Conditional density estimation for the origin of social media geolocation

Hayate ISO[†], Shoko WAKAMIYA[†], and Eiji ARAMAKI[†]

[†] Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan
E-mail: †{iso.hayate.id3,wakamiya,aramaki}@is.naist.jp

Abstract Despite the importance of the tweet geolocation, most of the research estimates the user geolocation through user timeline, not the tweet geolocation, because it is difficult to detect precise geographic coordinates from only a single tweet. In light of this, this study centers on estimating the single tweet geolocation through a density estimation approach. The advantage of density based model expresses the uncertainty of the tweet geolocation as the spread of the distribution. The proposed model reveals that a high consistency with human judgment, indicating the practical availability.

Key words Microblogs, Natural language processing, Geographic Identification, Density estimation, Machine Learning

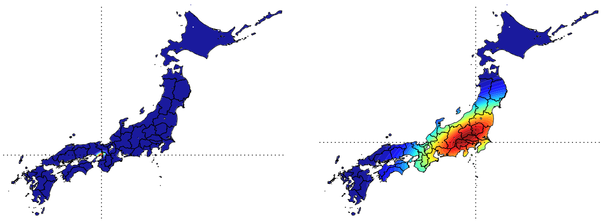
1. Introduction

The recent rise in popularity and scale of social media has created a growing necessity for social media-based applications. One of the strong advantages of social media applications is the availability of various meta information, such as temporal information [5], [7], spatial information [2], [9] and so on. Although temporal information is widely available, spatial information is often inaccessible for privacy reasons, causing a difficulty of precise location-based applications. In fact, a study reported that fewer than 0.5% of tweets include GPS information [3], thus motivating many location estimation studies [1], [4].

Nowadays, two types of location estimation tasks have been tackled; (1) to estimate the most probable user location from a user timeline (the *user level geolocation* task) and (2) to estimate the most probable geolocation for each tweet (the *message level geolocation* task). Although both of the tasks respectively have their own importance, most studies focus on the former task because of the task difficulty of the latter task.

Figure 1 shows the typical examples for the tweet and estimated geolocation. Although the former tweet is easy to identify the correct geolocation, the latter one is hard to that.

To quantitatively evaluate the difficulty of the message level geolocation, we have created a human annotated data that a worker labels geolocation for each tweet. The data reveals that only 10% of the tweets are annotated the geolocation by humans (*low recall*) (Table 1). In addition, 12% of annotated data were miss judged by humans. It is lower precision than our expectation. (*low precision*).



(a) *Nishinomiya is super hot.*
Error Distance: 7.7 (km)
Gold: Hyogo, Human: Hyogo

(b) *I'm really hungry.*
Error Distances: 77.4 (km)
Gold: Tokyo, Human: UNK

Figure 1: Estimated geographic distribution of two different tweets. "Gold" and "Human" represent the true location and the human estimation, respectively. The intersection of dotted line is the true geolocation.

According to this observation, this research employs a model to represent the certainty of the location as a mixture of probabilistic distributions. The density based approach tracts the geolocation uncertainly shown in Figure 1b. This density-based approach has another practical advantage; it enables the probabilistic values as follows:

- *posted at Kyoto prefecture with over 50%,*
- *posted within 50 km around Mount Fuji with over 20%.*

These probabilistic values are useful for various social media services, location dependent advertising optimization, and voting prediction.

In experiments, we evaluate our model under the hierarchical test dataset based on the human inferences. The experiments also showed a gap between human inference and actual locations.

Note that a recent study [8] was undertaken to estimate

Annotated ratio (%)		
Detailed (Prefecture)	Rough (Region)	Unknown
9.1	1.3	89.6
Human agreement (%)	Precision (%)	
	Prefecture	Region
93.6	89.7	74.6

Table 1: Human annotation summary.

the location as a density estimation problem. Although their motivations are similar to ours, this research independently considers the word and n-gram for estimating each Gaussian Mixture Model (GMM) and combines these GMMs using several optimization methods. We introduce a more reasonable method that automatically estimates the parameters for GMM, weights, means and covariances depending on the feature vectors, which are often used for the document classification.

2. Materials

2.1 Tweet dataset

We used a tweet dataset consisting of 554,320 geotagged tweets in Japan (July 15, 2012 July 21, 2012). From the data, we have removed the application generated tweet (Foursquare, etc.). Consequently, we obtained 154,748 tweets consisting of only official clients: Twitter for iPhone/Android. We, then, normalized tweets by replacing a username to *@mention* and web link to *-URL-*. We split tweets into three sets: training (144,748 tweets), validation (5,000 tweets), and test (5,000 tweets).

2.2 Human Annotation

To investigate the performance of human estimation, two humans annotated the location onto the test data (5,000 tweets). The human labels have a structure of three levels: "detailed" (Japanese prefecture-level; 47-way), "rough" (region-level; 8-way)[注1] and "unknown". To reduce the bias based on the humans' familiarity with areas, we permitted them to search for the location words. The annotated results are shown in Table 1.

3. Method

3.1 Word-specific Gaussian Mixture Model (GMM)

In the existing approach [8] used GMM for geographic density estimation, each tweet was converted to n-gram features consisting of the number of n-gram occurrences in a corpus and geographic coordinates (longitude and latitude) of n-grams. The GMM applied for each n-gram w_j is defined as word-specific GMM (WGMM) as

$$p(\mathbf{y}|w_j) = \sum_{k=1}^{K} \pi_{kj} \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{kj}, \boldsymbol{\Sigma}_{kj}),$$

where $\mathbf{y} \in \mathbb{R}^2$ represents geographic coordinates (latitude and longitude). w_j represents a word indexed in j, π_{kj} is the weight as the word w_j is assigned to k-th mixture component, $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{kj}, \boldsymbol{\Sigma}_{kj})$ is multivariate Gaussian distribution with mean $\boldsymbol{\mu}_{kj}$ and covariance matrix $\boldsymbol{\Sigma}_{kj}$. After estimating GMMs for each n-gram, the weighted sum of word-specific GMMs is combined as

$$p(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{J} \pi'_j p(\mathbf{y}|w_j),$$

where $\mathbf{x} = \{w_1, \cdots, w_J\}$ represents n-grams in a tweet, π'_j is a weight of GMM on the word w_j.

In the above formula, it is important how to define each GMM's weight parameters π'_j. Even if we independently consider the weight π'_j for each word, the output model highly depends on the locally distributed word in each tweet.

Whereas the WGMM combines the independently estimated group of models to represent tweet distribution, our approach, Gaussian Mixture Regression, is able to represent it through one model.

3.2 Gaussian Mixture Regression

This study applies *Gaussian Mixture Regression* (GMR) [10] to represent tweets distribution. This enables a uniform way of handling tweet features for estimating the conditional distribution $p(\mathbf{y}|\mathbf{x})$ conditioned on the tweet feature $\mathbf{x}$.

Our approach does not need to prepare a specific evaluation index for estimating weight π'_j. We can derive the parameters, weights of conditional distribution $p(\mathbf{y}|\mathbf{x})$, from jointly estimated GMM $p(\mathbf{x}, \mathbf{y})$.

To obtain the GMR results, we estimate the joint probability distribution $p(\mathbf{x}, \mathbf{y})$ of p-dimensional tweet features $\mathbf{x}$ and two-dimensional geographic coordinate $\mathbf{y}$. In this study, we use the Dirichlet processes to estimate GMM parameters. The conditional distribution $p(\mathbf{y}|\mathbf{x})$ can then be analytically derived as follows:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{p(\mathbf{x}, \mathbf{y})}{\int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}} = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} \\ &= \frac{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}, \mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{k,\mathbf{x}}, \boldsymbol{\Sigma}_{k,\mathbf{xx}})} \\ &= \sum_{k=1}^{K} \pi'_k(\mathbf{x}) \mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}_{k,\mathbf{y}|\mathbf{x}}, \boldsymbol{\Sigma}_{k,\mathbf{y}|\mathbf{x}}\right) \end{aligned}$$

where

[注1] : http://www.japan-guide.com/list/e1001.html

Method	Prefecture		Region		Overall	
	Mean (km)	Median (km)	Mean (km)	Median (km)	Mean (km)	Median (km)
Gaussian Mixture Regression	$\underline{251}$	$\underline{154}$	$\underline{242}$	$\underline{134}$	278	214
Elastic Net	272	181	268	181	$\underline{272}$	191
Mean location	277	185	273	185	273	$\underline{187}$

Table 2: Error Distances between model estimation and true geographic coordinates: These test datasets consist of hierarchical structures; Prefecture ⊂ Region ⊂ Overall. The lower value is the better estimation.

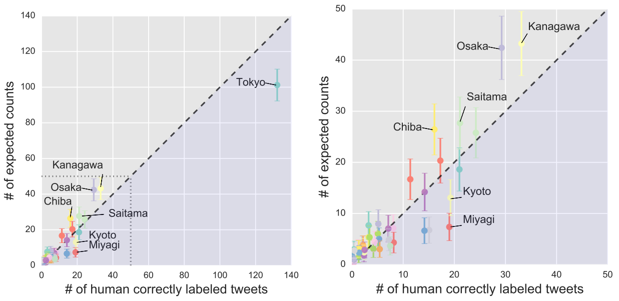


Figure 2: The count of each observed prefecture (x-axis) and expected ones (y-axis) in the human annotated data set. The right figure is an enlargement of the left one. The prefectures in blue shaded regions reflect that the amount of human inferable tweets is smaller than the overall geo-tagged tweets distribution and vice versa.

$$\boldsymbol{\mu}_k = \left[\begin{smallmatrix} \boldsymbol{\mu}_{k,\mathbf{x}} \\ \boldsymbol{\mu}_{k,\mathbf{y}} \end{smallmatrix}\right], \ \boldsymbol{\Sigma}_k = \left[\begin{smallmatrix} \boldsymbol{\Sigma}_{k,\mathbf{xx}} & \boldsymbol{\Sigma}_{k,\mathbf{xy}} \\ \boldsymbol{\Sigma}_{k,\mathbf{yx}} & \boldsymbol{\Sigma}_{k,\mathbf{yy}} \end{smallmatrix}\right],$$

$$\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{k,\mathbf{y}|\mathbf{x}}, \boldsymbol{\Sigma}_{k,\mathbf{y}|\mathbf{x}}) = \frac{\mathcal{N}(\mathbf{x}, \mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{k,\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{k},\mathbf{xx}})},$$

$$\boldsymbol{\mu}_{k,\mathbf{y}|\mathbf{x}} = \boldsymbol{\mu}_{k,\mathbf{y}} + \boldsymbol{\Sigma}_{k,\mathbf{yx}} \boldsymbol{\Sigma}_{k,\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{k,\mathbf{x}}),$$

$$\boldsymbol{\Sigma}_{k,\mathbf{y}|\mathbf{x}} = \boldsymbol{\Sigma}_{k,\mathbf{yy}} - \boldsymbol{\Sigma}_{k,\mathbf{yx}} \boldsymbol{\Sigma}_{k,\mathbf{xx}}^{-1} \boldsymbol{\Sigma}_{k,\mathbf{xy}},$$

$$\pi'_k(\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{k,\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{k},\mathbf{xx}})}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{k,\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{k},\mathbf{xx}})}$$

The key difference between WGMM and GMR is that the weights of mixture parameters $\pi'_k(\mathbf{x})$ are flexibly defined depending on the feature vector $\mathbf{x}$.

GMR is useful to represent the tweet as p-dimensional vectors. As described, we convert the tweet to vector representation learned by fasttext [6]. We set the dimension of the vector representation to 100, and compose a vector representation of a tweet by averaging all the word vectors in the tweet.

4. Results

4.1 Human Annotation Biases

We first investigate whether human classified tweets and overall geo-tagged ones are distributed in the same way. We also investigate which prefectures are easily or difficultly classified by humans. We implemented χ^2-test between human correctly classified data and overall geo-tagged tweet labels. For the overall geo-tagged data set, we employ training data. The P-value of χ^2-test is smaller than 0.001 that indicates the human inferable tweets differently distributed with the overall geo-tagged tweet.

As for exploring which prefecture influenced human decisions, we visualized the observed count in human correctly classified data and overall expected counts of geo-tagged tweets with standard deviation shown in Figure 2. When the observed count are larger than the expected value (blue shaded region), the number of human inferable tweets are larger than the overall geo-tagged tweets distribution and vice versa. The prefectures that belong to the blue shaded region tend to be the sightseeing spots. More users state in detail about the visiting places compared with the overall geo-tagged tweets. Conversely, the prefectures that belong to unshaded regions tend to be commuter towns. This means that many users state unrelated content with where that person is.

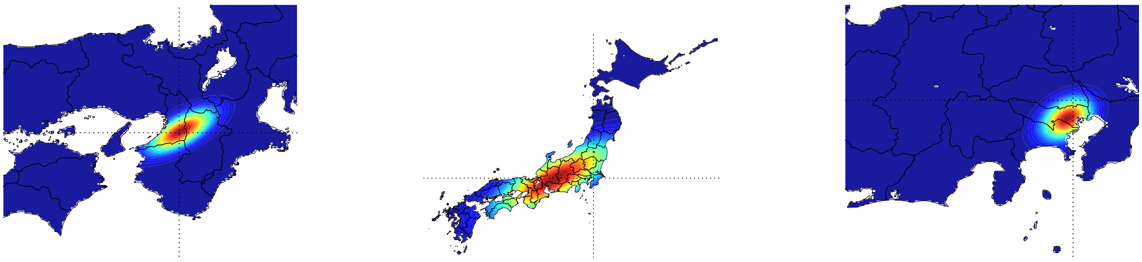
4.2 Evaluation

To compare our model with human inference, we evaluate prediction performance through several human inferable data set trained by the overall training data set. Note that our density-based model estimates only probability distribution, not the exact locations. Thus, we employ the mode value of estimated density as the estimated locations. To calculate the mode value of GMM, we need numerical optimization. As an approximation method, we employ the most probable mean in each mixture component.

We measure the mean and median values of error distances between the estimated location and the true tweet geolocation in the test data. As a baseline method, we use regularized linear regression method, Elastic-Net, using the same features and Mean locations. We optimize the baseline model using the validation set. Our results are presented in Table 2. Although our model performs worse than the baseline model for the overall test dataset, our model enhances prediction performance through human inference improvement.

Tweet	Gold Label	Human Estimation	Distance (km)	True label probability (%)
36 Kinki God of Fire pilgrimages.	Osaka	Osaka	3.2	42.6
Fuji can be seen before summer sunset!	Chiba	Kanto	250.7	1.06
Finish the work! Let's carouse all night at Ueno	Saitama	Tokyo	23.7	14.7

Table 3: Comparing the characteristic estimated density examples between human annotation and model inference



(a) Gold: Osaka, Human: Osaka

(b) Gold: Chiba, Human: Kanto

(c) Gold: Saitama, Human: Tokyo

Figure 3: Visualizing the characteristic estimated density examples. The color of the highest probability regions is red.

4. 3 Density-based Representation

We show some typical density representation of the tweet in Table 3 and Figure 3. As the first example in Figure 3a shows human label as coincident with the true tweet geolocation in the finest level, prefecture level. In such cases, our model estimates not only that the mode value of estimated density is near from the true tweet geolocation, but also that its variance is small.

In contrast to the first example, as for the second example in Figure 3b, humans are puzzled when specifying the location and labeled the wider level label, region level. Our model spreads to cover the human designated region.

The last example reveals the superiority of density-based estimation. The tweet's user stated about “Ueno”, which is a part of Tokyo. Thus, the human estimated the label as Tokyo for this tweet. However, the tweet's user was at the neighboring of Tokyo; the true tweet geolocation is actually within Saitama prefecture. Although a conventional classification approach hardly captures the proximity between the true tweet geolocation and the estimated area, our density-based method easily grasps it.

4. 4 Discussion

We have shown that the density-based approach is consistent with human estimation. Whereas conventional methods aren't interfered the performance which part of test data is used to predict, our density-based approach dramatically improves the prediction performance. In contrast to the conventional approach, each mixture component contributes a different role for modeling the data. For instance, noisy tweets belong to the broadly scattered mixture component, and similar tweets in terms of both content and geographic coordinates belong to the condensed one in the joint distribution $p(\mathbf{x}, \mathbf{y})$. Therefore, our model properly expresses the consistent estimation with human inferred geolocation.

5. Conclusion

In this study, we demonstrated that Gaussian Mixture Regression providing a new perspective for estimating tweets' geolocations. We discovered that the density-based approach exceeded the conventional approach in the several aspects; (1) the proximity of estimated area even for the miss classified tweet, and (2) the improvement of the prediction performance along the improvement of the human inference.

References

[1] Ajao, O., Hong, J., and Liu, W. A survey of location inference techniques on twitter. *Journal of Information Science 41*, 6 (Dec. 2015), 855–864.

[2] Broniatowski, D. A., Paul, M. J., and Dredze, M. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE 8*, 12 (12 2013).

[3] Cheng, Z., Caverlee, J., and Lee, K. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proc. of CIKM 2010*.

[4] Han, B., Cook, P., and Baldwin, T. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research 49* (2014), 451–500.

[5] Iso, H., Wakamiya, S., and Aramaki, E. Forecasting word model: Twitter-based influenza surveillance and prediction. In *Proc. of COLING 2016* (2016), pp. 76–86.

[6] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

[7] Paul, M. J., Dredze, M., and Broniatowski, D. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*

(2014).

[8] Priedhorsky, R., Culotta, A., and Valle, S. Y. D. Inferring the origin locations of tweets with quantitative confidence. In *Proc. of CSCW 2014*.

[9] Sadilek, A., Kautz, H. A., and Silenzio, V. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI* (2012).

[10] Sung, H. G. *Gaussian mixture regression and classification*. PhD thesis, Rice University, 2004.