Identifying Bloggers' Residential Areas

Norihito Yasuda, Tsutomu Hirao, Jun Suzuki, and Hideki Isozaki
NTT Communication Science Laboratories, NTT Corp.
2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, JAPAN, 619-0237
{n-yasuda,hirao,jun,isozaki}@cslab.kecl.ntt.co.jp

Abstract

This paper proposes a method to infer bloggers' residential areas. Identifying bloggers' residential areas will be useful as another axis to retrieve weblogs or for tasks that resolve ambiguous objects in terms of geographic contexts. Our method focuses on the local context of geographic location terms and uses binary classifiers to decide whether the context is indicating the writer's residential areas. Experimental results show that the method correctly estimated the residential prefecture out of 47 prefectures in Japan at a 50.7% accuracy rate for bloggers who wrote a geographic location term at least once.

Introduction

In this paper, we propose identifying the residential area of bloggers or writers of weblogs. Identifying residential areas is a task to identify the fixed unit of area such as the state or prefecture where the blogger is living. Recently an increasing number of weblogs provide us with an enormous amount of language resources. However, retrieving information from weblogs is not well developed yet. Introducing living information can be one criteria for information retrieval from the chaotic blogsphere, meaning that such queries such as "dentist, ache *written by someone living in Kyoto*" would be acceptable with residential area identification. In addition to such straightforward usage, residential area identification will also be useful for indirect uses in other applications. One characteristic of weblog texts is that fairly amount of them contain description about personal events or offer opinions. Blood (Blood 2002) distinguished three main uses of blogs: filters, personal journals, and notebooks. In their sample, Herring et al. (Herring *et al.* 2004) reported that 70% of blogs were typed as personal journals, which can be seen as "on-line diaries". Blogs containing personal events or opinions are important targets of sentimental or subjectivity analyses, which have recently attracted much attention (e.g., (Pang & Lee 2004; Turney 2002)).

One interesting usage of such personal opinions is retrieving subjective information from blogs based on geographic criteria. Opinions or events in limited local areas, which will never be covered by the mass media are useful for people living in that area. For example, descriptions such as feelings about a local dentist or a sale at a local clothing store, will not be included in a newspaper article, but are still useful for local people.

In addition to personal opinions, we think the location information of blogs is also interesting for general information retrieval geographic queries, so-called geographic information retrieval (Larson 1996; McCurley 2001).

To realize the applications described above, it is important to relate the terms in the text to real world entities, including both toponym (Baker 2003; Smith & Mann 2003; Leidner 2004) or anaphora resolution. Newspaper articles tend to write geographic expressions in strictly unambiguous forms, which we cannot expect from weblogs. So naturally we take into consideration that weblogs are written as online diaries. Moreover, we can find many examples more difficult than toponym disambiguation. Consider the following example:
"Today I went to the newly opened library, and I felt that the design was simple, functional, and elegant."
'The library' in this case must be a specific library for the writer, but it is not explicitly written in the text. Assuming that no antecedents hints exist in the text, identifying this location is virtually impossible unless we have knowledge external to the text. A blogger's residential area can serve as a last resort hint for such difficult tasks including exophora resolution. Since a blogger must live in a specific location and his/her active field is usually fixed, we can limit the candidate of the referent to such an area.

When considering the above applications, we think that precise inferences are more important than wide coverage for two reasons. First, since so many people are writing blogs, even a small portion of them still contains a hefty mass of data, which is sufficient for most applications. Second, in the above applications, residential area identification works as a backend. Less precise inferring causes leveraged errors.

Maybe the simplest inferring is assumes that geographic location terms appearing in the blog indicate the blogger's residential area. However, not all geographic location terms indicate residential areas; for example, some may write about sightseeing trips to other areas, and others may offer political analysis about an accident that occurred elsewhere. Such blog entries will contain many geographic location terms, but they do not imply the blogger's home base.

Even if many geographic location names are included, humans tend to be able to infer residential areas through the text. Consider the following examples:

1. I went to Osaka Station by bicycle to see the latest commuter train model.
2. I went to Osaka Station for the first time. It was pretty big and has many platforms.
3. I went to Osaka Station yesterday.

Assume that we know that Osaka Station is in Osaka prefecture. By reading carefully, we notice that the writer of sentence 1) probably lives in Osaka prefecture because he/she can get to Osaka Station by bicycle. We can also infer that the writer of sentence 2) probably does not live in Osaka prefecture because usually residents in Osaka prefecture have been to Osaka Station at least one time. From sentence 3), however, we may not infer because people get on trains at Osaka Station for commuting or for sight-seeing. The above examples suggest that the local context of a geographic location term (Osaka Station in the above examples) can be used to determine whether the term is really indicating the writer's residential area. Therefore, we propose to focus on their local context instead of treating geographic location terms in broader contexts.

One characteristic of weblogs is that they are continuously written by the same individual. We do not have to identify only from a one-shot example because we can use many examples continuously written by the target blogger. To exploit local contexts and this characteristic, our method consists of three steps. The first step is dictionary-based candidate extraction, which extracts any occurrence of geographic location terms from weblog texts. In the second step, classify all candidates who appear in positive contexts that indicate the blogger's residential area or negative ones using a binary classifier. The third step is voting on the classifier's outputs. Since the classifier will be trained without geographic location terms, we do not expect retrain it if a new area or a new location term is appended.

Corpus

For our study, we collected blog entries from 'goo blog' (http://blog.goo.ne.jp/), one of the biggest blog service sites in Japan, and most of the users are Japanese. As is common with many blog service sites, it offers some tools to make publishing blogs easier, including a 'profile tool' that enables bloggers to easily make their profiles public. Hence published web pages using this tool have special boundaries that indicate which part of the page is output from the profile tool from which we can automatically extract user profiles. We selected this site for our corpus because the residential areas provided by this tool are limited to a predefined set of candidates, so we can avoid the ambiguity of location names freely filled by each user. The set is basically one of the 47 prefectures in Japan. Filtering is another reason we selected this site; in blog service sites that do not filter the contents, a vast amount of computer-generated adult blogs are posted periodically. On the other hand, this blog site apparently manually filters salacious contents including computer-generated pornographic sites, resulting in collected data that is predominantely written by humans.

Location name	Prefecture
...	...
Shibuya	Tokyo
Odaiba	Tokyo
Minato-Mirai	Kanagawa
Kan-nai	Kanagawa
Kita-Shinchi	Osaka
...	...

Table 1: Sample entries of location dictionary

The total number of collected blog entries was 5,278,107, and the total number of bloggers was 74,155. The dates of the collected blogs ranged from 1 Jan 2005 to 30 Sep 2005. By parsing profiles produced by the profile tool, 40,354 (54.4%) of the prefectures of the bloggers were made public.

We extracted text bodies, excluding images, HTML tags, dates, headlines, and other decorations. The resulting data size was about 3 Gigabytes with 2 bytes for each character. We segmented the extracted texts into words using the Chasen morphological analyzer (Matsumoto *et al.* 2003).

Figure 1 shows the distribution of the bloggers' residential prefectures. Since the distribution is relatively skewed, trivial guesses that always submit 'Tokyo' can achieve a certain level of accuracy.

Location Dictionaries

Since the residential areas obtained from the collected corpora are prefectures, we selected prefecture as the area unit of these location dictionaries. The dictionary hence consists of only two columns; a location term and its prefecture name. A sample entry of the location dictionary is shown in Table 1.

To see how the location dictionary affects our methods, we prepared three kinds of dictionaries according to size.

Small-Sized Dictionary

The small-sized dictionary consists of very well-known area names. To get them, we used a Japanese web-

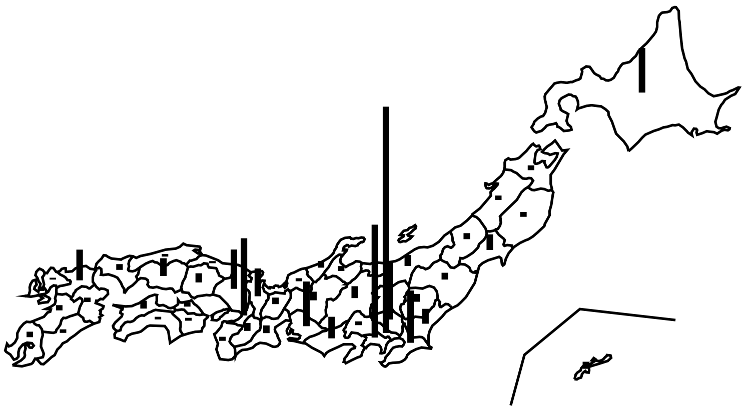

Figure 1: Residential prefecture distribution in our collected corpora

based travel guide site: http://machi.goo.ne.jp/. This site's directory is organized into three levels: zones, prefectures, and major areas in the prefecture. There are seven climes, 47 prefectures and, 147 area names. Entries for prefectures ranged from 1 to 21. For example, Tokyo (the most populous prefecture) has 21 entries, whereas less populous prefectures such as Toyama and Wakayama have only one entry, typically a station name.

We collected all 147 major names and used them as entries for a small-sized location dictionary.

Medium-Sized Dictionary

As with the small-sized dictionary, we also collected medium-sized dictionaries from a Japanese web-based travel information site: http://www.gojapan.jp/. This site, which provides various types of travel information including public facilities, is organized by prefecture and category of interest. There are 33 categories of interest from which we selected 15 that contain landmarks, facilities and special events in the area.

Since some terms obtained here were polysemic, we manually omitted terms that might be recognized as human names location name situated in another prefecture. Entries per prefecture ranged from 59 to 665. By combining the entries and the small-sized directory, eventually the vocabulary size became 7,531.

Large-Sized Dictionary

A large-sized dictionary was made from Japanese zip code lists, containing 121,161 area names . Each name consists of three levels of geographical terms: prefecture name, city or town name, and local area name.

	ambiguous/unambiguous
local area only	49694 / 64545
prefecture + local area	19296 / 94943
city + local area	6 / 114323

Table 2: Ambiguous terms in zip code lists

Although any area can be identified if these three components appear in a series, there are names that can be uniquely identified without prefecture or city names. Almost all city or town names (second level components) are unambiguous. Therefore, we also added names that are unambiguous themselves. Furthermore, since most of the local area names can be resolved by adding a prefecture name, we also added such names concatenated with prefecture names to the dictionary. In the same way, if the local area name can be uniquely identified by combining seen names with city or town names, we also added such combinations. Table 2 shows the number of ambiguous terms if concatenated with prefecture or city/town names. Note that about 60% of ambiguous toponyms listed in the Japanese zip code lists can be resolved if the prefecture is known.

We also combined the above entries with our small and medium-sized dictionaries. The resulting vocabulary size was 244,897.

Identifying Residential Areas

Our residential area location method consists of three steps. First, find any occurrence of the location terms and extract sentences that contain them. Second, classify whether each sentence indicates the blogger's residential area using binary classifier. Third, conduct vote using all classifier results for the blogger.

Identifying residential areas is basically a multiclass classification problem. However, generally speaking, classification is difficult if the number of classes is too large. However, the number of geographic locations is typically large. For example, even though prefectures are the largest administrative unit in Japan, the country has 47. By combining the location dictionary with a binary classifier, we can avoid treating this problem as a multiclass problem. Furthermore, since we trained the classifier without geographic location terms, the classifier is supposedly dependent on area or location terms, which means it should work when on unknown location term is input.

The content of this dictionary does not have to be an exhaustive list of geographic terms; instead, we believe it can be relatively small. The aim of this method is not locating objects that appeared in the blog and making exhaustive lists of geographic information, but just locating the residential area of the person who wrote the blog. Therefore eligible entries for this dictionary are ones that tend to represent a person's location.

An overview of the method's procedure is shown in Figure 2.

Identifying candidate location terms

The first step in our method is identifying all occurrences that potentially indicate a blogger's residential area. A series of blog entries written by the target blogger is supposed to server as input. Since our method focus on the local context of geographic location term, a relatively small unit, a sentence, will used as the context.

For all occurrences of location terms that are entries in the location dictionary, sentences that contain the location terms are extracted, and the candidate prefecture is marked by referring to location dictionary.

As described in the previous section, dictionary size basically does not matter, since the location dictionary need not be an exhaustive list; a small-sized dictionary is also applicable, even though it may sacrifice coverage. On the other hand, the quality of the dictionary does matter, since the relationship between a location term and its prefecture name is a key piece of information upon which the rest of our process relies; ambiguous or unreliable entries are not desirable.

To avoid uncertain relationships between location terms and its pefecture names, we augmented the dictionary for morphological analysis to avoid segmenting location terms in texts; besides a standard dictionary for morphological analysis, we added terms that appeared in location term dictionaries.

Binary classifier

The binary classifier estimates that the local context of a geographic location term indicates the residential area of the writer. We use the AdaBoost (Freund & Schapire 1996) algorithm with a bag-of-words represented vector. To avoid retraining the classifier when a new area or a new location term appears, we train the classifier without geographic location terms. The classifier is supposed to work with unknown location terms. Eventually, the bag-of-words representation of the words that surround the location term will be classified.

Location Term Weight

Although the classifier works independently of target location terms, each location term itself has tendencies that suggest whether it appears in positive or negative contexts. For example, a relatively small station mainly used by local people will tend to appear in texts written by residents of that area. On the other hand, a famous temple known as a sightseeing spot will tend to appear in texts written by tourists (non-residents).

Thus, we also use the location term weight combined with the output of the binary classifier. We determined the weight of term t as follows:

$$w(t) = \log \tfrac{P(t)+1}{N(t)+1},$$

where $P(t)$ denotes the number of times that the term t appeared in positive contexts that indicate the blogger's residential area, $N(t)$ denotes the number of times that the term t appeard in negative contexts.

For location terms that do not appear in the training set, we use 0 as $w(t)$.

Voting

One characteristic of weblogs is that they are continuously written by the same person. Thus, generally, we can obtain candidates that indicate location terms. To get more precise inferences by using plural outputs from a binary classifier, we conduct vote on these outputs and select an area that got the highest score as the blogger's residential area.

We prepared two kinds of voting schemata. The first uses the signs of the classifier outputs, which means that positive output from the classifier counts as +1 and negative as -1. The second selects the raw score from the classifier. In this step, location term weight is also considered. Thus we defined the voting score for residential area a in each schema as follows:

- voting schema 1

$$V(a) = \sum_{loc} ((\text{classifier sign for}\, s(loc)) + \alpha\, w(loc))$$

- voting schema 2

$$V(a) = \sum_{loc} ((\text{classifier score for}\, s(loc)) + \beta\, w(loc))$$

,

List any occurrences of location term:

(last night's game with the [Fukuoka] Hawks was terrible, ...)
(today I went [Shibuya] for a job interview for my next part-time ...)
([Namba] in Osaka prefecture is famous for ...)

Lookup location dictionary and add location weight for each vector

Tokyo	-0.18	(last night's game with the [LOCATION] Hawks was terrible,)
Tokyo	0.12	(today I went [LOCATION] for a job interview for my next part-time)
Osaka	0.33	([LOCATION] in Osaka prefecture is famous for ...)

Consult the binary classifier and get results with scores:

Tokyo	P (0.23)	-0.18
Tokyo	P (0.38)	0.12
Osaka	N (-0.24)	0.33

Voting

1) Tokyo: 0.55, 2) Osaka: 0.08

Figure 2: Procedure overview

dictionary	appeared sentences	appeared bloggers
small-size	172,647	19,663
medium-size	200,460	21,752
large-size	214,170	23,645

Table 3: Extracted sentences and their writers

where *loc* denotes that the geographic location term appeared in the target blogger's text, $s(loc)$ denotes the sentence in which the term *loc* appeared, $w(loc)$ denotes the lacation term weight of *loc*, and α and β are constants.

The area that gets the highest positive score will be selected as the final output of our method. If no candidate gets a positive score, 'reject' will be output. Since our method is dependent on location dictionary entries, if all occurrences of location terms are unrelated to the blogger's residential area, our method cannot produce correct answers; such cases are possible. We think that rejection is better than outputting answers likely to be incorrect.

Experiments

We conducted experiments on the collected corpora. First we only identified candidate location terms to see how dictionary size affects matching. Table 3 shows the number of extracted sentences and the number of different bloggers. Contrary to our expectations, the number of sentences were close to each other, while the size of the small and large dictionaries differed about 1,600 times. Hence we only used the medium-sized in the following experiments with 21,752 bloggers who wrote geographic location terms listed in the dictionary at least one time.

method	accuracy
always guessing 'Tokyo'	26.8%
dictionary-based voting	48.2%
proposed, voting 1	50.7%
proposed, voting 2	50.1%
proposed, voting 1, w/rejection	57.6%
proposed, voting 2, w/rejection	56.4%

Table 4: binary classifier performance

We compared with two naive methods. The first naive inference method uses the characteristics of bloggers' distribution by prefecture. As shown in Figure 1, the distribution is skewed, always guessing 'Tokyo' can get a certain level of accuracy.

The second compared method is simple dictionary-based voting. For each identified candidate location term, prefecture names corresponding to the location term will be counted, and the prefecture that acquired the most counts is selected.

As mentioned in the previous section, our method essentially does not work when none of the candidates are positive unless all other 46 prefectures are negated. Actually, for about 37% of these 21,752 bloggers, no candidate location terms were found that correctly indicated the blogger's prefecture. In such situations, rejection might be better than giving an desperate answer. We also measured performance assuming correct rejections when all of the candidate location term appeared in negative contexts.

Table 4 shows the results of experiments. The proposed method achieved slightly better results than the compared methods.

	Top 12 Positive Features	Top 12 Negative Features
2	*Jitensha* (bicycle)	*Onsen* (spa)
3	*Minami-Guchi* (south gate)	*Tokushima* (prefecture name)
4	*Awa-Odori* (festival name)	*Hoteru* (hotel)
5	*Futan* (baggage or load)	*Sen* (fight or game)
6	*Hanabi* (fireworks)	*Ryoko* (travel)
7	*Shinema* (cinema)	*Mito* (city name)
8	*Nishi-Guchi* (west gate)	*Aomori* (prefecture name)
9	*Ba* (bar)	*Gumma* (prefecture name)
10	(alphabet character F)	*Kyoto* (prefecture name)
11	*Chome* (... street or ... quarter)	*Miyage* (souvenir)
12	*Ensen* (along a railway of ...)	*Ken* (prefecture)

Table 5: Highly weighted features

Discussion

Although peformance improvement is not so significant, the trained weighted features suggest some interesting things. Table 5 shows examples of highly weighted positive and negative features. In positive features, we can see such places to which we usually walk. Note that 'F' (10th in the list of positive features) maybe appeared because it is used to indicate floor number in the buildings. In negative features, prefecture names frequently appear, perhaps because when we mention unfamiliar places we tend to write more specificlly, while for familiar places we tend to write shorter. In other words, the features listed on the postive side tend to specify small districts, while bigger districts are listed on the negative side.

Conclusion

We proposed a method to estimate bloggers' residential areas, which will be useful as an another axis to retrieve weblogs or tasks that resolve ambiguous objects in terms of geographic contexts.

Since geographic location terms which does not relate to the blogger frequently appear, we introduced a binary classifier that estimates whether the local context surrounding the location term is indicating the blogger's residential area.

Although experimental results were not so significant, weighted features suggests us some interesting characteristics appeared in the local context of the location term.

Currently, blogs that does not contain any location dictionary entries cannot be handled by our method, and it causes less coverage. As future work, to compensate for the coverage, we plan to combine the proposed method with text categorzation-based methods, which do not rely on location term dictionaries.

References

Baker, E. R. M. B. K. 2003. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References.*

Blood, R. 2002. *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog.* Perseus Publishing.

Freund, Y., and Schapire, R. E. 1996. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 148–156.

Herring, S. C.; Scheidt, L. A.; Bonus, S.; and Wright, E. 2004. Bridging the gap: A genre analysis of weblogs. In *HICSS*.

Larson, R. R. 1996. Geographic information retrieval and spatial browsing. http://sherlock.berkeley.edu/geo ir/PART1.html.

Leidner, J. L. 2004. Toponym resolution in text: "which sheffield ist it?". In *Proceedings of the the 27th Annual International ACM SIGIR Conference (SIGIR 2004).* Abstract, Doctoral Consortium.

Matsumoto, Y.; Kitauchi, A.; Yamashita, T.; Hirano, Y.; Matsuda, H.; Takaoka, K.; and Asahara, M. 2003. Japanese morphological analysis system chasen version 2.3.3.

McCurley, K. S. 2001. Geospatial mapping and navigation of the web. In *World Wide Web*, 221–229.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 271–278.

Smith, D., and Mann, G. S. 2003. Bootstrapping toponym classifiers.

Turney, P. D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *CoRR* cs.LG/0212032.