue operations.

In this paper, we discuss the language expressions that are used, in particular those representing a "specific location". For example, expressions that refer to a location (henceforth referred to as "location reference expressions", **LRE**) are often mentioned in such SNS posts, and if it were possible to associate a specific set of coordinates with an area (grounding), this text information could be transferred to a map. By mapping tweets posted during disasters on time and spatial axes, it would possible to gain an improved understanding of a disaster situation.

In this case, it seems that it would be possible to use GPS information that has been attached as metadata to tweets. However, whether GPS information is included in tweets is controlled by the user, in their client settings. It was reported in a recent study [4] that less than 1% of tweets have GPS information appended to them. LREs are expressed in natural languages in the tweet, and an analysis would make it possible to map the actual spatial entity. Even though there is a large demand for this kind of application, a corpus that annotates geographical entities to LREs in microblog texts does not currently exist.

In this paper, we report the results of the trial that was conducted with the aim of creating a corpus that annotates specific entity information with the coordinate information to LREs appearing in Japanese texts sampled from microblogs. We provide details as to how we made the decisions on the various design aspects, how we built the entity gazetteer, and how we defined the representation of the annotated target. In addition, we describe how the validity of the proposed schema was verified by having it annotated by multiple people and we describe the problems identified from the results of this verification.

As will be discussed later in this paper, not only location names, but also facility names often appear in microblog texts. We compiled a large (more than 5 million entries) gazetteer of locations and facility entities from data obtained from the Web, and managed to annotate about 40% of these entities (an eight-fold increase on previous work) with facility names for which the writer assumes a specific location.

Finally, we analyzed part of our corpus to enable us to discuss the technical problems that would need to be resolved to perform the grounding of LREs. The resulting corpus, documentation, and annotation guidelines are available on our web site[*1] and following DOI: https://doi.org/10.5281/zenodo.161645.

2. Related Work

Studies that automatically annotate location information according to text are basically divided into the following types: The first is **Document Geolocation**, that is, inferring the location information for the whole of the given text. A typical example of

[1] Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi 980–8579, Japan
[2] Artificial Intelligence Research Center (AIRC), AIST, Koto, Tokyo 135–0064, Japan
[a)] matsuda@ecei.tohoku.ac.jp

[*1] http://www.cl.ecei.tohoku.ac.jp/~matsuda/LRE_corpus/
[*] This paper is a revised and extended version of [7].

this form of research is the automatic annotation of location information in Wikipedia articles, or inferring the residency of a Twitter user. This approach is mainly used for supervised learning, with text converted to feature representation. However, it has been reported that this method does not work well on short documents such as tweets [8].

A contrasting approach assigns specific geographical entities by automatically analyzing LREs to identify information such as a toponym that appears in the text (**Geoparsing, Toponym Resolution**) [9]. Reference [10] proposed a supervised learning method by using an indirect supervision technique. Reference [11] proposed a gazetteer independent method by using density estimation techniques.

These studies were evaluated by using a reference corpus such as the TR-CoNLL [9] or LGL (Local-Global Lexicon) [12] corpus. However, these corpora are annotated only by location entities, and not by facility entities. In addition, existing corpora have mainly been compiled from the newspaper domain.

Our main aim is the analysis and mapping of social media text; therefore, we need to investigate the behavior of different toponym resolution methods on social media text. This prompted us to annotate text sampled from SNSs.

Reference [13] annotated location information to text, by annotating both the location and facility entities, but their corpus is sampled from the ACE corpus, which is drawn mainly from broadcast conversations and news magazines. However, in our investigation of their corpus, out of all the LREs in the expressions that were annotated, only 5% were tagged as "Facility", and these were only very popular entities such as "the Pentagon" and "the White House".

In contrast, as our corpus study reveals below, real-life microblog texts include as many mentions referring to facilities whose location can be uniquely identified as are mentions referring to location entities. The annotation of these facility-referring mentions poses interesting research challenges, which motived our corpus study reported in this paper.

Recently, Ref. [14] annotated Twitter messages, but their annotation focus is limited to toponyms, and facility names are not annotated. Examples of geoparsing for Japanese text, GeoNLP [15] exist, but there are no reports of quantitative evaluations of the performance, because there is no corpus for evaluation.

3. Pilot Categorization: Why Do We Focus on Individual LREs?

In this work, we focused on the identification and disambiguation of LREs in a tweet into entities in a gazetteer in order to infer the context of the tweet (e.g., location where the tweet is posted, the identity of the author). As mentioned in the previous section, there are several studies for the task. The majority of these studies models the location inference as a multi-class classification problem on the grid over geo-spatial areas or cities on the gazetteer such as GeoNames and DBpedia [12], [16].

However, there exist many other useful linguistic expressions that can infer the current location of the author of a tweet. For example, a tweet describes the impressions of a location that the author visited in the past, or intends to visit in the near future.

Table 1 List of target noun on pilot categorization task.

Proper Nouns	秋葉原 (Akihabara) 仙台 (Sendai) 渋谷駅 (Shibuya-Station) 清水寺 (Kiyomizu-dera) スカイツリー (Skytree)
Common Nouns	病院 (hospital) 市役所 (city hall) 交差点 (crossing) 改札 (ticket gate) 動物園 (zoo)

Table 2 Result of pilot categorization task.

Total Tweets		1000
Label	`Present`	261 (26.1%)
	`Past`	179 (17.9%)
	`Future`	211 (21.1%)
	`Non-Temporal`	288 (28.8%)
	`Other`	61 (6.1%)

Likewise, there exist a large amount of tweets that contain LREs mentioning specific locations, without implying the user's current location, impression, or intention.

In order to verify the above intuition, we categorized 1,000 randomly sampled Japanese tweets posted during May–October 2015 that meet the following criteria.

- A tweet must that contain at least one LRE mention listed in **Table 1** in body text of the post.
- A tweet must be posted by one of the clients in the white list. This excludes automatically generated tweets by templates or bots (spam)[*2].
- A tweet must be a regular post, not as a retweet (RT) nor reply (@username).

We asked a Japanese independent linguistics analyst, who is not listed as an author, to categorize these tweets into the following four categories:

- `Present`: the author of the tweet is present in or close to the location expressed by LRE.
- `Past`: the author is not present in the location, but the tweet indicates that the author was present there in the past.
- `Future`: the author is not present in the location, but the tweet indicates that the author will visit there in future.
- `Non-Temporal`: the author only mentions the location referenced by the LRE, and was not nor will not be present at that location.

Table 2 shows the result of the manual categorization. We found that only 26.1% of the tweets indicate the presence of the user in the location referred by LREs (labeled as `Present`) and about 40% of the tweets were labeled as `Past` and `Feature`. In other words, even if a tweet contains a location name, it may not express the user's actual location.

This result has an important implication about applications. The previous work mostly identifies a location to a tweet but not to each LRE mention in the tweet. However, as shown in Table 2, 74% of tweets include an LRE that refers to the location where the author is not currently present. This means that it is impractical to estimate the user's location only by predicting a location for a tweet as a whole. Instead, we need to analyze textual clues around an LRE (e.g., "I love to visit Sendai.") to predict the cur-

[*2] Filtering script is available on https://bitbucket.org/conditional/tweet_utils/src/40610d5198874458446eac491b657238ab1b4d5d/filter_whitelist.py

rent location of an author. Thus, it is necessary to identify each LRE in a tweet so that we can predict the author's profile. In order to realize this, we decided to annotate each LRE with its corresponding entity. The subsequent section describes our annotation procedure and corpus created by this study.

4. Challenges in Annotating LREs

In this section, we describe the new research challenges associated with annotating geographical entities in text and our policies for addressing these issues.

4.1 Systematic Polysemy of LREs

One prominent issue in annotating facility entities is so-called *systematic polysemy* inherent in mentions referring to facilities (see, for example, Ref. [17]). For example, the mention "the Ministry of the Environment" in sentence (1) below refers to a specific location while the mention "the Ministry of the Environment" in (2) should be interpreted as an organization and does not refer to the location of the organization.

(1) 午後は 環境省 にいます / I'll be at the Ministry of the Environment this afternoon.

(2) これから 環境省 の職員に会ってきます / I will go to meet a staff member of the Ministry of the Environment.

This distinction can be crucial in potential applications of annotated geographical entities. In our annotation guidelines, ambiguities of this nature need to be resolved.

4.2 Analysis of Not Annotated Examples

Another issue in annotating facilities in SNS text is how to manage cases in which a mention refers to a certain (unique) facility entity, but the reader (annotator) cannot resolve it to any specific entry in the gazetteer by only using the information from the local context. For example, the mention "the park" refers to a certain unique location but the local context provides insufficient information for identifying it.

(3) 公園 でスケボーしてる人達眺めてる / I'm looking at the people skateboarding in the park.

According to our corpus study, roughly 50% of facility-referring mentions in our microblog text samples cannot be resolved to a specific entry in the gazetteer. One straightforward way to manage these type of mentions is to discard all common noun phrases from the targets of our annotation. However, since one can also quite often find common nouns that can be resolved to a specific gazetteer entry as in **Fig. 1**, it is intriguing to see the distribution of such cases through a large corpus study and consider the task of building a computational model for analyzing them. Motivated by this consideration, we incorporate the following two tags in our annotation guidelines:

Underspecified (UNSP) indicates that the tagged segment refers to a certain unique geographical entity but is not identifiable (i.e. cannot be resolved to any entry from the gazetteer).

Out of Gazetteer (OOG) indicates that the referent of the tagged segment is a geographical entity and can be identified, but is not included in the gazetteer.

Fig. 1 Overview of the corpus.

In contrast to the news domain, a tweet may include LREs that a third party find it difficult to identify their locations without knowing the author, especially when the tweet is dedicated to friends or circles close to the author. Therefore, there are many instances that are underspecified by the author. This is contrast to the news domain where the author provides sufficient information for every reader. For this reason, UNSP tag is essential for our study.

4.3 Building a Gazetteer of Facility Entities

Another problem we faced was how to build a gazetteer. For location entities (toponyms), it tends to be easier to find a comprehensive list from public databases such as GeoNames [4], [9]. For facilities, on the other hand, since the referents of LREs in microblogs include a broad variety of facilities, including stations, restaurants, shopping stores, hospitals, and schools, it is not a trivial job to build a comprehensive list of those facilities with a sufficient coverage even if the targets are limited to a single country.

For our corpus study, we were fortunate to be able to use the data collection from the Location Based Social Networking Service (LBSNS) as reported in Section 5.2. However, our corpus study suggests that our gazetteer still needs to be extended to ensure improved coverage. In addition, we also had to determine ways in which to share the database with other research sites.

5. Annotation Specifications

In this section, we provide an overview of the specifications of our annotation schema based on the issues discussed in Section 4.

5.1 Annotation Guidelines

In the existing named entity tagged corpora in Japanese, expressions are annotated with a named entity class and its boundaries. However, the corpora does not contain annotations as to whether each of the expressions actually relates to an entity. Partly following the annotation guidelines in TAC KBP [18][*3],

[*3] http://www.nist.gov/tac/2014/KBP/ColdStart/guidelines/TAC_KBP_2014_EDL_Query_Development_Guidelines_V1.5.pdf

Table 3 Definition of the tags used in our annotation.

Tag	Example	Description
LOC (Location)	埼玉県 / Saitama-prefecture, 仙台市 / Sendai-city	Specific geographical area
FAC (Facility)	仙台駅 / Sendai-station, 九州大学 / Kyusyuu University, 南 武 線 / Nanbu-line, 東 北 道 / Tohoku-expressway	Facility/Road/Railroad entity that has a specific location

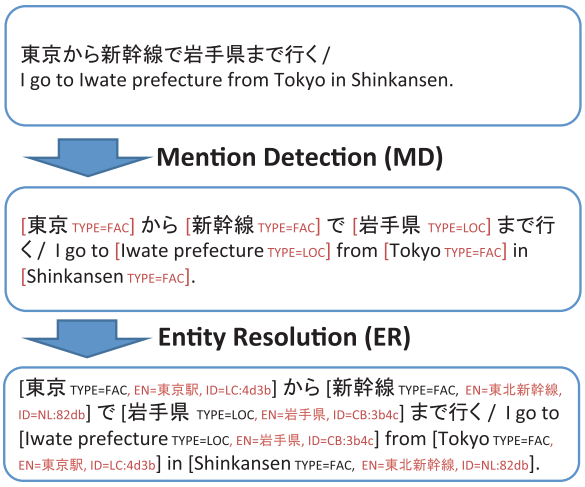


Fig. 2 Flow of our annotation scheme.

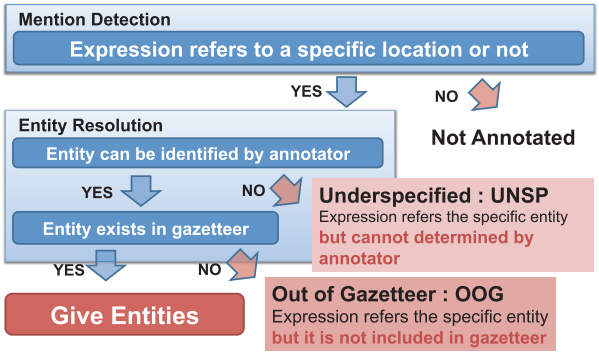


Fig. 3 Description of OOG and UNSP tag.

the extended named entity tag set [19] and the Japanese extended Named Entity-tagged corpus, we adopted the approach illustrated in **Fig. 2** to annotate microblog texts. Note that most existing entity linking corpora such as TAC KBP adopt either the Wikipedia or the Freebase[*4] as the gazetteer of entities. However, these databases do not cover local LREs and cannot be used for linking supermarket branches, local clinics, etc.

The annotation task consists of the following two subtasks:

Mention Detection (MD) Given a microblog text (i.e., a tweet), an annotator annotates all the mentions which refer to specific geographic entities with a predefined set of tags given in **Table 3**.

Entity Resolution (ER) For each detected mention, an annotator searches the gazetteer for its referred entity and annotates the linking. We allow a mention to be linked to multiple gazetteer entries. If the referent cannot be found in the gazetteer, annotate the mention as **OOG**, and if the referent is not identifiable, annotate the mention as **UNSP**, as depicted in **Fig. 3**.

In our annotation, all potential LREs in the text are annotated. Following Ref. [13], non-referring expressions,such as "town" and "city" in "It is better to live in a small town than in a big city", are not annotated. Deictic references such as "there" and pronouns are not annotated. The annotators are allowed to use the information from the writer's profile for reference purposes.

In the ER step, a mention may have a number of possible interpretations (entity assignments) in general. In our annotation, we did not enumerate these entities exhaustively; however, we took a balanced approach between the annotation workload and the speed. When an LRE can be mapped to multiple entities, our guidelines asked annotators to assign as many entities as possible. However, when the number of candidate entities are too large (roughly ten or more), an annotator can assign a UNSP tag to the mention. For example, the two mentions in the example below have multiple possible entities.

(4) [東京 TYPE:FAC, EN={ 成田空港, 羽田空港 }] から [大阪 TYPE:FAC, EN={ 伊丹空港, 関西国際空港 }] まで飛行機で向かいます。/ I'll go to [Tokyo TYPE:FAC, EN={Narita International Airport, Haneda International Airport}] from [Osaka TYPE:FAC, EN={Itami Airport, Kansai International Airport}] via airplane.

In addition, our annotation guidelines asked annotators to choose the best specific entity as possible. For example, we prefer facility entity "Shibuya station" to location entity "Shibuya ku" because the former is more specific.

5.2 Gazetteers

In Japan, under open data initiatives, government agencies have released data with the specific latitude and longitude for the name to be used as a postal address, such as the prefecture and city (City-block level location reference information[*5]). Therefore, this can be used as the location name gazetteer. However, for facility entities, no existing comprehensive database is available. We used data crawled from Yahoo! Loco[*6], which is one of the Location Based Social Networking Services (LBSNSs). This is a large, but noisy, amount of data, which contains many duplicate records of the entity and surface variations. Therefore, we cleaned up entries that were ambiguous or those whose name was either too short or too long by using several handwritten rules. In addition, we used entities downloaded from "National Land Numerical Information"[*7] for railroad data. **Table 4** presents an overview of the resulting entity gazetteer. The Location entity gazetteer includes prefectures, cities, and other administrative areas such as "大字 (oaza)" (sections) and villages. The Facility entity gazetteer includes a broad variety of facilities including stations, restaurants, shopping stores, hospitals, and schools. As a result, we compiled a large (more than 5 million entries) gazetteer

[*4] In 2016, Google is shutting down Freebase Search API. Data dumps provided in http://freebase.com/.

[*5] http://nlftp.mlit.go.jp/isj/

[*6] http://loco.yahoo.co.jp/

[*7] http://nlftp.mlit.go.jp/ksj-e/

Table 4 Overview of entity gazetteer used in our annotation.

Gazetteer Type	Source	Number of Entries
Locations	City-block level location reference information	147774
Facilities	Yahoo! Loco, National Land Numerical Information	4990239

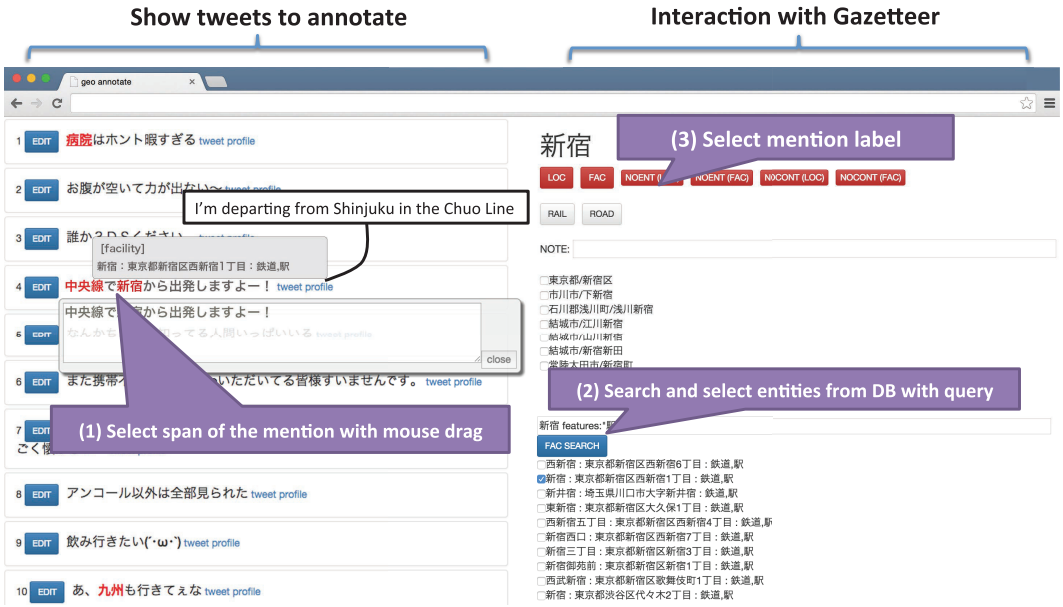


Fig. 4 Screenshot of annotation tool.

of location and facility entities in Japan.

Each entity is formatted as GeoJSON Feature object[*8], as this format is easy to use with other GIS applications.

5.2.1 Selection of Gazetteer

In this study, we used Yahoo! Loco data as the gazetteer of facility entities; at that time, we also considered other resources, such as Foursquare, as a candidate of a gazetteer. The Foursquare database is constructed with social effort, based on an individual user's contribution (also called as "check-in"). In contrast, the Yahoo! Loco database is constructed from several domain specific databases including the telephone directory, which is regarded as a highly trusted list in Japan[*9].

In literature of LBSNS, it is reported that location sharing (check-in) actions are biased by the purposes of visits to location [20]. The authors pointed out that the primary motive for sharing location is to interact with the social circle by sharing a 'positive' experience. Therefore, gazetteers built from a location sharing service may be biased to locations where people receive positive experiences. For this reason, we annotated tweets using Yahoo! Loco data, which is expected to include unbiased locations of high coverage in a variety of domains.

5.3 Two Sub-corpora for Annotation

We performed annotations for 10,000 randomly sampled tweets that were tweeted during a specific time period (**RANDOM**), However, this trial revealed that random sampling may not be a suitable method for collecting tweets for development of annotation scheme because randomly sampled tweets rarely contain an LRE (only around 10% of tweets in RANDOM sub-corpus contain an LRE), the yield ratio of entities is low and inefficient. Therefore, we performed annotations for another set of 1,000 tweets (**FIL**) for development and evaluate annotation scheme, which were sampled by the following rules: (1) Tweets must include two or more potential location names that can be verified by simple string matching with the location entity gazetteer. (2) One of the location names of rule (1) must be the location name of a prefecture in which the annotator resides. These rules increased the LRE density, and enabled us to efficiently collect instances of geographical entity annotation, which were then used as a basis for refining the annotation guidelines. In a later section, we discuss the inter-annotator agreement in the FIL sub-corpus.

5.4 Tool for Corpus Annotation

Compared with mention detection, entity resolution tends to be considerably more expensive particularly when the gazetteer at hand has a large coverage. For a given geographical mention, the gazetteer may have dozens of candidate entries, from which the annotator would have to select the correct one. The tasks of searching for the candidate entries and choosing the most appropriate one from among them can be substantially supported with an adequate computational environment. For this purpose, we created an annotation support tool especially designed for our annotation schema. Unlike tools devised in prior work [9], our tool stores the entire data of our gazetteer (including, for example, the postal address, ontological category, etc., for each facility entity) on a standard full-text search engine and allows the use to search for candidate entries with an arbitrary query string, as illustrated in **Fig. 4**.

Our annotation tool has the following UI requirements:

- The user can annotate the offset and length of a mention span via a drag-and-drop action on the text (Fig. 4 (1)).
- For each selected text span, the annotator can associate a set

[*8] http://geojson.org/
[*9] http://pr.yahoo.co.jp/release/2011/0601b.html

Table 5 Number of tagged expressions in annotated corpus.

Tag	#tagged expression	#tagged with entity	OOG	UNSP
LOC	406	298 (73.4%)	14 (3.4%)	94 (23.2%)
FAC	545	221 (40.6%)	43 (7.9%)	281 (51.6%)
TOTAL	951	519 (54.6%)	57 (6.0%)	375 (39.4%)
#Tweet		10000		
#Character		332739		

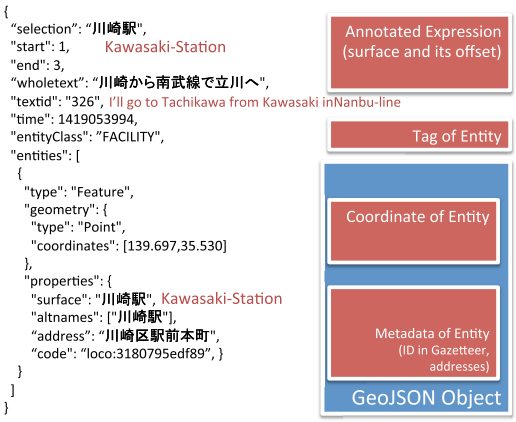


Fig. 5 Example of annotated data.

of any entities via a list of candidates (Fig. 4 (2)).

- The tool can retrieve a list of candidates from a backend gazetteer database, based on the selected mention of string.
- If no appropriate entity exists in the candidate list, the user can either give another query to the database to find an appropriate entity or assign a UNSP or OOG tag to a mention (Fig. 4 (3)).

In Fig. 4, the annotator assigns entities to the mention "新宿 (Shinjuku)" in the sentence 4.

The difference between a general purpose annotation system such as brat[*10] and ours is that the annotator can lookup entity candidates using any structured query which is supported by a full-text search engine[*11]. Thus, the annotator can filter based on specific ontological category information. For example, the entity has an ontological category "駅 (Station)" which is stored together with the entity name and the coordinates of the entity, as shown in Fig. 4 (2).

This tool works as a Web application, and is capable of working with more than one person at the same time. **Figure 5** shows an example of the annotated data, in which the annotated entities are represented by the list of GeoJSON objects, and each object has an ID that uniquely corresponds to an entity in the gazetteer.

6. Corpus Annotation and Evaluation

Using the annotation tools mentioned in the Section 5.4, we annotated 10,000 tweets randomly selected from tweets sent during 2014.

In actual annotation, two Japanese (in the Miyagi prefecture, Japan) graduate students annotated 5,000 tweets each. As mentioned in Section 6.2, although the annotation bias caused by background knowledge of the annotator was expected, we did not question at this time. In addition, the sample tweets were filtered using the client list described in Section 3. The client list does not have any location related SNS client such as Foursquare. Thus, our corpus does not contain trivial tweets such as a check-in tweet generated by these services.

Table 5 shows the number of tagged expressions in the **RANDOM** sub-corpus.

This result shows that our approach annotating not only geographical location entities but also facility entities could extract a number of LREs from Twitter corpora. We think annotating facility entities are useful in that they are more specific and informative than location entities in general. Location entities are sometimes ambiguous because the major location entities are prefectures and cities (e.g., "Miyagi" and "Sendai"), and may refer to large areas. In contrast, a facility entity (e.g., "Shibuya Station") usually refers to a specific point with a relatively small area. Thus, our annotation scheme can find location information more effectively, and capture an important factor that were overlooked in the previous corpora.

As an evaluation of the coverage of the gazetteer, we calculated those location and facility names which are annotated with entities in the gazetteer. This result shows that 519 out of 951 (54.6%) LREs were annotated with entities. We manually analyzed the instances that were not linked to entities and obtained the following findings:

Location These instances mainly suffer from an absence of foreign location names such as "Rome", "New York", consisting of surrounding areas such as "Higashi Mikawa", and tourist resorts such as "Mount Zao".

Facility In most cases, highly ambiguous instances, such as "house", "McDonald's", and "workplace", were difficult to annotate with an entity. As these instances are dependent on the context of the writer, a third person would be unable to guess the specific entity despite considering the whole text.

6.1 Quality of Annotation: Mention Detection

To examine the problems underlying the annotation guidelines, we additionally asked two annotators to annotate 200 tweets in **FIL** sub-corpus independently to see how they agree with each other.

First, the two set of annotations were converted into IOB2 codings at the character level, and assuming that the annotation on one side is correct, we then calculated the precision, recall, and the F1-Score of the annotation on the other side. For reference, comparing two annotations at the character level, Cohen's Kappa was 0.892. **Table 6** shows the evaluation results of the inter-annotator agreement. This indicated that the annotation is generally successful, but the annotation quality of the FAC tag is slightly lower. As mentioned above, in this annotation, annotators need to interpret the intent of the writer of a text (irrespective

[*10] http://brat.nlplab.org
[*11] We used Elasticsearch (https://www.elastic.co/products/elasticsearch).

Table 6 Evaluation results of inter-annotator agreement (assuming the annotation on one side is correct) measured on **FIL** sub-corpus.

Tag	Precision	Recall	$F_{\beta=1}$
LOC	87.68% (178/203)	97.27% (178/183)	92.23
FAC	89.25% (83/ 93)	72.81% (83/114)	80.19
Overall	88.18% (261/296)	87.88% (261/297)	88.03

of whether a specific location is assumed).

(5) これでもう 大学図書館 から取り寄せてもらわなくていいのね … / I don't need to order from university library anymore.

In this example, one annotator judged "university library" as a facility name, while the other judged it as an organization and did not annotate it as an LRE. This arrangement probably makes annotation harder; hence, we would have to re-examine this guideline for future work.

6.2 Quality of Annotation: Entity Resolution

To evaluate our entity resolution annotation scheme quantitatively, we compare entities annotated by two annotators for an LRE mention. As metrics for measuring inter-annotator disagreements, we use the Average Error Distance (AED) and Median Error Distance (MED) , following the with related work.

Each entry of the gazetteer has a single coordinate (latitude, longitude) pair based on the WGS 1884 coordinate system. For a facility entity, the coordinates are obtained from the geocoded address information of the Yahoo! Loco database. For a location entity, the coordinates are obtained from "National Land Numerical Information", which is roughly consistent with the centroid of the city block. In exceptional cases, for administrative divisions such as a prefecture or a city, the gazetteer has the location of the administrative office such as a city hall or prefectural government as the representative coordinate of the location entity. The disagreement of an LRE disambiguation between the two annotators is measured by the absolute distance of the coordinates to which the two annotators associated for the LRE by choosing entities in the gazetteer. AED and MED are defined as the average and median, respectively, of the distance values for all LREs. These measures are widely used in this research for the automatic annotation of geographical information to unstructured data including text [9], [10], [11] and images [21]. The distances are calculated using the Haversine formula[*12].

Each of the two annotators annotated 243 expressions, and the AED was determined as 1648 meters, whereas the MED was found to be 0 meters. **Figure 6** shows the distribution of Error Distance as histogram. This figure indicates that the majority of errors is very close.

Of these 243 instances, 199 (81.9%) show an error distance of 0 meters. In other words, two annotators annotated exactly the same entity for these instances. The following example shows instances with large errors in the distance. This instance indicates that the two annotators made different interpretations, and thus the annotations differed. We denote the annotators as A and B.

(6) (Error Distance: 70.8 km) 江坂周辺、[淡路 A:LOC/兵庫県淡路市 B:FAC/淡路駅 (大阪市東淀川区)] 周辺、西中島南方周辺、新大阪周辺でバイト見つけたい / I want to work in a part-time job near Esaka, [Awaji A:LOC/Awaji-shi, Hyogo B:FAC/Awaji Station (Yodogawa-ku, Osaka-shi)], Nishi-Nakajima, or Shin-Osaka.

[*12] https://en.wikipedia.org/wiki/Haversine_formula

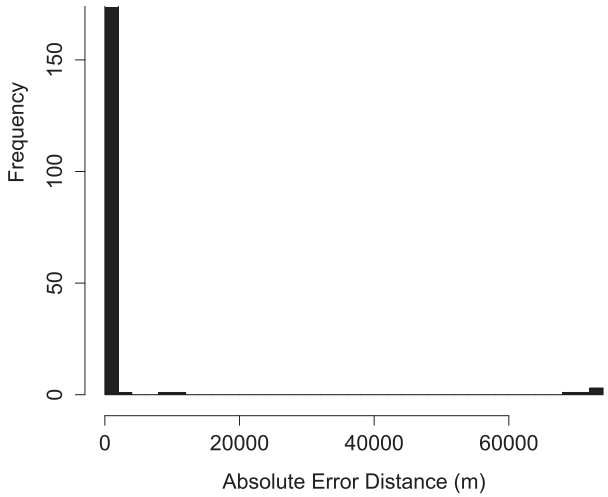


Fig. 6 Histogram of Absolute Error Distance in 243 entity pair annotated by two annotatior in **FIL** subcorpus.

According to the two annotators, one annotator interpreted each location name in this example literally and confirmed that these location names belong to "Kansai region", then annotated "Awaji-shi", which has the largest population. The other annotator perceived that these location names are station names in a specific region, then interpreted "Awaji" as a station name in "Osaka-shi".

(7) (Error Distance: 68.9 km) [福島 A:FAC/福島第一原子力発電所 B:LOC/福島県福島市] の事故で風評被害じゃないんだよ。/ It is not a harmful rumor, but [Fukushima A:FAC/Fukushima I Nuclear Power Plant B:LOC/Fukushima-shi, Fukushima]'s accident.

In example (7), to reason that "福島/Fukushima" means "福島第一原子力発電所/Fukushima I Nuclear Power Plant", the annotator needs the background knowledge that there are harmful rumors caused by the accident of Fukushima.

We plan to discuss how much reasoning or background knowledge should be used for annotation.

6.3 Required Clues for Entity Resolution

As we show below, although some LREs need complex reasoning and annotations for them disagree, we also find there are also LREs which are considered to be easily identified by a simple clue. We investigated the annotated entities in the 10,000 tweets in **RANDOM** sub-corpus, judged what types of clues are required for manual entity resolution, and examined the distribution. When we performed manual judgement, we assumed that the LRE tag (location or facility name) and the boundary is given, and then we focused on the types of clues required for entity resolution, which can require multiple clues. In addition, LREs annotated with a single entity are subject to investigation. Therefore, 267 location names and 169 facility names were investigated. **Table 7** shows the result. This table enables us to make the following observations.

Table 7 Required Clues for entity resolution.

Clue	LOC	FAC	TOTAL
(1) No ambiguity (There was only one candidate entity in the gazetteer, and it was the correct entity)	85 (31.8%)	48 (28.4%)	133 (30.5%)
(2) Candidate entity which has the largest population is the correct entity	151 (56.6%)	0 (0.0%)	151 (34.6%)
(3) Need to deal with abbreviations or variations of surface form	5 (1.9%)	74 (43.8%)	79 (18.1%)
(4) Resolved by considering other LREs in the text	25 (9.4%)	17 (10.1%)	42 (9.6%)
(5) Resolved by considering contextual information in the text	0 (0.0%)	34 (20.1%)	34 (7.8%)
(6) Resolved by considering global context (profile data, URL, photo, and so on)	1 (0.4%)	11 (6.5%)	12 (2.8%)

Nearly 30% of location names presented no ambiguity, and more than half of these were annotated with the candidate entity with the largest population. Therefore, as for location names, population seems to be a good baseline for entity resolution. This result is consistent with those of Ref. [9], which targeted the newspaper domain. However, in the case of facility names, entity resolution was more complicated. Although the proportion considered to be unambiguous is virtually the same as that of the location names, there are no existing metrics, such as population, for facility entities. Therefore, defining metrics, such as population, is desirable. For that purpose, we would prefer to consider a term such as "popularity". To calculate these metrics, the check-in counts of a Location Based Social Network Service (LBSNSs), such as Foursquare[*13], appear to be useful.

In addition, 40% of facility names require the ability to process abbreviations and variations of surface forms. For example, "Hama-sta" in the following text seems to refer to "Yokohama Stadium"; however, it is not possible to look this up directly in the facility entity gazetteer.

(8) ハマスタ で試合観戦なう / I'm watching a game at Hama-sta.

To address this, we would have to consult the gazetteer flexibly, by using methods such as approximate string matching [22]. As this is a widespread problem with facility names, it would have to be addressed to enable grounding to be performed.

Moreover, 20% of facility names required local context in the text (other than LRE). The following is an example.

(9) 山手線で 東京 から品川に向かっています / I'm going toward Shinagawa From Tokyo.

In this example, "Tokyo" seems to refer to "Tokyo Station", considering the local context in the text. As far as we searched, most of the entities requiring local context were station names such as "Tokyo Station".

7. Limitations and Future Work

In this study, we independently annotated individual tweets, without considering the context of the author of a tweet. Existing studies, which only annotated locations mentioned in a news corpus, reported that geographical entities have consistency or minimality (e.g., location mentions of geographical entities co-occurring in same document are located in a similar geographical region). However, these consistencies do not exist in our corpus because the unit of a document in our corpus is an individual tweet. In addition, we used the proprietary gazetteer as the facility entity gazetteer owing to its high coverage; therefore, it is difficult to re-distribute the gazetteer. In future, we will attempt to annotate with more open resources such as GeoNames, and compare the coverage of these gazetteers.

From the NLP perspective, this corpus can be used as training/test data for information extractor where LREs are automatically identified from raw tweets. Similarly, this data can be used as training/evaluation data for the automatic entity resolution (in similar task called as Entity Linking or Named Entity Disambiguation). Some might think the corpus created by this study is small. However, we argue that the previous studies have never been specialized for geo-location entities, building corpora for generic entities (e.g., PERSON, ORGANIZATION) with a few thousands of mentions. Thus, our corpus is the largest resource for bridging geo-locational mentions to actual entities. We expect that the approaches for Entity Linking, for example, unsupervised graph-based method [23] and coherence based collective resolution [24] will also work for resolving geographical entities.

8. Conclusion

This paper discusses the problems associated with the task of annotating geographical entities on Japanese microblog texts and reports the preliminary results of the actual annotation. All the annotation data and the annotation guidelines are publicly available for research purposes from our web site.

The annotation task consisted of two subtasks: mention detection and entity resolution. Our corpus study showed that our annotation scheme could achieve a reasonably high inter-annotator agreement.

The scope of the annotation was extended to facility entities by introducing the **OOG** and **UNSP** tags. The distributions of these tags obtained through our corpus study will provide useful implications for our future work for an improved annotation setting.

We also investigated the types of clues that are considered useful for entity resolution and found that the task of identifying facility entities poses interesting research issues including abbreviations, variations of surface forms, and the popularity of each facility. In particular, the popularity appears to be important in resolving facility entities.

Acknowledgments This research was supported by the program *Research and Development on Real World Big Data Integration and Analysis* of the Ministry of Education, Culture, Sports, Science and Technology, Japan and by the Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Agency (JST).

[*13] https://foursquare.com/

References

[1] Li, J., Ritter, A. and Hovy, E.: Weakly Supervised User Profile Extraction from Twitter, *Proc. 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), Association for Computational Linguistics, pp.165–174 (2014).

[2] Signorini, A., Segre, A.M. and Polgreen, P.M.: The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic, *PLoS ONE*, Vol.6, No.5, p.e19467 (online), DOI: 10.1371/journal.pone.0019467 (2011).

[3] Collier, N.: Uncovering text mining: A survey of current work on web-based epidemic intelligence, *Global Public Health*, Vol.7, No.7, pp.731–749 (online), DOI: 10.1080/17441692.2012.699975 (2012). PMID: 22783909.

[4] Middleton, S., Middleton, L. and Modafferi, S.: Real-Time Crisis Mapping of Natural Disasters Using Social Media, *Intelligent Systems, IEEE*, Vol.29, No.2, pp.9–17 (online), DOI: 10.1109/MIS.2013.126 (2014).

[5] Ohtake, K., Goto, J., De Saeger, S., Torisawa, K., Mizuno, J. and Inui, K.: NICT Disaster Information Analysis System, *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, Asian Federation of Natural Language Processing, pp.29–32 (2013).

[6] Varga, I., Sano, M., Torisawa, K., Hashimoto, C., Ohtake, K., Kawai, T., Oh, J.-H. and De Saeger, S.: Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster, *Proc. 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), Association for Computational Linguistics, pp.1619–1629 (2013).

[7] Matsuda, K., Sasaki, A., Okazaki, N. and Inui, K.: Annotating Geographical Entities on Microblog Text, *The 9th Linguistic Annotation Workshop* (*LAW IX*), pp.85–94 (2015).

[8] Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J. and Muhlhauser, M.: A Multi-Indicator Approach for Geolocalization of Tweets., *ICWSM*, The AAAI Press (2013).

[9] Leidner, J.L.: Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding, *SIGIR Forum*, Vol.41, No.2, pp.124–126 (online), DOI: 10.1145/1328964.1328989 (2007).

[10] Speriosu, M. and Baldridge, J.: Text-Driven Toponym Resolution using Indirect Supervision, *Proc. 51st Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers, Association for Computational Linguistics, pp.1466–1476 (2013).

[11] DeLozier, G., Baldridge, J. and London, L.: Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles, *Proc. of AAAI 2015*, The AAAI Press (2015).

[12] Lieberman, M.D., Samet, H. and Sankaranarayanan, J.: Geotagging with local lexicons to build indexes for textually-specified spatial data., *ICDE*, pp.201–212, IEEE (2010).

[13] Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S. and Clancy, S.: SpatialML: annotation scheme, resources, and evaluation, *Language Resources and Evaluation*, Vol.44, No.3, pp.263–280 (online), DOI: 10.1007/s10579-010-9121-0 (2010).

[14] Zhang, W. and Gelernter, J.: Geocoding location expressions in Twitter messages: A preference learning method, *J. Spatial Information Science*, Vol.9, No.1, pp.37–70 (online), DOI: 10.5311/JOSIS.2014.9.170 (2014).

[15] Kitamoto, A. and Sagara, T.: Toponym-based Geotagging for Observing Precipitation from Social and Scientific Data Streams, *Proc. 2012 ACM Workshop on Geotagging and Its Applications in Multimedia, GeoMM'12 (co-located with ACM Multimedia 2012)*, pp.23–26, ACM (2012).

[16] Wing, B. and Baldridge, J.: Hierarchical Discriminative Classification for Text-Based Geolocation, *Proc. 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Association for Computational Linguistics, pp.336–348 (2014).

[17] Peters, W. and Peters, I.: Lexicalised Systematic Polysemy in WordNet, *Proc. Second International Conference on Language Resources and Evaluation* (*LREC'00*), European Language Resources Association (ELRA) (2000).

[18] Ji, H., Dang, H., Nothman, J. and Hachey, B.: Overview of TAC-KBP2014 Entity Discovery and Linking Tasks, *Proc. Text Analysis Conference* (*TAC2014*) (2014).

[19] Sekine, S., Sudo, K. and Nobata, C.: Extended Named Entity Hierarchy., *Proc. Third International Conference on Language Resources and Evaluation* (*LREC-2002*), Las Palmas, Canary Islands — Spain, European Language Resources Association (ELRA) (2002).

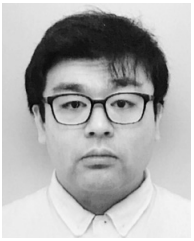
[20] Patil, S., Norcie, G., Kapadia, A. and Lee, A.: "Check out Where I Am!": Location-sharing Motivations, Preferences, and Practices, *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, New York, NY, USA, ACM, pp.1997–2002 (online), DOI: 10.1145/2212776.2223742 (2012).

[21] Hays, J. and Efros, A.A.: IM2GPS: estimating geographic information from a single image, *IEEE Conference on Computer Vision and Pattern Recognition*, *2008*, *CVPR 2008*, pp.1–8, IEEE (2008).

[22] Okazaki, N. and Tsujii, J.: Simple and Efficient Algorithm for Approximate Dictionary Matching, *Proc. 23rd International Conference on Computational Linguistics*, COLING '10, Stroudsburg, PA, USA, Association for Computational Linguistics, pp.851–859 (2010).

[23] Pan, X., Cassidy, T., Hermjakob, U., Ji, H. and Knight, K.: Unsupervised entity linking with abstract meaning representation, *Proc. 2015 Conference of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies* (2015).

[24] Ling, X., Singh, S. and Weld, D.: Design Challenges for Entity Linking, *TACL*, Vol.3, pp.315–328 (2015).



Koji Matsuda received his bachelor's degree in engineering from Toyohashi Institute of Technology in 2006, and his M.E. degree in engineering from Tokyo Institute of technology in 2012. He has been a researcher of Graduate School of Information Sciences, Tohoku University since 2014. His current research interest is natural language processing, especially entity linking and automatic knowledge acquisition.

Akira Sasaki received his bachelor's degree in engineering from Tohoku University in 2013, and his M.S. degree in information science from Tohoku University in 2015. He has been a Ph.D. student of Graduate School of Information Sciences, Tohoku University since 2015. His current research interest is natural language processing, especially stance detection and sentiment analysis.

Naoaki Okazaki is an associate professor at Graduate School of Information Sciences, Tohoku University. Prior to his faculty position, he worked as a research fellow in National Centre for Text Mining (NaCTeM) (in 2005) and as a postdoctoral researcher in University of Tokyo (in 2007–2011). He obtained his PhD from Graduate School of Information Science and Technology, University of Tokyo in 2007. He has served as a technical consultant in SmartNews Inc. since 2013. He is also a visiting research scholar of the Artificial Intelligence Research Center (AIRC), AIST. His research interests include natural language processing, text mining, and machine learning.

Kentaro Inui received his doctorate degree of engineering from Tokyo Institute of Technology in 1995. He has experience as an assistant professor at Tokyo Institute of Technology and an associate professor at Kyushu Institute of Technology and Nara Institute of Science and Technology, he has been a professor of Graduate School of Information Sciences at Tohoku University since 2010. His research interests include natural language understanding and knowledge processing. He currently serves as the IPSJ director and ANLP director.