Japanese Place Name Disambiguation based on Automatically Generated Training Data

Seiji Okajima and Tomoya Iwakura

Fujitsu Laboratories Ltd.
1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki 211-8588, Japan
{okajima.seiji, iwakura.tomoya}@jp.fujitsu.com

Abstract. This paper proposes a Japanese place disambiguation method by estimating Japanese prefectures referred by text. For filtering out irrelevant candidate places given by a gazetteer, our method first estimates Japanese prefectures referred by a given text with a classifier created from automatically generated training data. We can efficiently filter out ambiguous Japanese places with estimated Japanese prefectures because there are almost no cities that have the same name in the same prefectures in Japan. We evaluate our method with manually labeled Japanese tweet data. The experimental results show that our method attains higher accuracy than a method for selecting the set of places that attains minimum area in the candidate places of a given text.

Keywords: place name disambiguation, machine learning, wikipedia

1 Introduction

The amount of texts on the Web, such as news article, blog, micro blog, and so on, are increasing. These texts include valuable information for knowing opinions and events. One of the examples is use of Twitter, which is a micro blog service, for knowing events such as earthquakes [16] and predicting flu epidemic [1, 3, 10]. For exploiting event information extracted from text, the place that each event happens is often required.

One of the widely used methods for identifying places referred by text is use of a gazetteer. However, there are ambiguities that the same place name exists in multiple regions. For example, a place name "Dublin" is located in "Ireland", "United States" and "Australia". These phenomena are not limited to English.

This paper proposes a Japanese place name disambiguation method based on the content of each text. Our method assumes a certain level of address hierarchy, such as Japanese prefecture, is useful for disambiguating ambiguous places. In other words, for the Japanese case, we assume landmark names and address have no ambiguity in a prefecture. Based on the assumption, our method uses a classifier that predicts the Japanese prefectures referred by a given text. With the classifier, we filter out irrelevant Japanese places given by a gazetteer with the estimated Japanese prefectures because there are almost no cities that have the same name in the same prefectures in Japan.

There are two contributions of our proposed disambiguation method.

- We propose a training method of a classifier that estimates Japanese prefectures with automatically generated data. This method reduces the cost of preparing training data.
- We experimentally demonstrate combination of a gazetteer-based candidate generation and an automatically trained filter attains higher accuracy than a gazetteer-based one and a disambiguation with collocated place names.

2 Related Work

Most previous researches for disambiguating places rely on heuristics using metadata from a gazetteer. The population of the candidate places are often employed by selecting candidate places [15, 2, 14]. These methods select the place that has the largest population in candidates.

A spatial relationship between the places in each text [15, 11, 18] has also been used. The set of candidates that has the minimum area consisted of the candidates. Hierarchical information of the places like London and United Kingdom are also used [15, 2]. In this method, the hierarchical relation of the other places with the non-ambiguous place in a given text is used. However, if there is no non-ambiguous place, the method does not work.

Another method is use of context information, like words related to each region, obtained with machine learning algorithms for estimating regions. In order to prepare training data for training a classifier based on contextual information, geocoded posts of micro blog are used [5, 4, 8]. Geocoded posts are assigned to geodesic grids and we can train classifiers with the posts, however, it is still difficult to obtain training data that cover all the target areas.

Methods like Entity linking or Wikification [12, 7, 13] also disambiguates place ambiguities by using context information. However, we have to prepare the context information of all places in gazetteer for linking/wikifying entities to corresponding articles, and we cannot apply these systems easily.

Therefore, in this paper, we proposed an automatic generation method of training data for estimating prefectures in Japan, which is administrative areas, as regions. In addition, we use the hierarchy of Japanese prefectures with the characteristic that Japanese address have no ambiguity in a prefecture.

3 Proposed Method

Figure 1 shows the procedure of our disambiguation system. At first, the system recognizes named entities (NEs) that are used as place name candidates from an input text. Next, metadata of candidate places are assigned to the extracted NEs by a gazetteer. The metadata includes address and geographical coordinates (latitude and longitude). If there are candidates given by the gazetteer, the system estimates Japanese prefectures referred by the text. Finally, the places are identified by filtering out the irrelevant candidates that do not belong to the estimated prefectures.

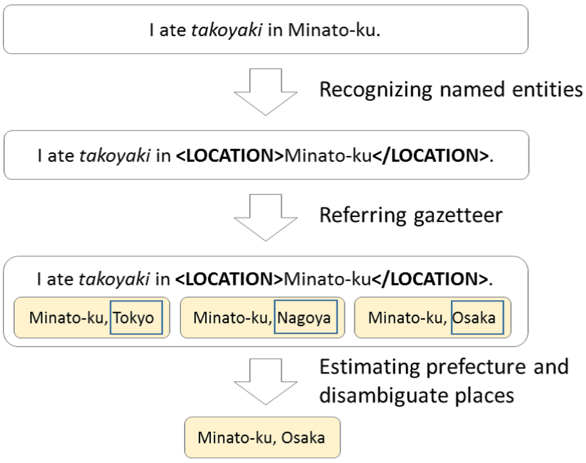
I ate *takoyaki* in Minato-ku.
Recognizing named entities
I ate *takoyaki* in <LOCATION>Minato-ku</LOCATION>.
Referring gazetteer
I ate *takoyaki* in <LOCATION>Minato-ku</LOCATION>.
Minato-ku, Tokyo
Minato-ku, Nagoya
Minato-ku, Osaka
Estimating prefecture and disambiguate places
Minato-ku, Osaka

Fig. 1. An overview of our disambiguation method.

3.1 Recognizing Place Names in Text

The first step is recognition of candidate places to be matched with a gazetteer. In order to recognize candidate places, we use an NE recognizer. We employ an NE recognition method [9] to recognize place names. The NER method extracts the eight types of NEs defined by IREX [1] and we use LOCATION and ORGANIZATION as place entities.

3.2 Gazetteer

We build a gazetteer consists of the address and its geographical coordinates based on following data sources.

- Location Reference Information Download Service (LRIDS) [2]
- Infoboxes of Wikipedia

LRIDS is published by *Ministry of Land, Infrastructure, Transport and Tourism*. LRIDS includes city/district level addresses with their position coordinates and street level addresses with their position coordinates. We used both of them for building our gazetteer.

Moreover, infoboxes of Wikipedia are used as another data sources. The keys of a gazetteer are derived from the combination of *prefecture*, *county*, *city*, *ward*, *street* and title of Wikipedia pages. The keys are normalized such as idiomatic orthographic variants, Arabic/Chinese numerals and one/double byte characters.

We built the gazetteer by using *LRIDS* of the 2015 October 18 edition and Wikipedia of the 2016 April 7 edition. In total, the gazetteer that have 1,679,853 records were created.

[1] http://nlp.cs.nyu.edu/irex/
[2] http://nlftp.mlit.go.jp/isj/

3.3 Automatic Generation of Training Data

We apply a machine learning algorithm to train a classifier to disambiguate place names. However, the cost of preparing training data is a problem. Therefore, we propose an automatic generation method of training data for estimating prefectures of Japan. To generate the training data, we use Wikipedia and our gazetteer.

Training Data from Wikipedia We use the abstracts of Wikipedia pages to create training data. Each abstract of Wikipedia is composed with the title, its URL, its abstract text, etc., and it is provided in XML format.

In order to generate training data from Wikipedia, we extract the pages including prefecture names as the instance of prefectures then nouns and compound nouns in each abstract are used as feature.

We don't use the pages of person as training data generation. Most person pages have hometown information that includes prefecture names but such pages isn't strongly related to the places. Person pages are detected by checking whether abstract of each page includes about 80 different words that indicate a person such as *poet*, *writer* and *cartoonist*, etc. or not. The pages judged as not related to prefecture and person pages are treated as the instance of out-of-prefectures class.

We built 92,296 instances of the Japanese 47 prefectures and 496,803 of out-of-prefectures class as training data from Wikipedia.

Training Data from Gazetteer We consider that training data of Wikipedia is not enough to estimate prefectures because Wikipedia only has pages of famous places. Therefore, we also create the training data from a gazetteer. We use the each records as instances and features are extracted from key, county, city, ward, street.

We built 780,655 instances of the 47 prefectures as training data from a gazetteer.

3.4 Training

We train a classifier with the automatic generated training data from Wikipedia and a gazetteer. As a machine learning, we use AROW [6] and we obtained three types of 48-class classifier. The detail of each classifier is described in the next section.

4 Experiments

4.1 Dataset

To evaluate accuracy of our method, we create the annotated data with Japanese addresses. We use 20,000 tweets from December 2012 to March 2013. The tweets were collected for avoiding the bias of the number of prefectures by populations of prefectures.

For example, when the tweet ”*There was a fire in Shibuya-ku*” is given, an annotator extracts ”*Shibuya-ku*” as a place name and assign the address ”Shibuya-ku (city), Tokyo-to (prefecture)” like this:

```
There was a fire in
 <place ad="Shibuya-ku, Tokyo-to">Shibuya-ku</place>
```

In this annotation process, we annotated 16,918 named entities with Japanese addresses in the 20,000 tweets.

4.2 Baseline Method

The procedure of the baseline methods is the same as proposed method until the gazetteer is applied to NEs but filtering methods are different. In this experiment, we employ following two methods as baseline filters.

Random:
A filter randomly selects a place from the candidates.

Convex hull:
A filter selects the set of places in candidate places of a text under the constraint to minimize a convex hull calculated based on a set of places [11].

4.3 Experimental Results

Evaluation of Prefecture Classification First, we evaluate the accuracy of prefecture classifier with the evaluation data set. We divide 20,000 labeled tweets into 5,000 and 15,000 tweets. The 5,000 tweets are used for evaluation of prefecture classification. Remaining the 15,000 tweets are used for evaluation of place name estimation and this results are described in sub-subsection 4.3. Macro precision, recall and F-measure of each prefecture are employed as evaluation metrics. In this experiment we examine three prefecture classifiers.

– PC1 is trained with Wikipedia only.
– PC2 is trained with mixture of Wikipedia and gazetteer.
– PC3 is trained with classifiers of Wikipedia and gazetteer combined by a stacking manner [17]. We first trained a gazetteer-based classifier. Then, the Wikipedia-based classifier is trained with the gazetteer-based classifier.

The experimental results of three classifiers are shown in Table 1. We see from Table 1 that PC3 is the best accuracy. We also see from the results of PC2 and PC3 that training data from gazetteer contribute to improved recall. However the precision of classifier PC2 is greatly reduced by increasing the size of the gazetteer. We think that this is because there is no contextual information in the gazetteer-based training data. In contrast, PC3 maintains precision while improving recall. This indicates a stacking training contributes to improved accuracy when we use training data including contextual information like the Wikipedia-based one and no contextual information like the gazetteer-based one.

In the following, we use the classifier PC3 in the experiments of the place estimation.

Table 1. Results of prefecture classification.

Classifier	Precision	Recall	F-measure
PC1	**0.94**	0.69	0.77
PC2	0.82	**0.74**	0.77
PC3	0.93	0.72	**0.80**

Evaluation of Place Name Estimation We evaluate the estimation accuracy of the places of addresses, prefecture, city, ward and street, by the baseline methods and the proposed method. We use 15,000 tweets as the test data except for the 5,000 tweets that was used in the prefecture classification experiment. Macro precision, recall and F-measure of estimating detailed addresses of each prefecture are also employed as evaluation metrics. The estimation results of each place are evaluated by following guidelines.

true-positive The place is extracted from the tweet and the estimation result of its address is correct.
false-positive The place is extracted from the tweet but the estimation result of its address is incorrect or unannotated place is extracted from the tweet.
false-negative The place is not extracted from the tweet.

Table 2 shows the evaluation results of baseline and proposed methods. Our proposed method achieved high accuracy. Especially, precision and F-measure are significantly improved compared with the baseline methods. However, recall of Convex hull is higher than that of proposed method. This is because Convex hull method always identify a disambiguated place from place name candidates but our proposed method sometimes reject all candidates. Our method shows 9.8 points and 8.4 points higher F-measure values than those of Random and Convex hull. This result indicates that filtering of the candidate places with estimated prefecture contribute to improved accuracy.

Robustness In order to evaluate the robustness against the gazetteer size, we measured transition of precision value when changing the size of gazetteer. In this evaluation, we use the 15,000 tweets same as place estimation experiments. Each size of gazetteer are created by random sampling from the prepared gazetteer

Table 2. Results of Place Name Disambiguation.

Method	Precision	Recall	F-measure
Random	0.604	0.827	0.695
Convex hull	0.603	**0.872**	0.709
Proposed	**0.772**	0.819	**0.793**

and the size of gazetteer is changed in increments of 50,000, from 100,000 to 1650,000. All gazetteer are sampled 10 times and the average precision and recall of the experiments are plotted in figure 2. Average f-measure is also plotted in figure 3. The results show that precision of the baseline methods are greatly reduced when the size of gazetteer is larger than 1,000,000. On the other hand, reduction of precision of proposed method is gradual and our method maintains a sufficient precision even at 1,000,000 scale of gazetteer. Therefore, F-measure of our proposed method is improved despite the size of gazetteer is increased over 1,000,000.

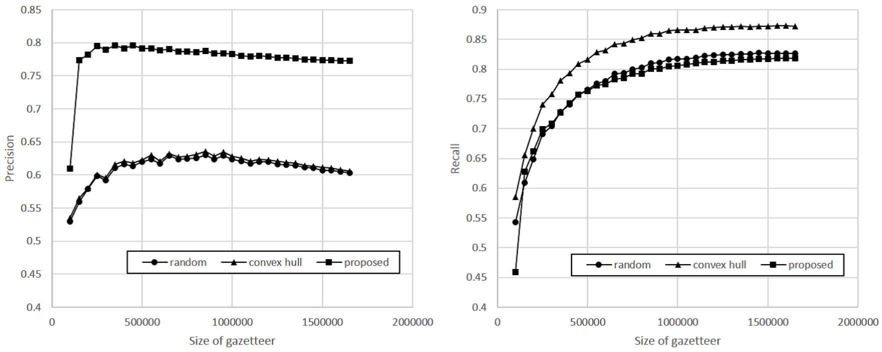

Fig. 2. Transition of precision (left) and recall (right) according to the gazetteer size.

Error Analysis We consider place name estimation errors can be classified into three types; (a) NE recognition error, (b) Prefecture classification error and (c) Lack of gazetteer data. We sampled 232 of false-positive and false-negative cases from experimental results and specify the reason of errors. Table 3 shows the investigation result of place name estimation errors. The result indicates that most errors are caused by lack of gazetteer. Especially, our method often cannot be estimate landmarks (such as store or buildings). We can solve this problem by increasing gazetteer data. The accuracy of prefecture classifier is not bad, however there are still room for improvement.

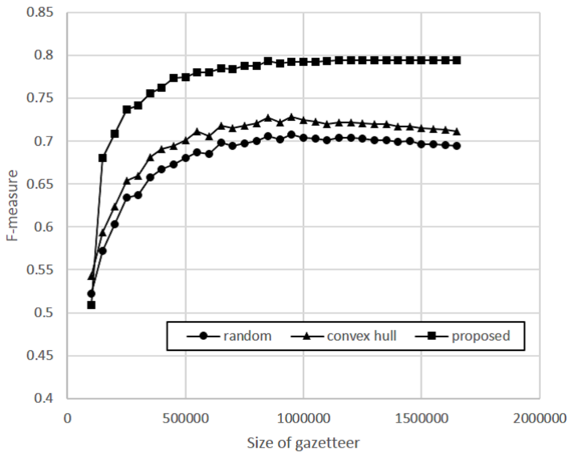

Fig. 3. Transition of f-measure according to the gazetteer size.

Table 3. Number of reasons of place name estimation errors.

(a) NER	(b) Classifier	(c) Gazetteer
10(5%)	41(17%)	181(78%)

4.4 Discussion

Figure 4 is F-measure values of place estimation about each prefectures. This result indicates proposed method improved accuracy of most prefectures compared with the baseline methods. However, the accuracy of some prefectures such as ”Tokyo” and ”Iwate” are degraded. This seems to be that correct candidates are filtered out by prefecture classifier. We will be able to improve the problem by enhancing the automatic generation method of training data. For example, there may be a way to generate training data by using full text of Wikipedia while currently we use only abstract or we can use not only noun and noun phrase but also the order of words as feature of prefecture classifier.

5 Conclusion

In this paper, we proposed a Japanese place disambiguation method by using classifier that is trained with automatically generated training data. The experimental results showed that our method attains higher accuracy than conventional methods. We also examine the robustness against the size of gazetteer and find out that proposed method keep enough precision even if increasing gazetteer size, in contrast conventional methods do not.

As future works, we have to add more landmark data to our gazetteer and improve the accuracy of the prefecture classifier because the accuracy of our

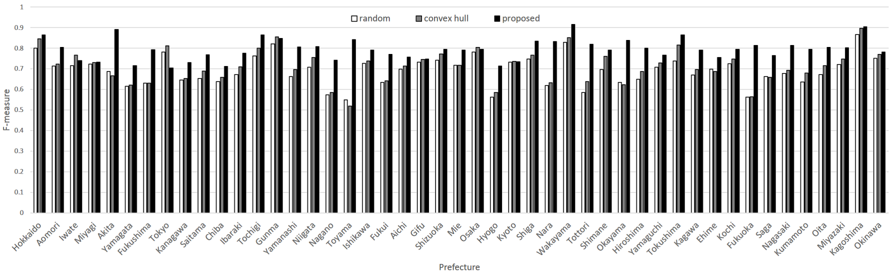

Fig. 4. F-measure results of place estimation about each prefectures.

proposed method is heavily relied on these components. We also would like to re-evaluate our method by comparing more strong conventional method. Moreover, we have to consider applying our method to the other languages and the places of other countries that have administrative area system like Japan.

References

1. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S., Liu, B.: Predicting flu trends using twitter data. In: IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2011. pp. 702–707 (2011)
2. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 273–280 (2004)
3. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: Detecting influenza epidemics using twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 1568–1576 (2011)
4. Chandra, S., Khan, L., Muhaya, F.B.: Estimating twitter user location using social interactions-a content based approach. In: PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT). pp. 838–843 (2011)
5. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM Conference on Information and Knowledge Management. pp. 759–768 (2010)
6. Crammer, K., Kulesza, A., Dredze, M.: Adaptive regularization of weight vectors. Machine Learning 91(2), 155–187 (2013)
7. Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data. In: EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 708–716 (2007)
8. Ikawa, Y., Enoki, M., Tatsubori, M.: Location inference using microblog messages. In: Proceedings of the 21st World Wide Web Conference. pp. 687–690 (2012)
9. Iwakura, T.: A named entity recognition method using rules acquired from unlabeled data. In: Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria. pp. 170–177 (2011)

10. Kitagawa, Y., Komachi, M., Aramaki, E., Okazaki, N., Ishikawa, H.: Disease event detection based on deep modality analysis. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. pp. 28–34 (2015)
11. Leidner, J.L.: Toponym resolution in text: annotation, evaluation and applications of spatial grounding. SIGIR Forum 41(2), 124–126 (2007)
12. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management. pp. 233–242 (2007)
13. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. TACL 2, 231–244 (2014)
14. Pouliquen, B., Steinberger, R., Ignat, C., Groeve, T.D.: Geographical information recognition and visualization in texts written in various languages. In: Proceedings of the 2004 ACM Symposium on Applied Computing (SAC). pp. 1051–1058 (2004)
15. Rauch, E., Bukatin, M., Baker, K.: A confidence-based framework for disambiguating geographic terms. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1. pp. 50–54. HLT-NAACL-GEOREF '03 (2003)
16. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. pp. 851–860 (2010)
17. Wolpert, D.H.: Stacked generalization. Neural Networks 5(2), 241–259 (1992)
18. Zhang, Q., Jin, P., Lin, S., Yue, L.: Extracting focused locations for web pages. In: Web-Age Information Management - WAIM 2011 International Workshops: WGIM 2011, XMLDM 2011, SNA 2011, Wuhan, China, September 14-16, 2011, Revised Selected Papers. pp. 76–89 (2011)