Annotating Geographical Entities on Microblog Text

Koji Matsuda[1], **Akira Sasaki**[1], **Naoaki Okazaki**[1,2], and **Kentaro Inui**[1]

[1]Graduate School of Information Sciences, Tohoku University
6-6-05 Aramakiaza-Aoba, Aoba-ku, Sendai, 980-8579, Japan
[2]Japan Science and Technology Agency (JST)
[1]{matsuda, aki-s, okazaki, inui}@ecei.tohoku.ac.jp

Abstract

This paper presents a discussion of the problems surrounding 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 not only annotate geographical location entities but also facility entities, such as stations, restaurants, shopping stores, hospitals and schools. We discuss ways in which to build a gazetteer, the types of ambiguities that need to be considered, reasons why the annotator tends to disagree, and the problems that need to be solved 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.

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 (Li et al., 2014), disease monitoring (Signorini et al., 2011; Collier, 2012), and disaster response (Middleton et al., 2014; Ohtake et al., 2013; Varga et al., 2013).

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 (Middleton et al., 2014) 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. As explained above, 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 de-

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 85–94,
Denver, Colorado, June 5, 2015.

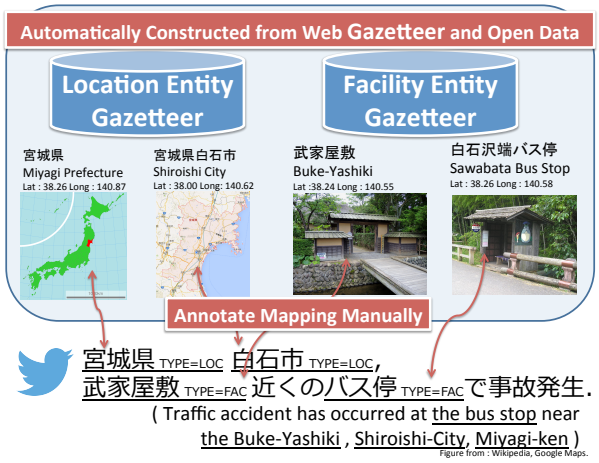

Figure 1: Overview of the corpus

sign 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 eightfold 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].

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 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 (Schulz et al., 2013).

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**) (Leidner, 2007). Speriosu and Baldridge (2013) proposed a supervised learning method by using an indirect supervision technique. DeLozier et al. (2015) proposed a gazetteer independent method by using density estimation techniques.

These studies were evaluated by using a reference corpus such as the TR-CoNLL (Leidner, 2007) or LGL(Local-Global Lexicon) (Lieberman et al., 2010) 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.

Mani et al. (2010) 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, Zhang and Gelernter (2014) annotated Twitter messages, but their annotation focus is limited to toponyms, and facility names are not annotated. Examples of geoparsing for Japanese text,

[1] http://www.cl.ecei.tohoku.ac.jp/~matsuda/LRE_corpus/

GeoNLP (Kitamoto and Sagara, 2012) exist, but there are no reports of quantitative evaluations of the performance, because there is no corpus for evaluation.

3 Challenges in Annotating LREs on Microblog Text

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

3.1 Systematic Polysemy of LREs

One prominent issue in annotating facility entities is the so-called *systematic polysemy* inherent in mentions referring to facilities (see, for example, Peters and Peters (2000)). For example, the mention "the Ministry of the Environment" in the 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.

3.2 Analysis of not annotated examples

Another issue in annotating facilities in microblogs 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 Figure 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 specifications:

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.

3.3 Building a Gazetteer of Facility Entities

Another problem we faced was to decide 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 (Leidner, 2007; Middleton et al., 2014). 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 4.2. 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.

4 Annotation Specifications

In this section, we provide an overview of the specifications of our annotation schema based on the is-

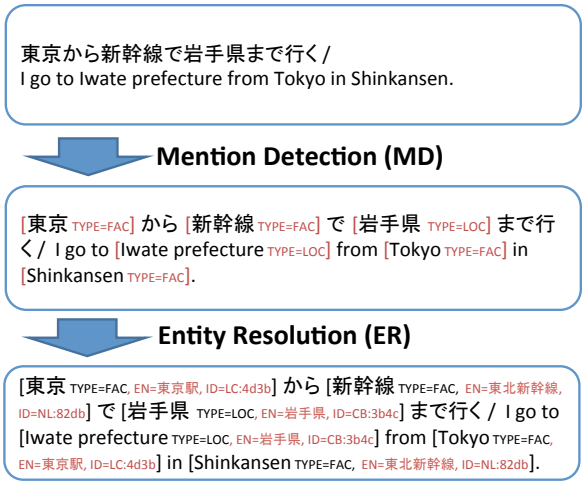

Figure 2: Flow of our annotation scheme

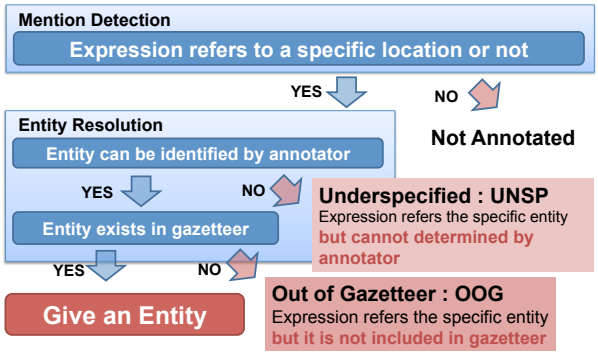

Figure 3: Description of OOG and UNSP tag

sues discussed in Section 3.

4.1 Annotation

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 (Ji et al., 2014)[2], the extended named entity tag set (Sekine et al., 2002) and the Japanese extended Named Entity-tagged corpus, we followed the approach illustrated in Figure 2 to annotate microblog texts. 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 1.

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**.

In our annotation, all potential LREs in the text are annotated. Following (Mani et al., 2010), 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.

4.2 Gazetteer

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[3]). Therefore, this can be used as the location name gazetteer. However, for facility entities, there is no existing comprehensive database. We used data crawled from Yahoo! Loco[4], 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 of which the name was either too short or too long by using several handwritten rules. In addition, we used entities downloaded from "National Land Numerical Information" for railroad data. Table 2 presents an overview of the resulting entity gazetteer. The Location entity gazetteer includes prefectures, cities, and other administrative areas such as "oaza" (sections) and villages. The Facility entity gazetteer includes a

[2] http://nlp.cs.rpi.edu/kbp/2014/elquery.pdf

[3] http://nlftp.mlit.go.jp/isj/

[4] http://loco.yahoo.co.jp/

Table 1: 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

Table 2: 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

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 of location and facility entities in Japan.

Each entity is formatted as GeoJSON Feature object [5], as this format is easy to use with other GIS applications.

4.3 Two Sub-corpora for Annotation

We performed annotations for 10,000 randomly sampled tweets that were tweeted during a specific time period (**RANDOM**), but this proved problematic for refining the annotation scheme rapidly. Because randomly sampled tweets very rarely contain an LRE, the yield ratio of entities is low and inefficient. Therefore, we performed annotations for another 1,000 tweets (**FIL**), which were filtered according to the following rules: (1) Tweets must include two or more potential location names that can be verified by peforming a simple string matching to 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 filters increase the LRE density, and enable us to rapidly advance the discussion to the annotation guideline. In a later section, we discuss the inter-annotator agreement in the FIL sub-corpus.

4.4 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 (Leidner, 2007), 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 Figure 5.

This tool works as a Web application, and is capable of working with more than one person at the same time. Figure 4 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.

5 Corpus Annotation and Evaluation

Using the annotation tools mentioned in the section 4, we annotated 10,000 tweets randomly selected from tweets sent during 2014. Table 3 shows the number of tagged expressions in the annotated corpus.

In addition, as an evaluation of the coverage extent 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

[5] http://geojson.org/

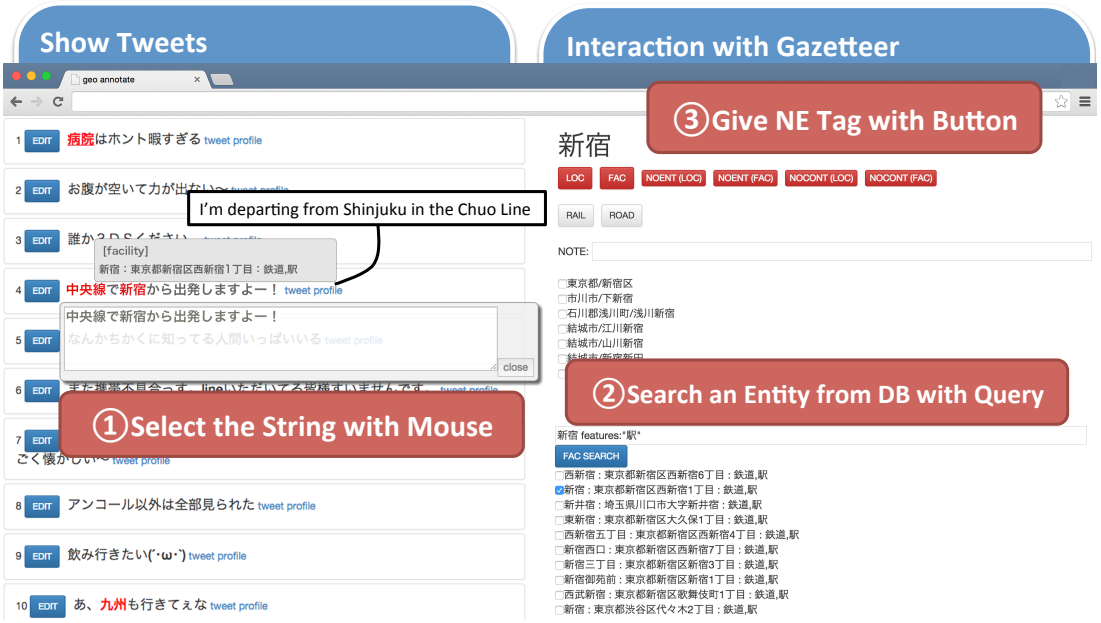

Figure 5: Screenshot of annotation tool

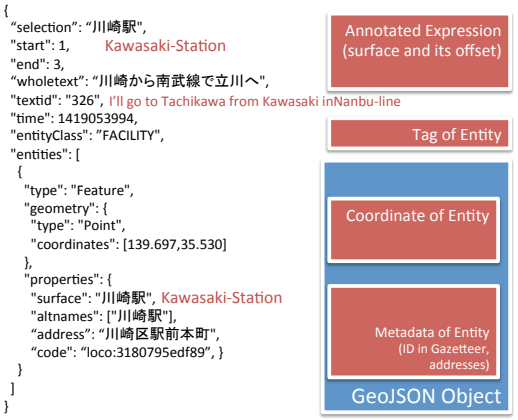

Figure 4: Example of annotated data

951 (54.6%) LREs were annotated with entities. As we analyzed instances without entities, we made the following observations.

Location These instances mainly suffer from an absence of foreign location names, 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.

5.1 Quality of Annotation: Mention Detection

To discuss the annotation specification, two annotators independently annotated 200 tweets.

First, two 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 4 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 of whether a specific location is assumed).

(4) 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.

Table 3: 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		

Table 4: Evaluation results of inter-annotator agreement (assuming the annotation on one side is correct)

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

In this example, one annotator judged "university library" as a facility name, on the other hand, 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.

5.2 Quality of Annotation: Entity Resolution

To evaluate our entity resolution annotation scheme quantitatively, we compare the coordinate pair of the entity that was annotated by two annotators, as described in the section 5.1. As error metrics, we use the Average Error Distance (AED) and Median Error Distance (MED) to ensure comparability with related work. 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. 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.

(5) (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.

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

5.3 Required Clues for Entity Resolution

As we show below, although some LREs need complex reasoning and annotations for them disagree, on the other hand, there are also LREs which are easily annotated by a simple clue. We investigated the annotated entities of 10,000 tweets in **RANDOM**, 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 5 shows the result. This table enables us to make the following observations.

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

Table 5: 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%)

result is consistent with those of (Leidner, 2007), 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[6] or Loctouch[7], 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.

(6) 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 (Okazaki and Tsujii, 2010). As this is a widespread problem with facility names, it would have to be addressed to enable grounding to be performed.

[6] https://foursquare.com/

[7] http://tou.ch/

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

(7) 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".

6 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. The automatic estimation of the popularity over a broad range of facilities may present an interesting research issue.

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).

References

Nigel Collier. 2012. Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health*, 7(7):731–749. PMID: 22783909.

Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of AAAI 2015*. The AAAI Press.

Heng Ji, HT Dang, J Nothman, and B Hachey. 2014. Overview of tac-kbp2014 entity discovery and linking tasks. *Proc. Text Analysis Conference (TAC2014)*.

Asanobu Kitamoto and Takeshi Sagara. 2012. Toponym-based geotagging for observing precipitation from social and scientific data streams. In Gerald Friedland Liangliang Cao, editor, *Proceedings of the 2012 ACM Workshop on Geotagging and Its Applications in Multimedia, GeoMM'12 (co-located with ACM Multimedia 2012)*, pages 23–26. ACM, 11.

Jochen L. Leidner. 2007. Toponym resolution in text: Annotation, evaluation and applications of spatial grounding. *SIGIR Forum*, 41(2):124–126, December.

Jiwei Li, Alan Ritter, and Eduard Hovy. 2014. Weakly supervised user profile extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 165–174. Association for Computational Linguistics.

Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In Feifei Li, Mirella M. Moro, Shahram Ghandeharizadeh, Jayant R. Haritsa, Gerhard Weikum, Michael J. Carey, Fabio Casati, Edward Y. Chang, Ioana Manolescu, Sharad Mehrotra, Umeshwar Dayal, and Vassilis J. Tsotras, editors, *ICDE*, pages 201–212. IEEE.

Inderjeet Mani, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy. 2010. Spatialml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280.

S.E. Middleton, L. Middleton, and S. Modafferi. 2014. Real-time crisis mapping of natural disasters using social media. *Intelligent Systems, IEEE*, 29(2):9–17, Mar.

Kiyonori Ohtake, Jun Goto, Stijn De Saeger, Kentaro Torisawa, Junta Mizuno, and Kentaro Inui. 2013. Nict disaster information analysis system. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, pages 29–32. Asian Federation of Natural Language Processing.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 851–859, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wim Peters and Ivonne Peters. 2000. Lexicalised systematic polysemy in wordnet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA).

Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mhlhuser. 2013. A multi-indicator approach for geolocalization of tweets. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press.

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-2002)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. 2011. 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*, 6(5):e19467, 05.

Michael Speriosu and Jason Baldridge. 2013. Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1476. Association for Computational Linguistics.

István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1619–1629. Association for Computational Linguistics.

Wei Zhang and Judith Gelernter. 2014. Geocoding location expressions in twitter messages: A preference learning method. *J. Spatial Information Science*, 9(1):37–70.