Novel Location De-identification for Machine and Human

Katsuya Taguchi
Nara Institute of Science and Technology
Ikoma, Nara
taguchi.katsuya.tb3@is.naist.jp

Eiji Aramaki
Nara Institute of Science and Technology
Ikoma, Nara
aramaki@is.naist.jp

ABSTRACT

In recent years, the protection of personal information has drawn much attention, requiring an advanced technology on de-identification to remove personal information from data. Among various personal information such as personal names, phone numbers, and so forth, this study focuses on location information. The conventional approaches to protect location information are to remove address expressions. However, there are complicated cases in which location information can be guessed with unexpected combinations of non-address words. For example, we can guess '*the most traditional city in Japan*' is Kyoto. To our knowledge, such location-inferable expressions have not been dealt with. This study handles this phenomenon by using a location classifier. In addition, we assume two levels of location inferance; (1) inferable by machine and (2) inferable by human. To build the first-level inferance, we employed a collection of tweets with geo-tags. To build the second-level inferance, we created a new corpus with a flag for whether tweets are location-inferable by human or not. By using the two types of corpora, we classified texts into several categories such as a machine-inferable but human-non-inferable tweet, and so on. We also could obtain de-identified tweets by iterations of removing the highest weighted words for classifiers. We believe our novel concepts of de-identification are essential for various privacy protection.

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; • **Social and professional topics** → **Identity theft**; *Social engineering attacks*; • **Computing methodologies** → Learning linear models;

KEYWORDS

De-identification, Location inference, SNS, Twitter, Natural language processing

UISTDA '18, March 11, 2018, Tokyo, Japan

1 INTRODUCTION

In recent years, de-identification techniques to delete sensitive personal information have been studied actively because of the growing interest in privacy protection. In most automatic de-identification technologies, sensitive personal information is regarded as identical to proper expressions such as personal names, organization names, phone numbers, ID numbers, and addresses. Therefore, Named Entity Recognition (NER) techniques have been applied to de-identification. As described in this paper, this conventional approach is designated as NER-based de-identification.

Actually, NER-based de-identification has an important limitation: an address is identifiable from non-named entity expressions. Sometimes, the combination of general terms can be a strong clue for identifying a specific location. Consider the following sentence: '*I'm excited to have dinner with my colleague on the riverbed*!' Because the *riverbed* is a famous spot in Kyoto and the location of *riverbed* in Kyoto is well-known, most Japanese people can guess that the person behind the tweet is located in Kyoto. This limitation of NER-based approaches becomes an important issue because many people unintentionally expose their location information to others. Sometimes the knowledge might be used illegally.

This study specifically examines automatic de-identification of messages in Twitter in terms of their location information. Our de-identification method has two novel features.

- This study handles location-inferable expressions (not only proper expressions but also non-proper expressions).
- This study assumes two levels of location inference: (1) inferable by machine and (2) inferable by humans.

Using the two viewpoints of inference, we were able to design several levels of de-identification: a level of a machine-inferable but human-non-inferable tweets, and so on. It is noteworthy that the proposed method is independent of any specific language.

The remainder of this paper is organized as follows. First, we construct a classifier to infer tweet locations using geotagged tweets in Twitter (Section 4). Next, we investigate

Table 1: Work related to location inference

	Home location	Tweet location	Mentioned location
Human network	Kong et al. [1]	Sadilek et al. [2]	Hua et al. [3]
Tweet content	Yamaguchi et al. [4] (word-centric) Cha et al. [5] (location-centric)	Flatow et al. [6] (word-centric) Kinsella et al. [7] (location-centric)	Li et al. [8]
Tweet context	Efstathiades et al. [9]	Dredze et al. [10]	Fang et al. [11]

whether the classifier can infer tweet locations that are de-identified by humans (Section 5). Then, we tag the tweets with whether a human can infer the locations (what we call **feasibility of location inference**). We compare the difference between the classifier and human (Section 6). Finally, we present a de-identification method considering the combination of words (Section 7).

2 RELATED WORK

Location Inference

Many methods for location inference have been proposed to date. They are classifiable by two aspects: location types to be estimated and material types to be used for location estimation, as presented in Table 1.

As for location types, roughly three types of locations have been considered to date as shown in Table 1: user home locations, tweet locations, and described locations. Home location is a location where a user lives or spends much time, including the address of a user's home or office. Tweet location is one from which a user has posted a tweet. A mentioned location is one that a user has described in a tweet. This paper represents an attempt to estimate tweet locations, which are our target of de-identification.

For location inference, three types of materials have been used as shown in Table 1: human network, tweet content, and tweet context. A human network is a relation between users in social networking services such as follower or followee in Twitter. Tweet content represents the content of a tweet message. Tweet context is information associated with a tweet such as a time stamp, geo-tag, or time zone. When inferring locations using tweet content, there are two major approaches distinguished by probabilistic models. One is called the *word-centric* model, calculating the probability $p(l|W)$ that a location l is labeled to a set of words W. The other is called a *location-centric* model. It calculates the probability $p(d|l)$ that each location's label l outputs a tweet document d. In this paper, the *word-centric model* is applied to analyze tweet contents and to construct a classifier to estimate a tweet's location.

The study by Flatow et al. [6] is similar to ours in that they attempted to infer tweet locations with the word-centric model and tweet content. However, the method cannot estimate locations that are identifiable by unexpected word combinations because a classifier is constructed using a word list that is appropriate to each area. By contrast, we propose a method to infer locations by considering word combinations.

De-identification

In the medical field, de-identification of patient data has been studied actively. A conventional approach, Named Entity Recognition (NER) based de-identification, deletes proper expressions that are capable of specifying individuals such as proper nouns: phone numbers and addresses. However, NER-based de-identification is insufficient for de-identifying location information. Moreover, in the medical field, a law exists to protect individuals' medical records and other personal health information: Health Insurance Portability and Accountability Act (HIPAA) [1], which was approved in the U.S.A. in 1996. As for de-identification of social media contents including messages with location information, however, no criteria correspond to HIPAA. This paper therefore sets criteria for the de-identification of tweet locations by conducting experiments related to manual de-identification.

3 DATASET

This section describes our dataset consisting of tweets with location information and area division.

Tweets

Tweet data consist of 298,711 Japanese messages with geo-tags (hereinafter called 'tweets') posted within the central region of Kyoto City, Japan (latitude range = [34.93, 35.12] and longitude range = [135.67, 135.83]). This region includes popular landmarks, train stations, castles, shrines, temples, and so on, yielding a diverse mix of tweets. The tweets were collected about for a year between 2011/7/14 and 2012/7/31.

The tweet data are divided into training data and test data. Training data consisting of 179,227 tweets (60% of all data) are used to construct a classifier as described in Section 4.

[1] https://www.hhs.gov/hipaa/index.html

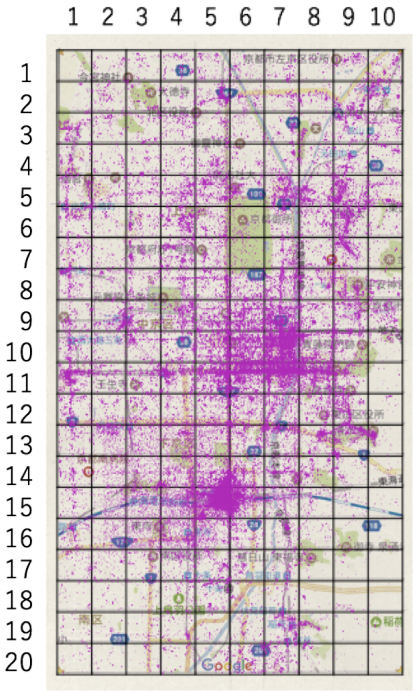


Figure 1: Geographical distribution of tweets in the central region of Kyoto City. The region is divided into 200 areas.

The test data consist of 119,484 tweets (40% of all data) used to evaluate the classifier's performance in Section 4. Some test data are used for experiments in Sections 5 and 6.

Area Division

The region described in Section 3.1, the central region of Kyoto City, Japan, was divided into 200 (= 20×10) areas ($a_{1\,1}, ..., a_{20\,10} \in A_{kyoto}$), as presented in Figure 1. Each area is 501 m × 547 m. This division was useful to separate two consecutive stations (Hankyu Kawaramachi Station and Hankyu Karasuma Station) into two areas. Both areas are located near the Hankyu Kawaramachi Station, which is well known as the busiest downtown area in Kyoto. Therefore, this manner of division is reasonable.

Figure 1 presents the geographic distribution of 298,711 tweets for the selected areas. 39,078 tweets (13.1% of all tweets) were posted around the area $a_{15\,5}$, where Kyoto Station (Kyoto's largest train station) is located. By contrast, only two tweets were posted in area $a_{17\,10}$, which is located southeast of Miterasennyuji Temple.

4 CONSTRUCTION OF LOCATION CLASSIFIER

This section describes a method to construct a classifier that estimates a tweet location and which shows the classifier performance. Each text is split into words using a Japanese morphological analyzer, MeCab[2]. All uni-grams and bi-grams are used as features for a bag-of-words representation. They are converted into vectors and are used for the training data. Each element of a vector was one or zero according to whether each feature appeared or not. Noises such as URLs (e.g. "https://XXX"), hashtags (e.g. "#hashtag"), or mentions to other users (e.g. "@username") were removed from each text[3]. Correct answer labels are set to each area (200 classes in total) and are attached to each tweet based on its geo-tag. The classifier is constructed based on a linear model trained by logistic regression.

To evaluate the constructed classifier, the test data were classified into 200 classes. Results show that the accuracy for the test data was 47.2%. If the classifier always outputs the area $a_{15\,5}$ having the highest tweet density in the training data, then the accuracy for the test data is 11.6%. Also, 47.2% is modestly high in spite of its simple structure.

5 PRELIMINARY EXPERIMENT: MANUAL DE-IDENTIFICATION

When using the classifier constructed in Section 4, it is necessary to define the state: 'a tweet is de-identified.' This section describes an experiment by which the state is defined.

Materials and Procedure

To define the state that a tweet is de-identified, the manually annotated corpus was created. 500 tweets from the test data were de-identified manually. First, participants observe each tweet and infer its location as precisely as possible. The participants are allowed to use search engines, etc. Then, they delete the minimum number of morphemes in a tweet until they ascertain that the tweet's location becomes ambiguous.

In this preliminary experiment, two annotators with knowledge about Kyoto City independently annotated 500 tweets of the test data. The tweet below is an example of the annotated tweets. Words to be deleted are crossed off. In this example, the annotators considered that Tweet (1) was de-identified by deleting '御池 (*Oike*)' and 'マザーズハローワーク (*Mother's Hello Work*)'.

(1) 烏丸~~御池~~プラザが本チャンやないんか？
@~~マザーズハローワーク~~烏丸~~御池~~
(*Is not the Karasuma ~~Oike~~ Plaza main? @~~Mother's Hello Work~~ Karasuma* 御池~~*Oike*~~)

Then, a threshold determining whether a tweet is de-identified or not is defined using the annotated tweets. Given a de-identified tweet, the classifier calculates the

[2]http://taku910.github.io/mecab/
[3]https://github.com/s/preprocessor

probability of location inference when the tweet is assigned to the 200 areas, respectively. The maximum of the 200 probability values can be regarded as a reference to the tweet's de-identification. Finally the average of the maximum values of the probability for all annotated tweets is used as the threshold. The tweets for which the probability is below the threshold were regarded as being de-identified.

Results and Discussion

The threshold value was set to 0.37 from the preliminary experiment's result. Therefore, the tweets for which the probability was less than 0.37 were regarded as being de-identified.

However, for some tweets, the classifier outputs show high probability but the annotators were uncertain about their location, or vice versa. Because of such a discrepancy, probably one can make two types of inference for de-identification. One is to prevent inference of the location itself. The other is to prevent the assumption that a location can be inferred. In the next section, we examine another classifier to infer the feasibility of location inference.

6 FEASIBILITY OF LOCATION INFERENCE

This section describes a preliminary experiment to construct a classifier that infers the feasibility of location inference and the actual construction.

Materials and Procedures

To construct a classifier that infers the feasibility of location inference, a corpus annotated with the feasibility of location inference is generated. We first used 1,000 tweets selected randomly from the test data in Section 3.1. Then, binary classification tasks were conducted according to whether or not the locations can be inferred. To gather a large amount of experimental cooperation, the tasks were conducted through crowdsourcing. 100 participants answered each tweet as to whether or not the location can be inferred. The tweets for which 10% or more participants answered that they can be inferred were defined as tweets with feasibility of location inference. The others were treated as those without feasibility of location inference.

Results and Discussion

246 of 1,000 tweets showed the feasibility of location inference. Considering the two classification methods, whether the classifier in Section 4 can infer locations of tweets and whether tweets have feasibility of location inference, or not, the 1,000 tweets were classified into four classes. The results are presented in Table 2.

For some tweets, the classifier can infer their location, but those without feasibility of location inference are presented below.

Table 2: Inference of feasibility of location inference by the constructed classifier (machine) and human

		Classifier	
		infearable	not inferable
Human	inferable	216	30
	not inferable	258	496

(2) 風が強いです (>_<) 今日も明るく元気にお昼の営業開始です！
(*The wind is so strong* (>_<)*. I am about to start my lunch-hour business brightly and cheerfully as usual!*)

(3) おはようございます (^^) 今日の日中は雨予報ですね。 気温も 20 ℃まで行かないようです。 今日も明るく元気に！忙しく楽しい一日になるよう頑張ります p(^_^)q
(*Good morning* (^^)*. It is supposed to rain during the day. The temperature will not reach 20◦C. Let's be bright and cheerful! I try to be busy and enjoy my day* p(^_^)q*.*)

These tweets include fixed phrases for advertising stores, e.g. the latter part of Tweet (2), *'I am about to start my lunch-hour business brightly and cheerfully as usual!'* It seems that several tweets with typical phrases by a specific store are included in the training data. However, humans cannot read and learn so many tweets. Therefore, they believe that such tweets have no feasibility of location inference. The tweets below are examples for which the classifier cannot infer their location, but humans determine that they have feasibility of location inference.

(4) やっとお昼ご飯。つばめ
(*Finally, lunch time. Tsubame*)

(5) 河合塾の向かいのサブウェイなう！
(*I am at the subway station across the street from Kawaijuku now!*)

Tweet (4) is a case in which the proper noun 'つばめ (*Tsubame*)' is also a common noun. Considering such cases, data tagged with feasibility of location inference are apparently necessary. Tweet (5) represents a case in which the location is inferable by a combination of '河合塾 (*Kawaijuku*)' and 'サブウェイ (*Subway*)'.

Construction of Classifier for Inferring Location Inference Feasibility

A classifier was constructed with 1,000 tweets tagged using feasibility of location inference. Of the 1,000 tweets, 900

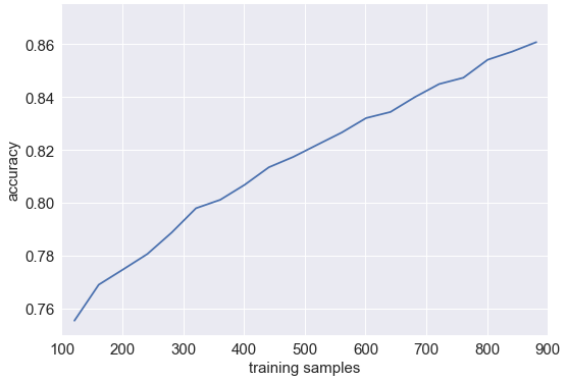


Figure 2: Relation between the number of training samples and the accuracy.

were training data. The other 100 were test data. As described in Section 4, each tweet was analyzed using MeCab. All uni-grams and bi-grams were used as features of Bag-of-Words. Training was performed using logistic regression. Accuracy obtained using the test data was 86.0%.

7 DISCUSSION

From the result, de-identification of two kinds apparently exists. For that reason, it is necessary to use the system properly according to the purpose of de-identification. For example, the classifier in Section 4 de-identifies tweets for sales purposes. Because it is not necessary to de-identify such tweets, tweets can be de-identified using both the classifiers in Sections 4 and 6. Here we propose a method to de-identify tweets related to a combination of words.

Method

Below is the algorithm of de-identification using the classifier constructed in Section 4.

Step 0: Substitute 1 for m, the number of morphemes to be deleted.

Step 1: Delete m morpheme(s) from an original tweet s_{org}. When the number of the morphemes of s_{org} is n, the number of possible patterns is ${}_nC_m$. Group the ${}_nC_m$ tweets into one group, S (= $\{s_1, ..., s_{{}_nC_m}\}$).

Step 2: For each tweet s_i in S, find the maximum value of the probabilities the classifier outputs for its location ($a_{1..200}$).

$$prob(s_{i,a_j}) = p(a_j \mid s_i)$$
$$maxprob(s_i) = \max(prob(s_{i,a_1}), ..., prob(s_{i,a_{200}}))$$

Step 3: Let the tweet with the least maxprob(s_i) be s_{new}, where the following holds.

$$s_{new} = \arg\min_{s_i} maxprob(s_i)$$

Return s_{new} if $maxprob(s_{new})$ is below the threshold (=0.37). Otherwise, increment m by 1 and back to Step 1 when m is less than n.

Results and Discussion

We present a part of the result of de-identification by the proposed method. The tweets below are samples of the de-identified tweets. Words to be deleted are crossed off.

(6) まだまだ 新幹線~~京都~~駅
(*It is still a long way to the ~~Kyoto Shinkansen~~ Station.*)

(7) 5 年ぶり 京都 ~~タワー~~
(*It has been five years since I came to Kyoto ~~tower~~.*)

(8) 清~~水~~の舞台から 1 枚 京都の街が一望だね
(*I took a picture from the top of ~~Kiyomizu~~. It has a full view of Kyoto.*)

(9) ランチ (at なか卯 河原町~~五条~~店) ~~折田~~先生なう
(*I am having lunch at the Nakau ~~Gojo~~ branch in Kawaramachi with ~~Orita~~-sensei now.*)

(10) 京都~~御所~~一般公開中
(*Kyoto ~~Imperial Palace~~ is now open to the public.*)

(11) 阪急~~河原町~~なう
(*I am at Hankyu ~~Kawaramachi~~ now.*)

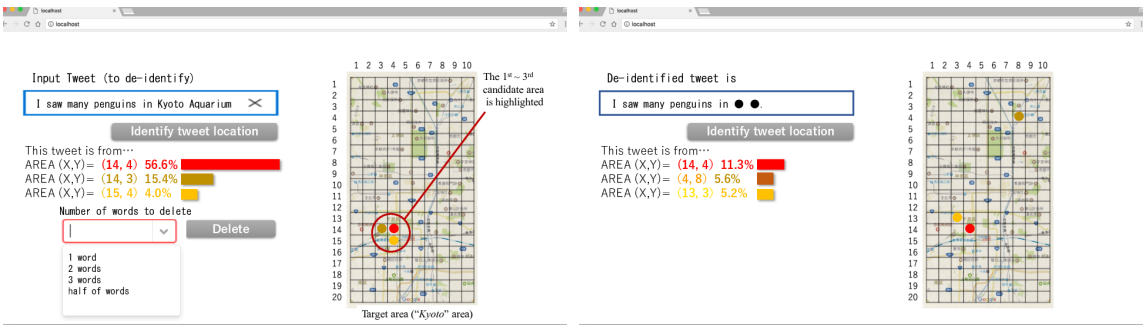
'新幹線京都駅 (*Shinkansen Kyoto Station*)' is a proper noun, but there are many stations in Kyoto City. Therefore ideal de-identification is achieved by deleting '新幹線京都 (*Shinkansen Kyoto*)'. In the case of '京都タワー (*Kyoto tower*)', an ideal de-identification system will delete 'タワー (*tower*)' because it is the only tower in Kyoto City. The result (5) is a successful example. Using the proposed method, ideal de-identification can be achieved in that this algorithm does not delete the whole proper noun.

With adequate training data, the method would work ideally, but a failure example exists as follows.

(12) 撮り飽きもせず 撮り足りもせず ~~京~~都御苑
(*I never get tired of and never get enough of taking pictures in ~~Kyoto~~ Gyoen.*)

The algorithm should delete '御苑 (*Gyoen*)', but it actually deletes '京都 (*Kyoto*)'.

Furthermore, we investigated the relation between the size of the training data and the accuracy. Figure 3 shows that more training data are necessary. Some difficulty arises



(a) Before: Original tweet and its location inference **(b) After: De-identified tweet and its location inference**

Figure 3: Images of the system's user interface

with obtaining sufficient amount of data because the way to make these data involves manual labeling.

8 APPLICATION

A system for inference and de-identification of tweets can be built based on the proposed de-identification method. Figure 3 presents screenshots for the system. Inputting any tweet, this system infers its tweet location and de-identifies it according to a selected number of morphemes to be deleted. This process supports both machine and human inference. For both (a) raw and (b) de-identified tweets, the location inference results are presented on each map.

9 CONCLUSION

This paper proposed a novel de-identification method to anonymize tweet locations. Two kinds of tweet location inference were presented. One is inference of a location itself. The other is inference of the feasibility of location inference. These location inferences are based on the respective definitions of de-identification. The former tends to regard tweets from stores as identifiable because such tweets are posted from only one place many times. The latter tends to regard tweets in which common nouns are used as proper nouns, as identifiable. Therefore, in practical use, it would not be sufficient to apply common concepts for de-identification of location. Our algorithm of de-identification based on the hypothesis that locations of tweets are inferable with combinations of words, partially brought expected results. Future tasks involve how to incorporate consideration of contexts. The analyses described in this paper investigated each tweet as a Bag-of-Words, and did not use information of relations of morphemes. This problem is expected to be resolved by consideration of the syntax structures of tweets.

ACKNOWLEDGEMENTS

This work is supported in part by Japan Agency for Medical Research and Development (16768699), Strategic Information and Communications R&D Promotion Programme (SCOPE), the Ministry of Internal Affairs and Communications of Japan.

REFERENCES

[1] L. Kong, Z. Liu, and Y. Huang. Spot: Locating social media users based on social network context. In *Proc. of the VLDB Endowment*, 7(13): pp.1681–1684, 2014.

[2] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proc. of the Fifth Intl. Conference on Web Search and Web Data Mining*, pp. 723–732, 2012.

[3] W. Hua, K. Zheng, and X. Zhou. Microblog entity linking with social temporal context. In *Proc. of the 2015 ACM SIGMOD Intl. Conference on Management of Data*, pp. 1761–1775, 2015.

[4] Y. Yamaguchi, T. Amagasa, H. Kitagawa, and Y. Ikawa. Online user location inference exploiting spatiotemporal correlations in social streams. In *Proc. the 23rd ACM Intl. Conference on Information and Knowledge Management*, pp. 1139–1148, 2014.

[5] M. Cha, Y. Gwon, and H. T. Kung. Twitter geolocation and regional classification via sparse coding. In *Proc. of the Ninth Intl. Conference on Web and Social Media*, pp. 582–585, 2015.

[6] D. Flatow, M. Naaman, K. E. Xie, Y. Volkvich, and Y. Kanza. On the accuracy of hyper-local geotagging of social media content. In *Proc. of the Eighth ACM Intl. Conference on Web Search and Data Mining*, pp. 127–136, 2015.

[7] S. Kinsella, V. Murdock, and N. O'Hare. I'm eating a sandwich in Glasgow: modeling locations with tweets. In *Proc. of the Workshop on Search and Mining User-Generated Contents*, pp. 61–68, 2011.

[8] G. Li, J. Hu, J. Feng, and K.-l. Tan. Effective location identification from microblogs. In *Proc. of the 30th Intl. Conference on Data Engineering*, pp. 880–891, 2014.

[9] H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos. Identification of key locations based on online social network activity. In *Proc. of the 2015 IEEE/ACM Intl. Conference on Advances in Social Networks Analysis and Mining*, pp. 218–225, 2015.

[10] M. Dredze, M. Osborne, and P. Kambadur. Geolocation for twitter: Timing matters. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1064–1069, 2016.

[11] Y. Fang and M. Chang. Entity linking on microblogs with spatial and temporal signals. *Transactions of the Association for Computational Linguistics*, 2: pp. 259–272, 2014.