Spatial Statistics of Term Co-occurrences for Location Prediction of Tweets

Ozer Ozdikis (✉), Heri Ramampiaro, and Kjetil Nørvåg

Norwegian University of Science and Technology, Trondheim, Norway
{ozer.ozdikis,heri,noervaag}@ntnu.no

Abstract. Predicting the locations of non-geotagged tweets is an active research area in geographical information retrieval. In this work, we propose a method to detect term co-occurrences in tweets that exhibit spatial clustering or dispersion tendency with significant deviation from the underlying single-term patterns, and use these co-occurrences to extend the feature space in probabilistic language models. We observe that using term pairs that spatially attract or repel each other yields significant increase in the accuracy of predicted locations. The method we propose relies purely on statistical approaches and spatial point patterns without using external data sources or gazetteers. Evaluations conducted on a large set of multilingual tweets indicate higher accuracy than the existing state-of-the-art methods.

Keywords: Location prediction, tweet localization, spatial point patterns, feature extraction

1 Introduction

Explicit location information in terms of latitude-longitude associated with text messages and photos in social networks provides a valuable resource for a wide range of applications, such as event detection, targeted advertisement, and crisis management. One of the most popular of these social networking platforms is Twitter, which enables users to post 140-character tweets and share them with their followers. Its widespread adoption and the accessibility of tweets through public APIs make it an attractive resource for research. However, despite increasing availability of GPS-enabled mobile devices, geotagged tweets are reported to constitute only 1-3% percent of all tweets [1, 2]. As a result, predicting tweet location from its text has recently received considerable attention [1–7].

A widely adopted content-based approach for tweet localization is probabilistic language models. In this approach, the area of interest is partitioned into subregions, and terms in tweets that are posted in these regions are used for the training of text-based classifiers [3]. Specialized feature selection methods that prioritize geo-indicative terms have also been proposed in order to increase the prediction accuracy of these classifiers [5, 8, 9].

The hypothesis that we investigate in this work is that even if strong location-indicative terms are perfectly identified in tweets, other terms can still be important in the interpretation of spatial information. In other words, if each term in a tweet is considered independent from other terms, probability assignments may give misleading results. The method that we propose in this work explores and evaluates spatial relationships, namely *attraction* and *repulsion*, between co-occurring terms in tweets using spatial point patterns and statistical methods. Selected term pairs (bigrams) with clustering or dispersion tendency with respect to the underlying unigram distributions are included in feature space to improve the accuracy of prediction.

To explain our idea, consider an example where we want to predict the location of a tweet mentioning *heathrow* with high precision, e.g., within a tolerance of 1 km error distance. The term *heathrow* can be considered to provide strong evidence about the location of a tweet, probably supporting the region around the Heathrow Airport in London, which covers a relatively large area. In this example, if that tweet also mentions *terminal*, whose co-occurrence with *heathrow* has stronger clustering tendency than *heathrow* alone, evaluating these two terms together as a new feature can yield predictions closer to the actual location. On the other hand, the phrase *heathrow express* can have an opposite effect (i.e., dispersion) and repel the geographical focus of the tweet to a region away from the airport area. We find such repulsion patterns quite interesting since even if they do not point to a specific place, they can indicate where a region is less likely to be the actual location for a tweet. In this example, the tweet mentioning *heathrow express* is probably posted from somewhere in the city, referring to the train that rides to the airport. Our claim is that such co-occurrences in a tweet can make an attraction or dispersion influence that may affect the geographical interpretation of a single term.

The main contributions of our work can be summarized as follows: 1) we investigate the spatial attraction and repulsion patterns of term co-occurrences, and propose a method to extend the feature space with term pairs having significant clustering or dispersion tendency, 2) we develop statistical techniques that can detect relationships between various types of features including emojis and multilingual texts, 3) we integrate our method with other unigram feature selection techniques to obtain higher prediction accuracies. An important aspect of our approach is that we can achieve the improvement in location prediction using only the tweet text in our analyses, i.e., we do not rely on external data sources, gazetteers, or other tweet metadata.

The remainder of this paper is organized as follows: We present a summary of related work in Sect. 2. We describe our proposed method in Sect. 3, along with a summary of baseline classification and feature selection techniques. Section 4 is devoted to our experiments and evaluation results. Finally, in Sect. 5, we conclude the paper and discuss future research directions.

2 Related Work

Location prediction for tweets can be described as estimating the geographical origin where a tweet is posted from [4, 6, 10]. Various techniques from the areas of information retrieval, machine learning, and natural language processing have been proposed to make accurate predictions [2–4, 7, 11–14]. One of the widely adopted techniques to that aim is probabilistic language models. In this technique, probability distributions are assigned for different subregions in an area using the textual content of georeferenced tweets in a training set [5, 9, 15]. Based on this trained model, per-region probabilities are then determined for non-geotagged tweets to be localized. A significant advantage of content-based approaches is that they can make predictions even in the absence of any other geographical cues [8].

Recent efforts to improve the accuracy of content-based approaches employ feature selection techniques, most of which have previously been used in similar text categorization problems [16]. The objective of these improvements is to determine location indicative terms in tweets by ranking them according to a metric. Top-n ranked features are then used in the training of language models, rather than using the complete vocabulary. Among recent studies in that direction, Cheng et. al [8] determine local words according to an analysis of frequency and dispersion. In [5], the authors experimented with numerous feature selection methods, such as information gain, information gain ratio, χ^2 statistic, geographical spreading, and Ripley's K statistic, and showed that information gain ratio outperforms their benchmark prediction methods in terms of accuracy. In that work, the authors use unigrams and also note that their preliminary results with named entities and higher order n-grams were not satisfactory. In a similar study [9], the authors employed Kernel Density Estimation (KDE) and Ripley's K statistics in order to improve the performance of location estimation for Flickr photos, particularly when only few terms can be selected for prediction. Their experiments revealed that the optimal results using geographical spreading was approximately the same as the optimal results based on KDE and Ripley K. However, geographical spreading showed more sensitivity to the number of features used in prediction.

The main objective in these previous feature selection efforts is to select location-indicative terms and eliminate common words that presumably have no spatial dimension [5]. Our approach essentially differs from these studies by evaluating spatial interactions between term pairs, even if a term appears to have no explicit spatial dimension. The method we adopt in our solution uses Ripley's K function [17], which has widespread usage in characterizing the spatial distribution patterns of objects in two-dimensional space [9, 18]. To the best of our knowledge, our work is the first to analyze spatial patterns of term co-occurrences with respect to the underlying term distributions, and use them in the location prediction of tweets.

3 Spatial Co-occurrence Patterns in Location Prediction

In this section, we briefly describe our baseline model, and then explain the details of our location prediction method. Adhering to the probabilistic language model, the region of interest is discretized into mutually exclusive subregions and a Multinomial Naive Bayes (NB) classifier with additive smoothing is trained using terms (unigrams) in tweets in a training set [5,9]. We use Multinomial NB classifier mainly because it incorporates class priors in prediction and is reported to perform well even on scarce training data [5]. We adopt a grid-based approach to define subregions, since we aim to make fine-grained predictions, such as at the level of a place in a city [2,15].

Improvements over this classifier apply feature selection and use only the selected location-indicative terms for training. These methods were categorized as statistical, information-theoretic, and heuristic in [5], and we implemented different methods from each category as our baselines (explained in Sect. 4). Our proposed method can also use the results of these term selection methods, and identify spatially significant bigrams according to the selected unigrams. In the remainder of the section, we explain how we detect spatially significant bigrams and use them in the enhancement of feature space for location prediction.

3.1 Detection of Significant Spatial Co-occurrence Patterns

Ripley's K-function, represented by K_λ, is a statistical method to evaluate spatial patterns of points in a region [9,17,18]. The function calculates a value that is proportional to the number of point pairs that lie within a distance of λ to each other. In practice, it is widely applied to analyze spatial patterns of a set of objects having a certain property in order to determine whether these objects have a clustering or separation tendency.

We use Ripley's K-function to analyze the geographical distribution of specific terms (and term pairs) in tweets based on the latitude-longitude coordinates of these tweets. The K-function is defined as:

$$K_\lambda(X_t) = A \times \frac{|\{(x_i, x_j)|x_i, x_j \in X_t, x_i \neq x_j, d(x_i, x_j) < \lambda\}|}{|X_t|^2} \tag{1}$$

where X_t=$\{x_1, ..., x_m\}$ with m=$|X_t|$ represents the set of tweets that include the term t, A represents the area of our grid, and $d(x_i, x_j)$ is the distance between two tweets x_i and x_j according to their coordinates. The value of $K_\lambda(X_t)$ is proportional to the number of tweet pairs in X_t that are within a distance of λ to each other. The λ parameter enables the evaluation of spatial relationships at different distance scales.

In an environment where the underlying population distribution is non-homogeneous, the value of the K-function for a specific set of objects may be affected by the population distribution. Therefore, comparison with the underlying point pattern should also be performed in order to evaluate the clustering

Algorithm 1 Find co-occurrences with attraction/repulsion w.r.t. feature space

```
Input1: Set of terms in feature space and set of all terms in the training corpus
Input2: Distance range $\lambda$ for Ripley's K-function
Input3: Number of Monte Carlo simulations to execute, denoted by $M$
Output: Set of bigrams B=$\{\langle t_p, t_c\rangle \mid X_{t_p t_c}$ has either clustering or repulsion tendency with respect to $X_{t_p}$ $\}$
for each primary term $t_p$ in feature space do
    Find the set of tweets $X_{t_p}$ that include $t_p$
    for each distinct term $t_c$ in the corpus do
        Find the set of tweets $X_{t_p t_c}$ for which $t_p$ is followed by $t_c$ in the tweet text
        Apply K-function in Eq. (1) on $X_{t_p t_c}$ to get $K_\lambda(X_{t_p t_c})$
        for $i$=1...$M$ do
            Randomly sample $n$ tweets from $X_{t_p}$, where $n$=$|X_{t_p t_c}|$
            Let $X^i_{t_p}$ denote this sample, apply K-function on $X^i_{t_p}$ to find $K_\lambda(X^i_{t_p})$
        end for
        Calculate upper boundary ($u$) and lower boundary ($l$) of envelop using $K_\lambda(X^i_{t_p})$ values with 0.05 confidence interval
        if $K_\lambda(X_{t_p t_c}) > u$ or $K_\lambda(X_{t_p t_c}) < l$ then
            Insert tuple $\langle t_p, t_c\rangle$ to the set of selected bigrams $B$
        end if
    end for
end for
```

and dispersion tendency of objects with respect to the population. This is usually achieved by executing a stochastic process, namely the Monte Carlo simulation [9]. The simulation mainly consists of taking random samples from the population, applying K-function on the samples, and calculating a confidence envelope with upper and lower bounds. A point pattern with K_λ value above the upper bound indicates clustering tendency (attraction), whereas the K_λ values below the lower bound is interpreted as dispersion (repulsion).

Our proposed method employs a similar approach to analyze the spatial patterns of term co-occurrences. However, rather than using the whole tweet set in the corpus as the underlying distribution, we compare the spatial distribution of co-occurring term pairs (bigrams) with the spatial patterns of corresponding single terms (unigrams). In other words, we measure the clustering and dispersion tendency of a bigram with respect to the spatial point pattern of each term in the bigram. This can be considered as a conditional analysis of the bigram's spatial distribution. Our algorithm to find term co-occurrences having significant attraction or repulsion pattern with respect to their unigrams is presented in Algorithm 1. For each unigram in the feature space, which we call *primary term* and denote by t_p, the algorithm finds co-occurring terms t_c in the training corpus that follows a primary term and exerts an attraction or repulsion influence on t_p. Specifically, if the spatial pattern of tweets with bigram $t_p t_c$ has significantly higher K_λ value compared to K_λ values of the tweet samples with t_p alone, $\langle t_p, t_c\rangle$ is regarded to have a clustering tendency in relation to t_p. Similarly, if the $K_\lambda(t_p t_c)$ value is below the lower boundary, $\langle t_p, t_c\rangle$ is selected as a repulsion co-occurrence for t_p. If spatial patterns of tweets that include $t_p t_c$ and t_p have no significant divergence, we do not perform any further analysis on $t_p t_c$. The reason we make a separate definition of primary term is to enable using the aforementioned feature selection methods for unigrams (e.g., χ^2, in-

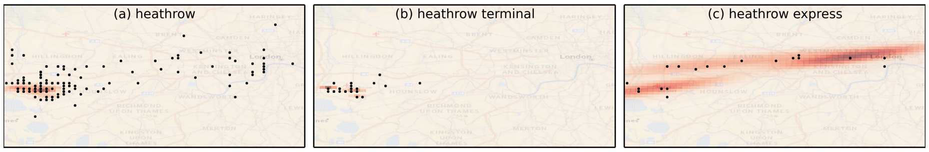

Fig. 1. Tweets mentioning (a) heathrow, (b) heathrow terminal, (c) heathrow express

formation gain). That means, a primary term t_p is taken from the feature space, which may be a selected subset of all distinct terms in the corpus, whereas t_c can simply be any term in the corpus.

We present the example in Fig. 1 to explain our findings. The dots in the figure represent locations of tweets in our grid for London area, and shadings in red color are generated by KDE for visualization purposes. Figure 1(a) shows the locations of tweets mentioning *heathrow* in our data set. Although they are slightly scattered, these tweets still exhibit a high concentration around the Heathrow Airport, as expected. Figure 1(b) presents a subset of these tweets, specifically the ones that include the bigram *heathrow terminal*. The density distribution of these tweets noticeably focuses on a more specific region in Heathrow. On the other hand, Fig. 1(c), which is generated for the bigram *heathrow express*, depicts a remarkably different distribution. This is probably due to the fact that people can post tweets about a train line that goes to the airport from places distant to the airport. As a result, although *heathrow* is a strongly descriptive term for location, we observe that its co-occurrences with other terms can change the spatial interpretation remarkably. Our experiments revealed that we can distinguish such attraction and repulsion patterns by applying Algorithm 1.

As noted in [9], when comparing two K-function values, the number of data points that are used in these calculations can affect the results. More specifically, $K_\lambda(X_1)$ and $K_\lambda(X_2)$ would not be comparable if $|X_1|$ and $|X_2|$ were different, since a larger dataset is more likely to yield a higher K_λ value. We do not observe this issue in Algorithm 1, since each simulation takes exactly $n{=}|X_{t_pt_c}|$ samples from X_{t_p}, as described in line 11. This is enabled by $X_{t_pt_c}$ being a subset of X_{t_p}. This also provides an advantage in terms of computational cost. In fact, except for a few term pairs that co-occur very frequently in our corpus (e.g., *United Kingdom*), we observe that $|X_{t_pt_c}|$ is remarkably lower than $|X_{t_p}|$ in most cases, which resulted in acceptable computation times in our experiments. Moreover, we transform latitude-longitude coordinates of tweets into three-dimensional Euclidean coordinates and index them in a k-d tree [9]. This transformation provided us noticeable performance improvement in the calculation of K_λ values.

The steps in Algorithm 1 describe the analysis of bigrams in the form of t_pt_c (i.e., t_p is followed by t_c). Similar procedures are also executed to identify spatially related bigrams where a primary term t_p is preceded by t_c. Ordering of terms in bigrams should be taken into consideration since we examine the distribution of a bigram conditioned on the distribution of t_p. We explain the

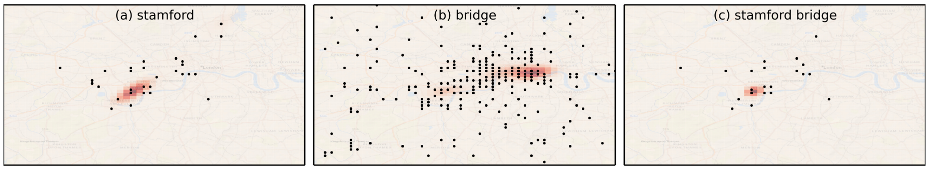

Fig. 2. Distribution of tweets mentioning (a) stamford, (b) bridge, (c) stamford bridge

effect of ordering on an example. Figure 2(a) and 2(b) demonstrate the distributions of *stamford* and *bridge*, respectively. Although tweets mentioning *bridge* are scattered in a large area, it exhibits a strong clustering tendency when preceded by the term *stamford* in tweets. In other words, when *stamford* and *bridge* are considered as t_c and t_p, respectively, distribution of $t_c t_p$ given in Fig. 2(c) leads to a significant clustering tendency with respect to the distribution of t_p. The density of tweets in (c) actually points to the region around the stadium Stamford Bridge. Similar relationships are also detected between emojis and terms, such as *heathrow* and ✈, since we primarily use statistical methods in our analyses without applying any restriction on the content of a tweet. The next section explains how we use these detected bigrams in the enhancement of our feature space.

3.2 Enhancement of Feature Space

The enhancement of feature space is performed by adding bigrams from B, which were found by Algorithm 1 above, as additional features to tweets. Specifically, given a tweet x with n terms in its text, denoted by $[t^x_1, t^x_2, ...t^x_n]$, if a bigram $\langle t^x_i, t^x_{i+1} \rangle$ exists in B (i.e., having significant spatial relationship), that bigram is added to the tweet as a new feature. Our rationale in this operation is that, if there is a clustering or dispersion tendency of a bigram $t_i t_j$ with respect to t_i or t_j, we can make more reliable estimations if we also have $t_i t_j$ in the tweet.

We exemplify the expansion operation on a hypothetical tweet with terms $[a,b,c,d]$. Assume that the tuple $\langle a, b \rangle$ exists in B, i.e., a and b were found to have significant spatial relationship in Algorithm 1. In this case, applying expansion on this example tweet results in $[a,b,c,d,\langle a, b \rangle]$. Following this example, if B had also included the pair $\langle b, c \rangle$, that bigram would also be added to produce $[a,b,c,d,\langle a, b \rangle,\langle b, c \rangle]$. This means that we do not impose any restriction about mutual exclusion, and utilize a bigram if it has a spatially significant pattern with respect to its unigrams.

The complete process can be summarized as follows: Using a training tweet set, we obtain B using Algorithm 1, enhance the feature space by the expansion operation, and train the language model for classification. For a new tweet to be localized, we apply the expansion operation for it according to the trained model and estimate its location using a Multinomial NB classifier.

4 Evaluation

In this section, we present the evaluation results of our method and compare with state-of-the-art baselines. We evaluate our methods on regional datasets to predict tweet locations at fine-granular level. Accordingly, we selected the Greater London Area for our experiments and divided the area into equal-size grid cells to form a 100x100 grid. This resulted in cells covering approximately an area of 0.25 km^2 each. We collected public tweets between 3 October 2015 and 20 January 2016 using the Twitter Streaming API[1]. We filtered out tweets without explicit GPS coordinates (i.e., latitude-longitude) and obtained 4,040,775 geotagged tweets posted from our area of interest. Following the common practices in earlier similar studies, we eliminated exact duplicate tweets, Foursquare check-ins, and tweets from possible spammers [5, 8, 13]. To filter out spammers, we excluded tweets from users with more than 1000 friends or followers or who posted more than 300 tweets in our time window (approximately more than two tweets per day), and tweets with advertisement hashtags (e.g., #job, #realestate), since they have almost the same text and are usually posted from same places [8,13,19]. Finally, we obtained 489,466 unique tweets posted by 100,997 distinct users. We did not apply any restriction on the language of a tweet.

Each tweet in our dataset is assigned to a grid cell according to its GPS coordinates. In our experiments, we used randomly chosen 464,993 tweets for training and the remaining 24,473 for test. Tweet texts are divided into tokens by using Twokenize[2] library. For training data, we discard tokens that appear in less than five tweets, hyperlinks, and single characters to reduce sparsity [3, 13]. This yields a total number of 54,752 distinct tokens (unigrams) in our training set. Tokenized tweets in the training set and their assigned grid cells are used in building the probabilistic language models and for analyzing the spatial point patterns of bigrams.

4.1 Evaluation Methodology

We compared our proposed method with different baselines, including the full model (i.e., using all terms in tweets without making prior unigram feature selection) and four feature selection methods. Among a wide range of feature selection techniques, we implemented the following four as our baselines, since they are widely applied in state of the art:

1. IG: Information Gain (*information-theoretic*) [5]
2. IGR: Information Gain Ratio (*information-theoretic*) [5]
3. CHI: $\mathcal{X}^2$ statistic (*statistical*) [16]
4. GS: Geographical Spreading (*heuristic*) [9]

In our implementations, we followed the descriptions given in the cited papers above. In addition to these five baselines, we also made estimations using class

[1] https://dev.twitter.com/overview/api

[2] https://github.com/brendano/ark-tweet-nlp/

priors [5], which basically finds the grid cell with maximum number of tweets in the training set. This is used to show that assigning all test tweets simply to the most populous place does not yield useful results.

We evaluate the performance of these methods using the following three metrics: 1) *Accuracy:* proportion of tweets in the test set for which the true grid cell is correctly predicted, 2) *Accuracy@n:* proportion of tweets for which the estimated location is at most n kilometers away from the true location, 3) *Median:* median of the distances between the predicted location and the true location for test tweets. The distances in these metrics are calculated based on the centers of predicted and true grid cells for tweets.

For each feature selection method in our baselines, we first apply a ranking of tokens in the training set using the corresponding feature ranking metric. For example, IG ranks all tokens in the training set according to their information gain. Then, the training of a baseline predictor is performed by using top-n features in its ranking, and the location prediction for test tweets are executed with that setting.

We apply our proposed enhancement on each baseline separately. When applied on the full model, Algorithm 1 analyzes all terms in the corpus. When applied on a unigram selection method, the algorithm uses only the top-n unigrams as primary terms (explained in Sect. 3.1). We denote our enhanced methods with suffix *SCoP*, as an abbreviation for *Spatial Co-occurrence Pattern.* For example, $\text{IG}_{(n)}$+SCoP represents our enhancement where the top n of unigrams with the highest information gain are used as the feature space in Algorithm 1.

4.2 Evaluation Results and Discussion

Table 1 presents the minimum median error distances obtained by each method, along with the corresponding accuracies. Since each method can achieve its highest accuracy using different top-n features, we also indicate this value with a subscript. For example, $\text{IGR}_{(0.6)}$, the most accurate baseline, uses top 60% of unigrams with highest information gain ratio. $\text{IG}_{(1.0)}$ means that a feature selection based on information gain does not perform better than the full model for any selection of top-n. We observe that the evaluation results of our baselines are also consistent with the findings in previous studies [5, 9].

The table shows that our enhancement (denoted by *SCoP*) on each baseline results in better predictions, even when no prior unigram feature selection is applied (full model). The most accurate predictions are obtained by our enhancement when it is applied on $\text{IGR}_{(0.6)}$. In order to analyze the difference in error rates between a baseline and its SCoP enhancement, we employ McNemar's test on their predictions. The results of the test show that the improvement using SCoP is statistically significant for every baseline ($p \ll 0.00001$). In our experiments, the λ distance that we used in the calculation of Ripley-K values is 0.5 (in kilometers). We have also experimented with λ=2.0 and obtained similar results. Specifically, when λ=2.0 is used, $\text{IGR}_{(0.6)}$+SCoP predictions were made with a median error distance of 0.7430 and an accuracy of 0.432. Since λ=0.5 performed slightly better, we demonstrate the results that we found using λ=0.5.

Table 1. Comparison of methods with settings minimizing their median error distance

Prediction Method	**Median (km)**	**Accuracy**	**Acc@0.5km**	**Acc@1.0km**	**Acc@2.0km**
Class Prior	3.7743	0.049	0.065	0.119	0.310
Full Model	1.4860	0.363	0.391	0.442	0.530
Full Model+SCoP	1.2585	0.408	0.432	0.478	0.558
$\mathrm{IG}_{(1.0)}$	1.4860	0.363	0.391	0.442	0.530
$IG_{(1.0)}+SCoP$	1.2585	0.408	0.432	0.478	0.558
$\mathrm{IGR}_{(0.6)}$	1.0831	0.407	0.436	0.491	0.586
$\boldsymbol{IGR_{(0.6)}+SCoP}$	**0.7429**	**0.435**	**0.460**	**0.510**	**0.594**
$\mathrm{GS}_{(0.3)}$	1.3657	0.389	0.415	0.462	0.535
$GS_{(0.5)}+SCoP$	1.2583	0.414	0.437	0.482	0.558
$\mathrm{CHI}_{(0.45)}$	1.2583	0.393	0.422	0.477	0.567
$CHI_{(0.45)}+SCoP$	1.0831	0.424	0.448	0.496	0.575

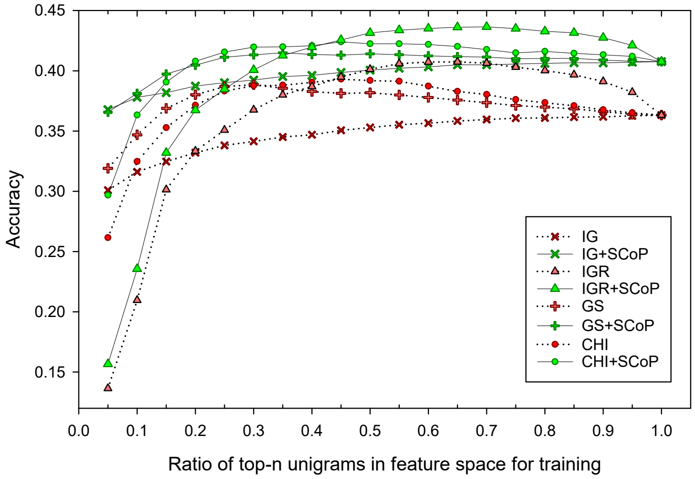

Fig. 3. Accuracies of four baselines and corresponding enhancements using SCoP. Baselines and SCoP enhancements are colored in red and green, respectively.

Figure 3 presents the detailed accuracies of predictions using different selections of top-n features. The results reveal that our proposed SCoP enhancement improves the accuracies of all baseline (unigram) feature selection methods for every setting of top-n. Among the four baselines, the highest accuracy is obtained by IGR, which is even further improved by applying SCoP on it. The results of GS are also worth discussing. The figure shows that if we had to use only the 5% of features (unigrams) for training, the highest baseline accuracy would be obtained by GS. That means, GS makes the most useful top 5% unigram selection among our baselines, and the figure shows that its predictions are further improved by using our proposed enhancement in GS+SCoP.

These results reveal that we can obtain accuracies with SCoP that could not be obtained by the unigram feature selection methods in our experiments. However, since our method adds new bigrams to the training model, we also analyze the size of increase in feature space. The baseline $\mathrm{IGR}_{(0.6)}$, which yields the most accurate predictions among the baselines, uses 32,851 tokens for training (60% of all unigrams). SCoP uses these tokens as primary tokens in Algorithm 1, which detects 10,095 bigrams with significant spatial relationship with respect to the 32,851 unigrams. As a result, IGR$_{(0.6)}$+SCoP uses 42,946 features in total. The number of added bigrams may vary depending on the baseline unigram selection method and the choice of top-n ratio.

Considering the spatial analyses in Algorithm 1, as expected, the detection of spatial co-occurrence patterns causes an increase in the training time of the overall model. We observe two important factors that can affect the training time: 1) the number of co-occurrences of term pairs, and 2) the number of simulations M in Algorithm 1. We refer to line 11 in the algorithm, where the simulation takes $n{=}|X_{t_p t_c}|$ samples from X_{t_p}. As a result of this sampling strategy, bigrams with high co-occurrence frequency negatively affect the execution time. For example, the most frequent term pair in our dataset was *United Kingdom* (n=16,377), and the analysis of this single bigram took more than 60% of the time spent to analyze all bigrams in our training data. Therefore, alternative sampling strategies may need to be devised for larger-scale analyses. Regarding the second performance factor, the number of simulations M in our experiments was 500. We also experimented the effect of using higher M values, and observed that using M=1000 made a change only for 0.7% of the bigrams that were identified with M=500. Therefore, we performed our experiments using M=500 with satisfactory results in reasonable time. We note that since there is no interdependency or sequential relationship in the spatial analyses of bigrams, these operations can also be parallelized and executed in distributed environments.

Finally, we evaluate the utility of our method by comparing it with a setup in which we do not make any particular selection among bigrams. The total number of distinct bigrams that occur in at least five tweets in our training set is 99,452. We trained the model using all of these bigrams and the unigrams selected by IGR$_{(0.6)}$ (i.e., without making the SCoP analysis in Algorithm 1). Our tests in this experiment resulted in an accuracy of 0.387, which is lower even than the baseline's. Therefore we can conclude that an effective analysis of bigrams, as we proposed in this paper, is critical to obtain accurate predictions.

5 Conclusion

In this paper, we introduced a new approach to detect term pairs that exhibit clustering or dispersion tendency in their geographical distribution in relation to the underlying single-term spatial patterns. We used the detected term pairs to improve probabilistic language models and increase the accuracy of content-based location prediction of tweets. We demonstrated that the effective selection of co-occurring terms yields significant improvement in location prediction accuracy. Using purely statistical methods and spatial point patterns enabled our methods to execute without any dependence on predefined gazetteers or external data sources.

In our future research, we plan to adapt our framework in distributed environments and apply our methods for fine-grained location prediction in global scale. Applying discriminative learning models for location prediction and investigating alternative methods to utilize detected attraction and repulsion patterns are also among our future research directions.

References

1. Li, W., Eickhoff, C., Vries, A.P.: Geo-spatial domain expertise in microblogs. In: Proc. of ECIR 2014. (2014) 487–492
2. Paraskevopoulos, P., Palpanas, T.: Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. Soc. Netw. Anal. Min. **6**(1) (2016) 89
3. Melo, F., Martins, B.: Automated geocoding of textual documents: A survey of current approaches. Transactions in GIS **21**(1) (2017) 3–38
4. Zheng, X., Han, J., Sun, A.: A survey of location prediction on Twitter. CoRR **abs/1705.03172** (2017)
5. Han, B., Cook, P., Baldwin, T.: Text-based Twitter user geolocation prediction. J. Artif. Int. Res. **49**(1) (2014) 451–500
6. Priedhorsky, R., Culotta, A., Del Valle, S.Y.: Inferring the origin locations of tweets with quantitative confidence. In: Proc. of CSCW '14. (2014)
7. Han, B., Rahimi, A., Derczynski, L., Baldwin, T.: Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In: Proc. of W-NUT. (2016)
8. Cheng, Z., Caverlee, J., Lee, K.: A content-driven framework for geolocating microblog users. ACM Trans. Intell. Syst. Technol. **4**(1) (2013) 2:1–2:27
9. Van Laere, O., Quinn, J., Schockaert, S., Dhoedt, B.: Spatially aware term selection for geotagging. IEEE Trans. on Knowl. and Data Eng. **26**(1) (2014) 221–234
10. Dredze, M., Osborne, M., Kambadur, P.: Geolocation for Twitter: Timing matters. In: Proc. of HLT-NAACL. (2016)
11. Hauff, C., Houben, G.J.: Placing images on the world map: A microblog-based enrichment approach. In: Proc. of ACM SIGIR '12. (2012) 691–700
12. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: Improving geographical prediction with social and spatial proximity. In: Proc. of WWW '10. (2010) 61–70
13. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: Proc. of EMNLP '10. (2010) 1277–1287
14. Miura, Y., Taniguchi, M., Taniguchi, T., Ohkuma, T.: A simple scalable neural networks based model for geolocation prediction in Twitter. In: Proc. of W-NUT. (2016)
15. O'Hare, N., Murdock, V.: Modeling locations with social media. Inf. Retr. **16**(1) (2013) 30–62
16. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proc. of ICML'97. (1997)
17. Ripley, B.D.: Modelling spatial patterns. Journal of the Royal Statistical Society. Series B (Methodological) **39**(2) (1977) 172–212
18. Ruocco, M., Ramampiaro, H.: Geo-temporal distribution of tag terms for event-related image retrieval. Inf. Processing & Management **51**(1) (2015) 92–110
19. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: Social honeypots + machine learning. In: Proc. of ACM SIGIR '10. (2010) 435–442