Predicting Event Mentions based on a Semantic Analysis of Microblogs for Inter-Region Relationships

Gwan Jang
School of Computing, Korea Advanced Institute of Science and Technology, Republic of Korea

Sung-Hyon Myaeng
School of Computing, Korea Advanced Institute of Science and Technology, Republic of Korea

Abstract
An ability to predict people’s interests in different regions would be valuable to many applications including marketing and policymaking. We posit that social media plays an important role in capturing collective user interests in different regions and their dynamics over time and across regions. Event mentions in microblogs of social media like Twitter not only reflect the people’s interests in different regions but also affect the posting of future messages as the content of microblogs propagates to others through an online social network. Differentiating from the various network analysis techniques that have been developed to capture people’s interests and their propagation patterns, we propose an event mention prediction method that utilizes an analysis of inter-region relationships. We first obtain regional user interests for each topic by applying Latent Dirichlet Allocation (LDA) to region-specific collections of tweets and then compute pairwise similarities among regions. The resulting similarity-based region network becomes the basis for constructing region groups through Markov Chain Clustering (MCL), which helps removing noise relationships among regions. We then propose a relatively simple regression technique to predict future event mentions in different regions. We demonstrate that the proposed method outperforms the state-of-the-art event prediction method, confirming that the novel method of constructing groups from region-based sub-topic interests indeed contributes to the increase in the prediction accuracy.

Keywords
event mention prediction; semantic analysis of microblog collections; interest-based region network and clustering

1. Introduction
Social media platforms such as Twitter generates millions of documents per day. Each user shares various topics and experiences through online social networks. Naturally social media has become a key resource for understanding people’s interests and making critical decisions in such areas as policymaking and business. As such, the propagation and diffusion of information through a social media network has been an important focus in social media research. However, it could be more useful if the change in people’s interests over time in different regions could be predicted beyond the analyses of existing content.

In this paper, we investigate how to predict people’s interests in an event. Social media report and propagate new events and express opinions about them, both of which are referred to as event mentions. These event mentions on social media represent events or news, public or personal, that have occurred in the real world, such as “Obama speaks”, “Ebola outbreaks”, and “watching football”. Propagation of information, particularly event mentions, in social media can help forecasting an actual event, similar to the spread of an influenza outbreak, as well as tracking past events and detecting salient current events through social media analyses [1]. Some recent studies have demonstrated that forecasting is a difficult but useful task [2, 3]; studies have also proposed an approach that considers both spatiotemporal and content features [4, 5].
Previous studies have introduced novel statistical methods that incorporate different types of social media features in forecasting tasks. They have focused on the Twitter content of individual regions and the burstiness of event mentions. In contrast, the proposed approach utilizes inter-region similarities and differences in terms of people’s interests in events, which are reflected in tweets, and constructs a region network and regions groups. The advantage of the proposed approach is that it considers a global view for local analyses and enables inference about inter-region relationships, even when the data is not sufficient for a particular region.

The foundation of the proposed method is related to a previous study where recognizing sub-topic changes in different regions over time provided useful insights into the development and diffusion of people’s interest in a certain topic [6]. For example, given a topic such as a typhoon, people near a beach might raise concerns about the possibility of a tsunami, whereas those in hilly areas might have greater interest in mudslides. However, for a topic related to an epidemic, the sub-topics would differ significantly depending on where its outbreak occurred and they would change as the epidemic spreads to different regions. It would be safe to assume that location-varying sub-topic differences and their changes could be related to the traits of the people in the regions, e.g. local politics, local events, geographic conditions, and so on. Besides, inter-region similarities and differences on a topic in the past can indicate how people would react to an event in the future. Therefore, forecasting an event or predicting event mentions in a particular area would benefit from modelling a global view of different regions as opposed to focusing on a particular region.

In predicting future event mentions (or interests), we exploit analyses that illustrate the changes of sub-topic distributions over time and proposed a prediction model based on a multivariate regression method. An important element in the prediction model is the use of the relationships among regions and the formation of their clusters. Identifying similar regions in terms of their sub-topic distributions enables prediction of event mention occurrences even when there is a lack of data available for a particular region belonging to the cluster. The experimental results to appear later highlight the importance of this capability for predictability.

In this study, Twitter is used because it is considered a valuable data source for understanding the interests of crowds. Because millions of posts are made each day, it has become possible and critical to analyse the data and extract the interests for applications in various fields such as politics, sports, entertainment, and academia.

The key contributions of this study are summarized as follows.

- We propose a novel method for predicting event mentions in a particular region by building and utilizing a global view of the regions, i.e. relationships among all regions based on the sub-topic distributions of individual regions. Unlike the usual event forecasting task, the proposed prediction method assists in predicting how people’s interest in an event can be propagated to different regions.
- As a way to improve the prediction quality, we propose an idea of building region clusters and utilizing inter-region similarities in predicting the diffusion of event mentions. It enables predictions even when the data is not sufficient for a region, in which case the propagation from the region to others would not be reliable.
- Through a series of experiments, we show the efficacy of the proposed method against a state-of-the-art method.

2. Related Work

The work on spatiotemporal event extractions and forecasting with social media is an emerging research topic [1, 4, 7], and many researchers have focused on the detection of events. For example, Ritter et al. [8] extracted events based on linguistic features, by parsing tweet strings to extract both entities and event phrases. Weng et al. [9] attempted to detect events on tweets with wavelet-based signals computed by frequencies of individual words. Similar signals were clustered and assigned an event. They captured emerging events but without considering locality or capturing region-specific events. Akbari and Chua [10] describe their attempt to detect events on a specific area, ‘wellness’, by utilizing a category hierarchy, three high level thematic categories and their 14 sub-categories, and applying a multi-task learning method to consider both task-specific and task-shared features together. The linguistic features include n-grams, named entities, gazetteer and modality. This approach is useful when thematic categories are available for a particular domain.

Spatial features have been utilized to detect events on specific regions. Zhou et al. [11] detected events by location-time constrained topic (LTT) model, where a social message was represented as a probability distribution over a set of topics. While the location-time constraints were used to better extract events, there was no explicit attempt to associate events with specific locations. Walther and Kaiser [12] first clustered tweets based on features derived from text, such as sentiment and subjectivity, and other non-textual features such as tweet counts and coordinates where the tweets were issued. The resulting spatio-temporal clusters were analysed with C4.5 in order to determine whether or not they constitute real-world events. Abdelhaq et al. [13] attempted to extract localized events from tweets by identifying low entropy
keywords that appeared in a few regions with high burstness. Selected keywords within a time window are represented as vectors of their frequencies across the regions and compared to form clusters. The keywords in a cluster are considered to refer to an event.

Zhao et al. [4, 5] proposed a multi-task learning method for spatiotemporal event forecasting across regions at once. They extracted static features with predefined vocabularies and dynamic features that appear frequently in a specific location and time. However, they forecast events for different regions independently from each other, without considering their relationships. In our study, we emphasize inter-region relationships computed based on the way similar content is shared among regions, rather than using the linguistic features available for individual regions.

Identifying local topics and their changes in social media can be important in predicting event mentions. There have been attempts to model and discover the relationships between topics and locations reflected on Twitter [14-17]. For example, Song et al. [14, 15] analysed tweets from spatiotemporal aspects to discover relationships among topics, i.e. queries on Twitter. They compared the spatiotemporal distributions of the documents associated with each topic. While the method can be used to relate topics by observing their spatio-temporal distributions, it does not reveal inter-region relationships, not to mention location-sensitive sub-topics. While these approaches analyse word occurrences in the spatiotemporal analysis of tweet topics, there have been other attempts either using hashtags or queries as the topic [18, 19] or using a topic modelling technique to identify topics [20, 21].

Akkbari and Chua [22] attempted to discover and profile communities among users based on tweets. They use seven different social behavioural features: content (words), hashtags, interactions (reply and retweet), network (following, co-following, and co-followed). They transform the large sparse matrices to low dimensional latent ones through non-negative matrix factorization (NMF) and combine them to discover user communities. Since the main objective of this work is to cluster users based on tweet content, its manipulation of the large sparse matrix may be applicable to our work. But, our work has reduced the region-word space into region-topics space to create a smaller size region-region matrix. The idea of using other types of features for users is interesting but left for future work as it is not clear, for example, how to use following/followee relationships among the users in a particular region.

Several studies [23-25] based on Hawkes Process [26] analysed and predicted the information diffusion through social media focusing on individual users. The Hawkes Process infers the intensity of an event in a timeline. Through the analysis, they determined who influences whom. The approach in [25] generated a directed acyclic graph (DAG) among users through tracking their communications. However, the events might occur simultaneously in a time window, increasing the difficulty of creating a DAG among them. It is also difficult to determine partial orders over regions because they mutually influence each other.

While the relationships discovered in these studies can be utilised for event mention prediction to a certain extent, the resulting relationships are too crude to be useful for sophisticated tasks such as event mention prediction. Furthermore, these methods focus on local analyses of topics without considering how locations are intertwined with common or different perspectives of the topics.

3. Event Mention Prediction

Since an event mentioned in a social media usually have happened in reality, the event itself would be associated with a specific location. However, such an event mention can be diffused to other regions through the social network, or other mentions of the same event can be posted. In either case, event mentions must be tied with users’ interests. As such, understanding people’s interests in the event itself or related content across different regions in the past can help predicting an event mention in a particular region. In other words, the dynamics of people’s interests in temporal and spatial dimensions must be an important factor for enabling a prediction on whether further interests about the event will appear in the social media content from the region.

3.1. Prediction Task

The event mention prediction task is defined over event dynamics that captures the changes of event mentions on social media in the regions over time. We define $O_t$ for an event $e$ as an event mention distribution vector at time $t$ as follows:

$$O_t(e) = (o_1, o_2, \ldots, o_n) \quad (1)$$

where $o_i$ is the frequency of an event mention in region $i$, and $n$ is the total number of regions considered.

The prediction task is to obtain $\hat{O}_{t+1}(e) = (\hat{o}_1, \hat{o}_2, \ldots, \hat{o}_n)$ at time $t + 1$ when the event $e$ is given. If the target event is expected to occur in region $i$ at time $t + 1$, $\hat{o}_i$ becomes 1. Otherwise, $\hat{o}_i$ becomes 0. The task is not to be precise in the
number of occurrences but to determine whether there will be mention text about the event. We measure the prediction performance in precision, recall and F1 by checking whether a predicted mention for an element in $\hat{O}_{t+1}$ is correct.

\[
\text{Precision} = \frac{\# \text{ of correct predictions}}{\# \text{ of predictions}} \tag{2}
\]

\[
\text{Recall} = \frac{\# \text{ of correct predictions}}{\# \text{ of answers}} \tag{3}
\]

\[
F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}
\]

### 3.2. Prediction Model

Our interest lies in making a prediction based on $O_t$, which is constructed from an accumulation of the past tweets in different regions. While it is possible to build a model with time-segmented data that may show the trend of changing interests in the regions, our initial attempt is to build a simple model that takes into account all the collected data of the past as the basis for prediction. That way, we can focus on the investigation of how the overall regional interests need to be captured and processed for the prediction task. Note that we attempt to capture fundamental relationships among the regions based on the semantics of the past tweets, instead of relying on textual features in the past as in the previous studies.

Our prediction model is based on logistic regression and expressed as a linear equation for predicting $Y$, a probability vector generated for an event mentioned at time $t+1$ (or $\hat{O}_{t+1}(e)$ above):

\[
\ln \left( \frac{Y}{1-Y} \right) = XB \tag{5}
\]

\[
\ln \left( \frac{Y_i}{1-Y_i} \right) = \beta_{i0} + \beta_{i1}X_1 + \beta_{i2}X_2 + \cdots + \beta_{ip}X_p \tag{6}
\]

\[
Y_i = \frac{1}{1+e^{-(-\beta_{i0}+\beta_{i1}X_1+\beta_{i2}X_2+\cdots+\beta_{ip}X_p)}} \tag{7}
\]

where $X$ is the vector of event mention frequencies or $O_t$ for the event. $B$ is the regression coefficient matrix, capturing other factors influencing the probability of dependent vector $Y$. $Y_i$ is a probability of the occurrence at location $i$, and $\beta_{ip}$ is the coefficient value for the event mention frequencies of location $p$. By applying a cut-off value, 0.5, predictions are made in a binary value, either positive or negative.

Our goal is to estimate $B$ in such a way that they represent event mention dynamics among the regions and ultimately make accurate predictions. It is conceivable to use a different length of the time span, such as just the past one month, for data accumulation and even consider multiple time spans for temporal dynamics. However, we simply assume the entire data reflects the overall tendency of individual regions about their ongoing interests since our focus in this paper is to investigate how the accumulated past data need to be processed for enhanced predictability. In other words, our question lies in understanding how the event mentions in other regions in the past can influence those of a target region ($Y_i$). An investigation of different prediction models or inference algorithms for further improvements is left for future research.

### 3.3. Inter-region Influences

We explore three ways the coefficient $B$ is computed to capture the way regions influence each other. The simplest is to compute the relationships amongst all the regions and make each region at time $t+1$ is affected by all the other regions. It assumes every region influences each other for prediction and the coefficient is obtained as in (7). Since a logistic regression model assumes linearity of each of the independent variables (or the past occurrences in each region) to the logarithm of odds (or the probability of occurrence in the target region divided by the probability of absence), this model assumes users in a region affects every other region. This simplest model is referred to as the no cluster (NC) case as in Figure 1(a).

For the remaining two cases, we cluster the regions so that we relax the assumption of the mutual influence among the regions. In these cases, regions are supposed to influence each other only when they belong to a mutually influential group (i.e. cluster). For the multivariate linear regression model, we divide the graph into a set of subgraphs resulting from the
clustering process, and the input data for each subgraph becomes $X_{c_i}$ for the coefficient matrix $B_{c_i}$. In this way, we only consider the independent and dependent variables that have sufficiently strong relationships. After the individual occurrence probability of $Y_i$, we union the binary vectors. The resulting equation takes the form in (8):

$$\ln \left( \frac{Y}{1-Y} \right) = \bigcup (X_{c_i}B_{c_i})$$

(8)

where $B_{c_i}$ represents the weight matrix for cluster $i$. Figure 1(b) depicts this situation where two different clustering methods are employed.

Figure 1. Three different methods for event mention dynamics calculation. NC (No Cluster): every region can affect every other region; TDC (Topic-Dependent Cluster) and TIC (Topic Independent Cluster): regions in the same cluster can only affect each other.

4. Generating Region Clusters

TIC assumes regions would show a similar behaviour regardless of the topics whereas TDC assumes closeness of regions would vary depending on the topics at hand. The difference between TIC and TDC lies in whether we use the entire tweet collection or just a subset for a particular topic in calculating the similarity values among the regions. Other than that, we employ the same clustering technique for both TIC and TDC.

To build the interest profiles for the regions and form clusters based on their similarities, we attempt to capture the sub-topic interests of a topic in each region. An interest profile therefore is a sub-topic distribution of region or the extent to which each region shows an interest in each sub-topic. This is an attempt to represent regional interests at a sufficiently meaningful level, i.e. in terms of sub-topics or perspectives of a topic (e.g. ‘football’, ‘Ebola’). By doing so, event mention prediction can be done for an event under a specific topic. The intention is to reduce potential ambiguity of an event mention. For TIC, we adopt the same process but use the entire dataset covering all the topics so that the topics emerging out of the topic modelling process are used, instead of the sub-topics for a specific topic, to represent regional profiles.

More specifically, the first step for the clustering process is to compute a semantic representation of the tweets generated from each region. For this, we follow a recent work on topic versatility [6] where different perspectives or sub-topics of a topic are captured to see how people’s views change over time. Given a set of tweets collected for a topic from across the regions, a sub-topic distribution $\theta = (\theta^1, ..., \theta^k)$ is computed with Latent Dirichlet Allocation (LDA), where $\theta^i$ is a vector representing the $i$-th topic distribution of a region and $k$ is the size of sub-topics. The sub-topic distribution $\theta$ represents different interests for a topic in a region. By doing so, we attempt to represent a regional interest in a naturally emerging sub-topics for any given topic. For LDA, we use mallet [27] with the size of sub-topics


set to 100. Given sub-topic distributions for all the regions, it becomes possible to construct a \( n \times n \) matrix where \( n \) is the number of regions and each element represents similarity between two regions.

From the sub-topic distributions obtained by LDA, we then construct an undirected graph \( G = (V, E) \) where \( V \) is a set of nodes representing the regions and \( E \) is a set of edges connecting node pairs whose similarity exceeds a threshold. The weight of each edge is the cosine similarity between the sub-topic proportions of the corresponding node pair and obtained as follows:

\[
e_{ij} = \frac{\theta_i \cdot \theta_j}{|\theta_i||\theta_j|} \quad (5)
\]

To cluster the regions from the graph, we apply the Markov Chain Clustering (MCL) method [28], a graph partitioning algorithm. Each cluster contains a group of regions with similar sub-topic distributions. Given an adjacency matrix \( G \) with weights, the algorithm as in Algorithm 1 produces a segmented network \( M_1 \) as output where each segment is equivalent to a cluster. The method focuses on the information flows through the network utilizing the Random Walk algorithm. Its main idea is that nodes become members of a cluster when they are the destinations from a specific vertex with the same probability. It means that the members of a resulting cluster should share the similar information, forming a community in the network.

**Algorithm 1. Markov Chain Clustering (MCL)**

Input: \( G = (V, E) \)

\( \Gamma \) is an inflation parameter

\( M_1 \) is a matrix of random walks on \( G \)

while (change) {

\[
M_2 = M_1 \times M_1 \quad \# \text{ expansion}
\]

\[
M_1 = M_2^\Gamma / \sum_r M_2^{1, r} \quad \# \text{ inflation}
\]

change = difference\((M_1, M_2)\)

}

set CLUSTERING to the components of \( M_1 \)

5. **Experimental Results**

We conducted two experiments. The first one is to establish a reasonable baseline of the proposed method in comparison with a state-of-the-art for event forecasting so that the subsequent experiments would be meaningful. The second is to evaluate the three different methods for weight matrix calculation, including the one used for the baseline in the first experiment. The main goal of the experiments together is to make a convincing argument that the sub-topic level interest dynamics among the regions indeed help improving effectiveness in predicting even mentions.

There are two reasons for conducting the staged experiments instead of a direct comparison between the state-of-art and the proposed method. First, it was not possible to apply the proposed method to the dataset used in the literature of the state-of-art method because there is no way to compute the relationships among the cities from the dataset used in Multitask Feature Learning (MTFL) [5], which are critical for the proposed method to be applied. Second, it was not possible to replicate the state-of-art method in our experimental setting because the term selection method was not well documented in the paper with some heuristics for selecting seed terms. Whatever method we would employ to implement the state-of-art method in our experimental setting would not be identical to MTFL, and the comparison result would not be meaningful. As a result, we first show that the baseline in our proposed approach is already compatible with the state-of-the-art method in a similar task (event forecasting) and then show that the more advanced methods outperform the baseline in a similar task.

5.1. **Experiment 1. Comparison against MTFL**

In order to establish a solid baseline, the method with no clustering (NC) was compared against Multitask Feature Learning (MTFL) where the features of various types in the training documents are combined to forecast an event in individual locations. For evaluation, we opted for the task of forecasting civil unrest events in Mexico as used in the article. The
source code and dataset were provided by the authors\textsuperscript{1}. The dataset consists of 1,217 events obtained from 51 million tweets, which are divided into the training set covering from July 1, 2012 to December 31, 2012 and the test set from January 1, 2013 to May 31, 2013.

The input data for MTFL is the frequency of keyword features for each location on each day. The output value indicates the occurrence of the unrest event for each location on the next day: ‘1’ if the event occurred or ‘0’ otherwise. The input data for the NC case is the occurrences of unrest events for all locations each day from the training data set. In other words, input vector $X_t$ at time slot $t$ is the occurrence vector $O_t$ at time slot $t - 1$. The output value is either 0 or 1 as in MTFL.

The MTFL method applies a regression method for forecasting an event in a location on the next day based on the term features used in that location. It selects text features that possess region-specific and global importance, but does not capture an inter-region effect with which an event message spread to other locations can make the event occur.

As in Table 1, our method significantly outperforms the MTFL method in terms of precision, recall and F1. From the result, we are confident that the proposed method of using the relationships among the locations is more effective in event forecasting than using the local features only. MTFL uses features revealing civil unrest on an individual location for forecasting itself at a later time, whereas the proposed method takes into account the event mentions in one location that might have influenced the other locations. The result strongly suggests that understanding the occurrences of an event in other locations at time $i$ is important in predicting its occurrence in the target location at $t + l$. In other words, it would be difficult to forecast an event in location $i$ without considering relevant features from the location $j$ that affects location $i$.

**Table 1.** Relative performance of the baseline of the proposed method for the event forecasting task

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTFL</td>
<td>0.718</td>
<td>0.793</td>
<td>0.753</td>
</tr>
<tr>
<td>NC</td>
<td>0.815 (+14%)</td>
<td>0.861 (+9%)</td>
<td>0.837 (+11%)</td>
</tr>
</tbody>
</table>

**5.2. Experiment 2. Effects of the clustering methods**

The second experiment was designed to compare the three different ways of using the inter-region relationships for the task of predicting event mentions. We use OSU Twitter NLP tools\textsuperscript{2} discussed in [8, 29] to extract event phrases from Twitter. Under the assumption that an event consists of an entity and an event phrase [8], the tools parse Tweet strings and extract entities and event phrases. Event phrases are annotated using the guideline similar to the way EVENT tags are used in Timebank [30]. We use these tools for the purpose of extracting event phrases for a designated topic; we ignore entities because our goal is to predict an event mention in the form of <topic, event phrase> pair (see Section 5.2.1 for details about how topics are extracted). For topics such as ‘obama’, ‘weekend’, and ‘Alabama’, for example, extracted event phrases would be ‘speak’, ‘go out’ and ‘job opening’, respectively. Note that the tools simply extract general event phrases without any guarantee that they refer to actual events that happened in the real world. In addition, they are not capable of recognizing a phrase representing different events or different phrases representing the same event. They focus on the ability of processing a large amount of data quickly rather than attempting to achieve the highest precision and recall in extracting event phrases. Table 2 gives a summary of the collection, showing the number of identified event phrases.

**Table 2.** A summary of the event mention collection

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Event-14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>2014.08.01-2014.10.03</td>
</tr>
<tr>
<td>#Topics</td>
<td>36</td>
</tr>
<tr>
<td># of Event Phrases</td>
<td>252</td>
</tr>
<tr>
<td># of Event Mentions</td>
<td>171,594</td>
</tr>
</tbody>
</table>

In this experiment, our task is to predict whether an event mention would occur in the next time window based on its occurrences in the past data. The time window in the current experiment was set to 12 hours. Please note that our goal is not to distinguish between the events that actually occurred in the real world and those just posted on Twitter but to just predict such event mentions in the SNS.
5.2.1. Data collection
Assuming that tweets posted by people in different regions and at different times reflect their interests, we obtained the data collection used in [6], the tweets in 266 US metro cities from Aug. 1, 2014 to Oct. 3, 2014 (referred to as Event-14 hereafter). The US metros were obtained from Google Trends that adopted those defined by Nielsen in accordance with the US federal government’s Metropolitan Areas. The most frequent 1,000 hashtags were used as the candidate topics and the sparse topics occurring in less than 20 time windows and five cities (or regions for consistency) were eliminated so that the evaluation would be done with statistical reliability.

For event phrases, we selected the 100 most frequent ones for each topic and then filtered out those occurring less than 20 consecutive time windows for the topic, resulting in 50 topics and 4,930 event phrases. By filtering out the events that appeared only in a short time period, we attempt to guarantee a sufficient amount of training data for each phrase. After filtering out the sparse events occurring in less than 10% of possible locations, we obtained 252 event phrases with 171,594 occurrences in total under 36 topics. By retaining the event phrases that occur in more than 10% of the locations during 20 consecutive days, we eliminate temporary and overly location-specific events in the prediction task. The data set was divided into training and test sets: the first 80 percent of data in time series was used for training as the past and the remaining 20 percent for testing or predicting the future.

5.2.2. Clustering results
We built two different clustering results: topic-independent and topic-dependent clusters as in Figure 2. The clusters represent the regions that share similar sub-topic distributions. The nodes with the same colour belong to the same cluster while the singletons are in white. Note that the coloured edges only help identifying the regions. The topic-independent clustering result indicates general homogeneity among the member regions across the topics whereas the topic-dependent clustering reflects the characteristics of regions for individual topics such as ‘Ebola’ and ‘Android’. Figure 1 shows clearly that the clusters are different not only between TIC and TDC but also between topics. The difference among the topics is a strong indication that the proposed method of using semantics of the tweets would be meaningful in distinguishing the interests of the people and hence predict the event mentions.
5.2.3. Event mention prediction
We calculate the coefficient matrix $B$ for each of the 252 event phrases, generating 252 linear models. A prediction of a given event mention in a location at time slot $t$ is correct if the mention actually occurred in the test set. Table 3 shows a prediction result for the Event-14 data collection, comparing three different clustering options: NC, TIC and TDC. The result shows a macro average in precision, recall, and F1, which is the average of individual models’ performances.

As expected, the clustering helped improving predictability. Contrary to our expectation, however, TIC gave a better performance than TDC in precision, recall and F1 on average, with the main difference in recall. A conclusion out of this result is two-fold. First, removing irrelevant regions for prediction based on inter-region relationships through clustering at the training time is useful in event mention prediction. Second, the clustering contributes mainly to recall enhancement for the TIC case. This result seems attributable to the experimental setting in which we selected the topics with highly frequent event phrases in order to ensure a sufficiently large number of training instances. As a result, TDC was no better
than TIC on average because the omnipresent event phrases across many regions did not need clustering. A detailed case analysis is described in section 6.1.

Table 3. The effect of region clustering

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Macro-Precision</th>
<th>Macro-Recall</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>0.772</td>
<td>0.798</td>
<td>0.785</td>
</tr>
<tr>
<td>TIC</td>
<td>0.776</td>
<td>0.878</td>
<td>0.824</td>
</tr>
<tr>
<td>TDC</td>
<td>0.773</td>
<td>0.821</td>
<td>0.796</td>
</tr>
</tbody>
</table>

6. Discussion

6.1. Topic Independent Cluster (TIC) vs. Topic Dependent Cluster (TDC)

While the average result shows TDC was worse than TIC, the former was better than the latter for 5% of the event phrases. Based on the observation that those event phrases are sparse in their frequencies, we investigated whether the performance had anything to do with sparsity of the event phrase occurrences. We first expanded the Event-14 collection dataset to build a more inclusive dataset consisting of 4,930 event phrases for 50 topics by skipping the filtering of the less frequent ones. Now the new dataset includes event phrases that appear only in a small number of regions or for a small number of time windows. A new experiment shows that the percentage of the event phrases for which TDC is better than TIC increased to 49% with the expanded dataset.

Another important factor in determining when to use TDC is the number of clusters generated for a topic. As in Figure 3 that shows individual topics plotted for the number of clusters (x-axis) and the ratio of TDC wins (y-axis), TDC tends to be more preferred as the number of clusters increases. More precisely, TDC is almost better when the number of clusters is greater than 25.

Figure 3. Relationship between the number of region clusters and the ratio of TDC wins for each of the 50 topics

The example in Figure 4 shows two clustering results that illustrate when TDC can be better: the first case in (a) where TIC is better (topic ‘baby’) and the other where TDC is better (topic ‘news’). Since the colours represent the regions in the same clusters, the topic ‘baby’ has one dominating region group with a few exceptions forming a minority group whereas the topic ‘news’ has more distinct groups of regions. Note that the existence of such a dominating group makes the clustering meaningless, making the result similar to the NC case. Therefore, it would be more meaningful to follow the general inter-region similarities resulting from TIC as in Figure 2 (a). The existence of a dominating group indicates that the regional differences in terms of sub-topic distributions are so small that it would be better to follow the overall differences and similarities among the regions unless it results from errors in the topic-specific clustering.
6.2. Event mentions in a minority group

An advantage of the cluster-based prediction method is that it becomes possible to identify and utilize minority groups in predicting event mentions. This is an important capability of the technique because minority opinions can be easily ignored when we only rely on a technique that only computes the overall trend of people’s interests. While it is easy to understand what sub-topics or issues exist using sub-topic analyses from topic versatility analyses [6], it is not so easy to discover a minority group consisting of a small set of regions without clustering all the regions as in the proposed method.

We ran a further experiment where we only considered 36 topics revealed minority groups. If a cluster contains less than six regions, it is regarded as a minority group. Like the aforementioned experiments, the first 80% of the occurrences of each event phrase were used for training and the remaining ones for testing, respectively. Table 4 presents the prediction result showing comparison of the performance of TDC against those of NC and TIC. For the prediction task using TDC, only the minority regions participated in generating the frequency vectors for TDC as well as were chosen as the targets. It clearly shows that TDC is helpful in predicting event mentions in minority regions although the overall performance for such a topic using TDC is inferior to TIC. Identifying a minority group is important in making region-specific decisions, particularly when there is little data available for a group member.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>0.789</td>
</tr>
<tr>
<td>TIC</td>
<td>0.822</td>
</tr>
<tr>
<td>TDC</td>
<td>0.846</td>
</tr>
</tbody>
</table>

7. Conclusion

For the event mention prediction task where the past history of an event mention in tweets is used to predict whether it would occur in future tweets from a particular region, the general strategy of considering another region as a message is propagated into other regions. Furthermore, with a relatively simple inference method using a multivariate regression model, the inter-region relationships computed from the sub-topic analyses of tweets can be considered. Moreover, Markov Chain Clustering is used in two different ways: one for all tweets regardless of the topic that they belong to (TIC) and one for those belonging to each topic or topic-dependent way (TDC).
In order to validate the method, the baseline of considering other regions is compared with a current method for event forecasting, and this demonstrated the superiority of the proposed method. It was also demonstrated that the clustering methods led to a significant improvement. While TIC was demonstrated to be superior on average, TDC was better for 15.5% of the event phrases. Further analysis indicated that TDC was effective when the event mentions were sparse across regions and time windows, and when predictions needed to be made in minority regions.

For future research, a more sophisticated inference mechanism beyond regression will be investigated. Furthermore, different methods for computing inter-region relationships relevant to events will be investigated. More elegant methods of extracting event mentions will be also considered.

Notes
2. https://github.com/aritter/twitter_nlp

References


