

Location Name Disambiguation Exploiting Spatial Proximity and Temporal Consistency

Takashi Awamura[†] Eiji Aramaki[‡] Daisuke Kawahara[†]
Tomohide Shibata[†] Sadao Kurohashi[†]

[†] Graduate School of Informatics, Kyoto University

[‡] Design School, Kyoto University

awa@nlp.ist.i.kyoto-u.ac.jp, eiji.aramaki@gmail.com,
{dk, shibata, kuro}@i.kyoto-u.ac.jp

Abstract

As the volume of documents on the Web increases, technologies to extract useful information from them become increasingly essential. For instance, information extracted from social network services such as Twitter and Facebook is useful because it contains a lot of location-specific information. To extract such information, it is necessary to identify the location of each location-relevant expression within a document. Previous studies on location disambiguation have tackled this problem on the basis of word sense disambiguation, and did not make use of location-specific clues. In this paper, we propose a method for location disambiguation that takes advantage of the following two clues: spatial proximity and temporal consistency. We confirm the effectiveness of these clues through experiments on Twitter tweets with GPS information.

1 Introduction

As the volume of documents on the Web increases, technologies to extract useful information from them become increasingly essential. For instance, information extracted from social network services (SNS) such as Twitter and Facebook is useful because it contains a lot of location-specific information. To extract such information, it is necessary to identify the location of each location-relevant expression within a document.

However, many previous studies on SNS rely only on geo-tagged documents (e.g., (Han et al., 2013; Han et al., 2014)), which include GPS information,

but these represent only a small proportion of the total.¹ To extract as much location information as possible, it is important to develop a method that can estimate locations from numerous documents without GPS information.

Previous studies on location disambiguation made use of methods for word sense disambiguation and are based only on textual information, i.e., the bag-of-words in a document. It is, however, difficult to solve this problem using only textual information in a relatively short SNS document. For example, it is difficult to identify the location of “Prefectural Office Ave.” from the following document based only on word information.²

“I arrived at Prefectural Office Ave. from Shuri Station!”

In this paper, we propose a method that identifies the locations of location expressions in Twitter tweets on the basis of the following two clues: (1) spatial proximity, and (2) temporal consistency. Spatial proximity assumes that all locations mentioned in a tweet are close to one another. In the above document, for example, we would assume that “Prefectural Office Ave.” is “Prefectural Office Ave. (Okinawa)” using the proximity between “Shuri Station” and “Prefectural Office Ave. (Okinawa)” The other clue is temporal consistency,

¹SemioCast reported that GPS information is assigned to only 0.77% of all public tweets.

²Although it is possible to learn a clue from “Shuri Station,” which is located in Okinawa Prefecture, it would require a large amount of training data to learn such lexical clues for each target location expression.

which assumes that the locations in a series of tweets are near to each other.

In our experiments, we learn a location classifier for each ambiguous location expression in Japanese. Hereafter, we call an ambiguous location expression, such as “Prefectural Office Ave.,” a **Location EXpression (LEX)**, and a location to which a LEX points, such as <Prefectural Office Ave. (Okinawa)>, a **Location Entity (LE)**, which is linked to its GIS information. We call a LEX linked to multiple LEs an **ambiguous LEX**, which is the target of our location name disambiguation system. That is unambiguous LEXs are not our target, such as “Tokyo Tower,” which points the LE <Tokyo Tower>.

We define a set of LEXs and LEs on the basis of Japanese Wikipedia. Training data for the location classifiers are created from tweets containing GPS information. The resulting location classifiers can be applied to LEXs in any tweets or documents without GPS information.

Our novel contributions can be summarized as follows:

- two novel clues for location disambiguation are proposed,
- training data is automatically created from tweets with GPS information, and
- our method can identify LEs of LEXs in any documents without GPS information.

The remainder of this paper is organized as follows. Section 2 introduces related work, while Section 3 describes the resources used in this paper. Section 4 details our proposed method and Section 5 reports the experimental results. Section 6 concludes the paper.

2 Related Work

The location name disambiguation described in this paper is closely connected with Word Sense Disambiguation (WSD), and so studies on WSD are discussed here. We describe studies in location name disambiguation and in the significance of location names in social media.

2.1 Location Estimation

Location name disambiguation has been studied for a long time. It includes estimating one’s place of

residence and the entity of an ambiguous LEX. Several approaches have been proposed. Although one of the simplest and most reliable is to use IP addresses, many problems can occur, e.g., the IP address of past content cannot be accessed, and this approach is becoming increasingly ineffective with the increased use of portable terminals. As a result, location name disambiguation should now focus on procedures that consider the original text. As information references, Web pages and change logs in Wikipedia have been used as the basis of location name disambiguation. These resources are homogeneous and manageable. In contrast, the numerous data on SNS often contain noise, which makes disambiguation unmanageable.

A number of studies have investigated location name disambiguation. Han et al. (2012) extracted location-indicative words from tweet data by calculating the information gain ratios. Their paper states that the words improved the estimation performance of the users’ location. They concluded that the procedure requires relatively little memory, is fast, and could potentially be used by lexicographers to extract location-indicative words. Backstrom et al. (2008) developed a probabilistic framework to quantify the spatial variation manifested in search queries. This allowed them to obtain a measure of spatial dispersion that indicates regional information.

Adams and Janowicz (2012) estimated geographic regions from unstructured, non georeferenced text by computing a probability distribution over the Earth’s surface. Their methodology combines natural language processing, geostatistics, and a data-driven bottom-up semantics. Chandra et al. (2011) estimated a city-level user location based purely on a content of tweets, which may include reply-tweet information, without the use of any external information, such as a gazetteer, IP information etc. Chang et al. (2012) proposed two unsupervised methods based on notions of Non-Localness and Geometric-Localness to prune noisy data from tweets. Kinsella et al. (2011) created language models of locations using coordinates extracted from geotagged Twitter data. Van Laere et al. (2014) assigned coordinates to Flickr photos and to Wikipedia articles with Kernel Density Estimation and Ripley’s K statistic. Although these studies have estimated

location names from location-indicative words or the degree of popularity, most studies neglect spatial proximity, i.e., the distance between two locations, and temporal consistency, i.e., previous tweets from the same user. This paper proposes a new method of location name estimation that considers both spatial proximity and temporal consistency.

2.2 The Importance of Location Name in Social Media

Several researchers have attempted to extract information from SNS such as Twitter. Sakaki et al. (2010) detected earthquakes from tweets containing geographic information system (GIS) information. They judged whether the tweet was posted just after an earthquake using a support vector machine (SVM), and determined the seismic center from the formatted tweets. In addition, they developed a system that raises the alarm about an earthquake from the predicted results. Bollen et al. (2011) extracted the social mood, and predicted the stock price fluctuation N days from the day of observation by using evaluated data of the 'mood-related' dictionary. As a result, they concluded that they could show the 3 days from the 'calm-mood' day might be able to predict the stock price fluctuation. Aramaki et al. (2011) predicted an influenza epidemic from tweets. They showed the possibility of information extraction from the tweets that reflects the actual world's situation by using language processing technologies. Boyd et al. (2010) examined a practice of retweeting as a way by which participants can be "in a conversation." Paul and Dredze (2011) considered a broader range of public health applications for Twitter and showed quantitative correlations with public health data and qualitative evaluations of model output. Baldwin et al. (2013) explored how linguistically noisy or otherwise it is over a range of social media sources empirically over popular social media text types, in the form of YouTube comments, Twitter posts, web user forum posts, blog posts and Wikipedia. Yin et al. (2012) constructed a system architecture for leveraging social media to enhance emergency situation awareness with high-speed text streams retrieved from Twitter during natural disasters and crises.

In these researches, the location of an SNS document plays an important role in extracting informa-

tion, and in most cases, rely on GPS function connected to the tweets. However, in fact, there are less than 1% of the entire tweets that are connected to GPS. In order to enhance the accuracy of such research, it is necessary to use the framework that enables to discriminate the location out of the texts and words of the tweets that do not contain GPS information.

3 Resources

3.1 LEX Database

First, it is necessary to define the LEXs and LEs handled in this study. We focus on LEXs and LEs that have GIS information on Wikipedia. In this paper, we call the database of LEXs and LEs **LEX database**, and use two methods to obtain the LEX database from Wikipedia according to the type of GIS data:

- Infobox
- Latitude/longitude information

3.1.1 Infobox

The Infobox is a meta-template on a Wikipedia page (as shown in Figure 1). Infobox, which the article of a location name has, sometimes contains its address and latitude/longitude. We extract entries that have such Infoboxes as LEs.

We ran this process on the Japanese Wikipedia, and extracted 759 LEXs and 884 LEs as a result.

3.1.2 Latitude/Longitude Information

The latitude/longitude information is often given at the top of a Wikipedia article about a location (as shown in Figure 2). We extract LEs and LEXs from Wikipedia articles that contain such GIS information. We extracted 17,140 LEXs and 17,426 LEs by applying this method to the Japanese Wikipedia.

We merged these two databases to generate our LEX database, deleting duplicate LEs in the process. In total, we obtained 17,724 LEXs and 18,256 LEs. Table 1 lists the LEs of "Prefectural Office Ave." Table 2 lists the frequencies of LEXs and LEs according to the number of LEs for a LEX. From this table, we can see that we have 462 ambiguous LEXs, which correspond to 994 LEs.

Times Square (Detroit People Mover)

From Wikipedia, the free encyclopedia

Times Square is a Detroit **People Mover** station in Downtown Detroit, Michigan. It is located on Grand River Avenue between Cass and Washington Boulevard. The station takes its name from nearby Times Square, which in turn, took the name from the defunct *Detroit Times* newspaper formerly headquartered there. It also serves as headquarters for the People Mover system and houses a maintenance facility.

The DPM's 12 car fleet are stored in an indoor carhouse at Times Square.

Stations [edit]

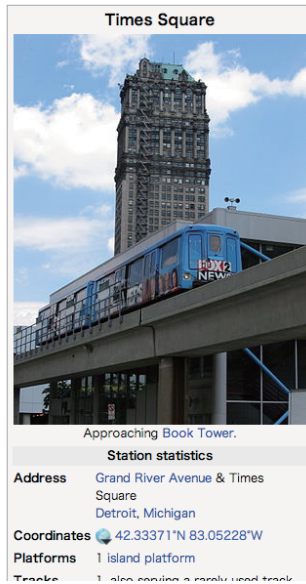


Figure 1: Infobox information

ID	LE	Lat	Long
1	Prefectural Office Ave. (Hyogo)	34.69	135.18
2	Prefectural Office Ave. (Chiba)	35.60	140.12
3	Prefectural Office Ave. (Toyama)	36.69	137.20
4	Prefectural Office Ave. (Hiroshima)	34.39	132.45
5	Prefectural Office Ave. (Ehime)	33.84	132.76
6	Prefectural Office Ave. (Kohchi)	33.55	133.53
7	Prefectural Office Ave. (Okinawa)	26.21	127.67

Table 1: LEs for the LEX “Prefectural Office Ave.”

In this study, a location name with parenthesis is used for an LE, such as <Times Square (Detroit People Mover)> and <Times Square (Hong Kong)> shown in Figure 1, 2, and a string without the part in brackets is used for a LEX, such as “Times Square.”

3.2 Corpus for Location Name Disambiguation

The disambiguation of LEXs requires a corpus in which each LEX is assigned to an LE. We extract this from Twitter data with GIS information. For example, given a tweet “Let’s meet at the Prefectural Office Ave.” that has GIS latitude and longitude information indicating Okinawa, it is natural that the “Prefectural Office Ave.” in the tweet indicates the LE <Prefectural Office Ave. (Okinawa)>. Therefore, we assign LEs to LEXs in tweets based on their GIS information using the following method.

Times Square (Hong Kong)

From Wikipedia, the free encyclopedia

Coordinates: 22°16′42″N 114°10′56″E﻿ / ﻿﻿ / ﻿

Times Square (Chinese: 時代廣場) is a major shopping centre and office tower complex in Causeway Bay, Hong Kong Island, Hong Kong.

The complex, owned by Wharf Properties Limited, part of The Wharf (Holdings) Limited group, was opened in April 1994.

Contents	[hide]
1	History
2	Project configuration
2.1	Shopping mall
2.2	Office buildings
3	Christmas and New Year celebrations
4	Public open space controversy
5	Transport
6	Cultural reference
7	See also
8	References
9	External links

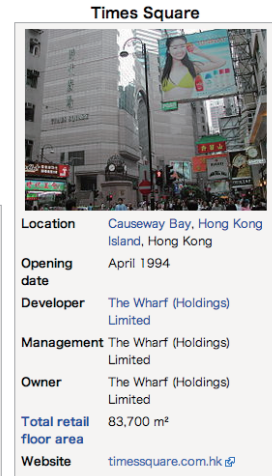


Figure 2: Latitude/longitude information

# of LEs	LEX	LE
1	17,262	17,262
2	412	824
3	38	114
4	8	32
5	2	10
7	2	14
Sum	17,724	18,256

Table 2: Statistics of LEX database

- **STEP 0 (pre-processing):** Preparation of tweets
We obtained tweet data containing GIS information from 2011/7/15 to 2012/7/31. We removed duplicate tweets.
- **STEP 1:** Extraction of tweets including LEXs
Tweets including ambiguous LEXs are extracted based on the LEX database described in Section 3.1. Tweets including unambiguous LEXs are not used for our target tweets but used for the clues of temporal consistency described in Section 4.3. This process searches for LEX strings within a tweet, and aggregates such tweets for each LEX. If several ambiguous LEXs are included in a tweet, this tweet is used for each LEX. For example, “I’ll go to Motomachi station from Prefectural Office Ave.” is used for “Motomachi station” and “Prefectural Office Ave.”
- **STEP 2:** Assignment of LEs
An LE is assigned to tweets for each LEX ex-

tracted in STEP 1. This process is conducted on the basis of the GIS information in the tweet and the LEs of the target LEX. Our idea is that if the distance between the tweet GIS and an LE GIS is short, the LEX in the tweet may point to this LE. For example, if the GIS of the tweet including “Prefectural Office Ave.” is near <Prefectural Office Ave. (Okinawa)>, this “Prefectural Office Ave.” may point to <Prefectural Office Ave. (Okinawa)>. In this paper, we set the distance threshold for the judgment of LEs to 10km. That is, if the distance between the tweet GIS and an LE GIS is less than 10 km, this LE is assigned to the tweet; otherwise this tweet is discarded. If the distance of several LEs is less than 10 km, the LE with the shortest distance is assigned to the tweet.

Approximately 180,000 tweets including ambiguous LEXs were obtained. Out of 462 ambiguous LEXs in the LEX database, 353 contain at least one tweet. We employed them as the gold standard data used in our experiments.

One of our novel contributions of this study is that we automatically constructed this large-scale corpus with GIS information, whereas previous studies on toponym resolution created a corpus by hand (Leidner, 2008).

4 Method for Location Name Disambiguation

We propose a method for location name disambiguation in tweets. Our approach automatically distinguishes LEs for a LEX in a tweet using a machine learning algorithm: SVM. The SVM classifiers are generated for each LEX. Each SVM classifier has the following features.

4.1 Baseline Features

We use the following two features as baseline features:

- (1) Lexical feature: bag of words in the tweet
- (2) Majority feature: frequency of LEs

4.2 Spatial Proximity Features

The distance between a target ambiguous LEX and an unambiguous LEX in the tweets is used for the

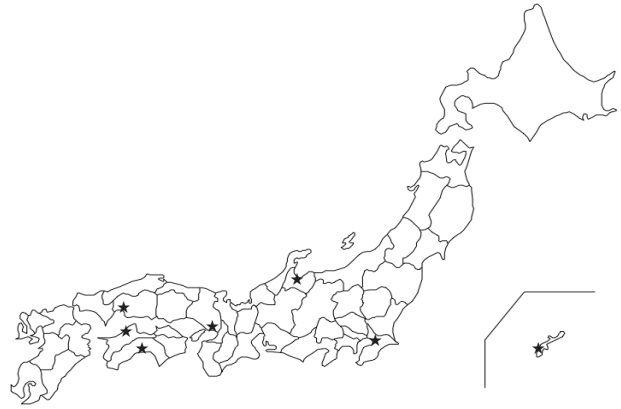


Figure 3: Locations of “Prefectural Office Ave.” in Japan spatial proximity features. An example is shown below.

- (1) It takes about 20 minutes to get from Shuri station to Prefectural Office Ave.

The ambiguous LEX “Prefectural Office Ave.” has seven LEs (shown in Figure 3).

In this example, it is difficult to estimate the LE based only on the lexical information. However, the relation between the LEX and other unambiguous locations in the same text provides a clue for the disambiguation of the LEX. In general, related LEXs tend to exist alongside the target LEX. Although the words in tweets may be learned implicitly from this relation by SVM, they cannot also be expected to occur. Thus, our method explicitly uses the distance between two locations as the relation. We assume that the distances between the LE of the target LEX and other LEs are short. For example, in the above example of “Prefectural Office Ave.,” <Shuri station> is relatively close to <Prefectural Office Ave. (Okinawa)>, but is not near <Prefectural Office Ave. (Chiba)>. Thus, it can be estimated that the LE of “Prefectural Office Ave.” is <Prefectural Office Ave. (Okinawa)>.

To assign the spatial proximity features to a tweet, we first check whether the tweet includes LEXs. If the LEXs are unambiguous, we then calculate the distance between the unambiguous LE and each target LE.³ Features depending on the distance are assigned to the tweet. If the LEXs are ambiguous, spatial proximity features are not used, because the LEs

³If there are multiple unambiguous LEs in the tweet, all of these are considered as features.

indicated by the LEXs cannot be determined.

For example, when a tweet with “Prefectural Office Ave.” contains the unambiguous LEX “Shuri Station,” the distance between <Shuri Station> and each LE indicating “Prefectural Office Ave.” is calculated. If the distance between <Shuri Station> and <Prefectural Office Ave. (Okinawa)> is 0~10 km and that between <Shuri Station> and <Prefectural Office Ave. (Chiba)> is 500~1,000 km, these distances are used as different features. The number of spatial consistency features is ld , where l is the number of LEs for the target LEX and d is the number of distance bins, which are described in Section 5.

4.3 Temporal Consistency Features

Until now, we have considered only a single target tweet to estimate locations. However, the target tweet sometimes contains few useful clues for LEX disambiguation because the tweet is too short. Therefore, this paper considers the preceding tweets posted in the previous t hours. The baseline features and the spatial proximity features are also extracted from these preceding tweets. An example is shown below.

- (2) I arrived at the Prefectural Office Ave.

Its preceding tweets are as follows:

- (3) I’m going to take an airplane. I’m looking forward to Okinawa!
 (4) I arrived in Okinawa!
 (5) I’m heading for Shuri Station by Yui Rail.

In such a case, useful information for location estimation can be obtained by considering these preceding tweets. For example, “Okinawa” is related to <Prefectural Office Ave. (Okinawa)>, and <Shuri Station> is near <Prefectural Office Ave. (Okinawa)>. Based on such information, it can be estimated that the LE of “Prefectural Office Ave.” is <Prefectural Office Ave. (Okinawa)>.

It is necessary to determine the time threshold t . This is because extremely old tweets are hardly related to the target tweet. We will discuss this issue in Section 5.

Method	Settings
SP (0~10, 100~500 km)	+10~100 km +10~50, 50~100 km +10~100, 500~1000 km +10~50, 50~100, 500~1000 km
Ts	Indefinite ~24 h ~12 h ~6 h ~3 h ~1 h

Table 3: Settings for SP and Ts

5 Experiments and Discussion

5.1 Experimental Settings

We create an SVM classifier for each LEX to solve location name disambiguation with the features described in Section 4. This classifier identifies the LE for an ambiguous LEX included in a tweet. Since location name disambiguation is a multi-class identification problem, we use the one-versus-the-rest method for the SVM classifier. For the gold-standard data, we used 70,184 tweets including the LEXs that are associated with ten or more tweets from the corpus described in Section 3.2. We conducted 5-fold cross-validation using this data. We adopted TinySVM,⁴ an SVM package with a quadratic polynomial kernel. For the segmentation of Japanese words, we used the Japanese morphological analyzer JUMAN.⁵

5.2 Methods for Comparison

We compare the following four methods in this study:

- Baseline (B): This method uses only the following two features: (1) lexical features, and (2) majority features. We used the base form of words in a tweet as SVM features. Here, we used only high-frequency words (top 100,000). We regard the frequency of a word in a tweet as the lexical feature.
- +Spatial Proximity (+SP): This method uses the baseline features and the spatial proximity features. The spatial proximity features are generated from the distance between the target LE and another unambiguous LEX (LE)

⁴<http://chasen.org/~taku/software/TinySVM/>

⁵<http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN>

mentioned in the same tweet (as described in Section 4.2). We examined four sets of distance bins as listed in Table 3 (default: 0~10, 100~500 km). Each feature of spatial proximity is considered separately according to the distance bins. The values are the number of LEs in the same tweet that satisfy the distance condition.

- **+Temporal Consistency (+TC):** This method uses baseline features and temporal consistency features. The temporal consistency features are generated from recent tweets (maximum of three), as described in Section 4.3. This feature disregards non-recent tweets. We investigated six definitions of recency as listed in Table 3.
- **+Spatial Proximity +Temporal Consistency (+SP+TC):** This method uses all features, i.e., baseline, spatial proximity, and temporal consistency features. The spatial proximity features are also generated from the preceding tweets that are used to generate the temporal consistency features.

5.3 Evaluation

The accuracy $\frac{s}{c}$ is calculated from the system output and the correct LEs, where s is the number of tweets whose output had the correct LE and c is the total number of tweets considered. Moreover, the accuracy is calculated separately for each number of tweets per LEX (10~100: rare LEX, 100~1,000: intermediate LEX, 1,000~: common LEX, 10~: all).

5.4 Experimental Results and Discussions

The results for all methods are compared in Table 4 with the following proximity and consistency features:

- 0~10, 10~100, 100~500 km
- ~6 h

The Majority Baseline (MB) is a baseline method that outputs the most frequent LE for each LEX.

Table 4 lists the accuracy of the estimated LEs considering spatial proximity and temporal consistency. In particular, considering the proximity improves the accuracy, regardless of the number of tweets for each LEX. Although the consideration of

# of Tweets (Sum)	Method	# of Correct	Accuracy
10~100 (4,891)	MB	4,171	0.8528
	B	4,485	0.9170
	+SP	4,515	0.9231 †
	+TC	4,491	0.9182 †
	+SP+TC	4,520	0.9241 †
100~1000 (25,758)	MB	22,477	0.8726
	B	24,725	0.9599
	+SP	24,752	0.9609
	+TC	24,708	0.9592
	+SP+TC	24,737	0.9604
1000~ (39,535)	MB	36,896	0.9332
	B	39,041	0.9875
	+SP	39,054	0.9878
	+TC	39,036	0.9874
	+SP+TC	39,054	0.9878
10~ (70,184)	MB	63,544	0.9054
	B	68,251	0.9725
	+SP	68,321	0.9735 †
	+TC	68,235	0.9722
	+SP+TC	68,311	0.9733 †

“†” means the superiority to B estimated at the 5% significance level and “‡” means that at the 1% level.

Table 4: Main results

temporal consistency also improves accuracy for rare LEXs. the accuracy is below the baseline for common LEXs. The accuracy considering both features outperforms the baseline by 7.13% for rare LEXs. In addition, a sign-test was adopted to demonstrate the significance of the results. This test was performed using R.⁶ “†” means the superiority to B estimated at a significance level of 5%, and “‡” means that at the 1% level. This test shows the significance of the proposed method, particularly for rare LEXs. Moreover, the accuracy with all tweets verifies the significance of the proposed method compared to the baseline.

The accuracy did not improve for common LEXs because of an imbalance in the tweet data. This study only uses tweet data that include LEXs and GIS information. Therefore, the LEs of the tweets are imbalanced for each LEX. The high accuracy of MB suggests this imbalance depends on the number of tweets for each LEX. Moreover, most tweets with GIS information are generated automatically by companies such as Foursquare. As a result, high accuracy is obtained in many cases without considering the proximity or consistency. Although this study used only tweets with GIS information, the accuracy could clearly be improved using tweets with-

⁶<http://cran.r-project.org/>

# of Tweets (Sum)	Proximity	# of Correct	Accuracy
10~100 (4,891)	+10~100 km	4,515	0.9231
	+10~50, 50~100 km	4,513	0.9227
	+10~100, 500~1000 km	4,519	0.9239
	+10~50, 50~100, 500~1000 km	4,520	0.9241
100~1000 (25,758)	+10~100 km	24,752	0.9599
	+10~50, 50~100 km	24,758	0.9612
	+10~100, 500~1000 km	24,744	0.9606
	+10~50, 50~100, 500~1000 km	24,746	0.9607
1000~ (39,535)	+10~100 km	39,053	0.9878
	+10~50, 50~100 km	39,069	0.9882
	+10~100, 500~1000 km	39,064	0.9881
	+10~50, 50~100, 500~1000 km	39,065	0.9881

Table 5: Comparison of SP features

# of Tweets (Sum)	Terms	# of Correct	Accuracy
10~100 (4,891)	indefinite	4,429	0.9055
	~24 h	4,491	0.9182
	~12 h	4,493	0.9186
	~6 h	4,491	0.9182
	~3 h	4,494	0.9188
100~1,000 (25,758)	indefinite	24,694	0.9587
	~24 h	24,700	0.9589
	~12 h	24,709	0.9593
	~6 h	24,708	0.9592
	~3 h	24,718	0.9596
1,000~ (39,535)	indefinite	38,988	0.9862
	~24 h	38,988	0.9873
	~12 h	39,036	0.9874
	~6 h	39,036	0.9874
	~3 h	39,033	0.9873
~1 h	39,034	0.9873	

Table 6: Comparison of TC features

out GIS information.

A comparison of the results for various proximity levels is shown in Table 5. As shown in Table 5, the accuracy of location name disambiguation with rare LEXs improves with the addition of the 500~1000 km bin. However, when many tweets are considered, the accuracy improves with the addition of 10~50 and 50~100 km bins. This implies that the LE estimation requires additional information when there are few tweets, and less information when many tweets are available.

A comparison of the results for different degrees of temporal consistency is shown in Table 6. Although there were few remarkable results, it is clear that the accuracy does not improve significantly when older tweets are considered. In particular, the poorest accuracy was achieved when specific terms

are not defined. This shows the validity of considering specific terms.

6 Conclusions and Future Work

In this paper, we presented a method for location name disambiguation for text snippets on SNS. We considered both the spatial proximity and temporal consistency to produce the estimates of LEs. As a result, our method substantially outperformed the baseline method that considers only lexical information. More specifically:

- Considering the spatial proximity improves the accuracy
- Considering the temporal consistency with many tweets improves the accuracy
- Considering both of the above outperforms the baseline by 7.13%

In future work, first, we plan to further investigate the cause of the decrease in accuracy when the temporal consistency feature considers many tweets.

Second, in this paper, only tweets including unambiguous LEXs are used to calculate the proximity feature for the target LEX. However, tweets including ambiguous LEXs could also be used if the LEXs have been disambiguated in advance.

In addition, we estimated the LEs of ambiguous LEXs, although the location estimation has several problems. One concerns whether the user posting the tweet including the LEX is actually at that location. Solving this problem is necessary for some applications specializing in GIS information. In future work, we aim to solve this problem using the proposed spatial proximity and temporal consistency.

References

- Benjamin Adams and Krzysztof Janowicz. 2012. On the geo-indicativity of non-georeferenced text. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press.
- Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. 2008. Spatial variation in search engine queries. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 357–366, New York, NY, USA. ACM.
- Timothy Baldwin, Paul Cook, Marco Lui, Andrew Mackinlay, and Li Wang. 2013. How noisy social media text, how different social media sources?
- J. Bollen, H. Mao, and X. Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, pages 1–8.
- Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA. IEEE Computer Society.
- S. Chandra, L. Khan, and F.B. Muhaya. 2011. Estimating twitter user location using social interactions—a content based approach. In *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*, pages 838–843, Oct.
- Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. 2012. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 111–118, Washington, DC, USA. IEEE Computer Society.
- Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, December.
- Bo Han, Paul Cook, and Timothy Baldwin. 2013. A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *J. Artif. Intell. Res. (JAIR)*, 49:451–500.
- Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. “I ’m Eating a Sandwich in Glasgow ’’: Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 61–68, New York, NY, USA. ACM.
- Jochen L Leidner. 2008. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal Press, Boca Raton, FL, USA.
- Michael Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *5th International Conference on Weblogs and Social Media*, pages 265–272. AAAI Press.
- Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. ACM.
- Olivier Van Laere, Jonathan Quinn, Steven Schockaert, and Bart Dhoedt. 2014. Spatially aware term selection for geotagging. *IEEE Trans. on Knowl. and Data Eng.*, 26(1):221–234, January.
- Jie Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. 2012. Using social media to enhance emergency situation awareness. *Intelligent Systems, IEEE*, 27(6):52–59, Nov.