Content-Based Geo-Location Detection for Placing Tweets Pertaining To Trending News on Map

Saurabh Khanwalkar[1], Marc Seldin[1], Amit Srivastava[1], Anoop Kumar[1], Sean Colbath[1]

[1]Raytheon BBN Technologies, Cambridge, MA, USA
{skhanwal,mseldin,asrivast,akumar,scolbath}@bbn.com

Abstract. In the last few years, social media services such as Twitter have proven to provide first-line information on current news events such as civil unrest, protests, elections, etc. The limited availability of self-identified geo-location information makes it challenging to place such current and trending news on the globe. In this paper, we demonstrate a novel approach for content-based geo-location of Arabic and English language tweets by collating contextual tweets into a document using a user-tweeting-frequency based temporal window. We compute the distances of the content-based geo-located tweets against the device-based geospatial points provided natively via Twitter API as well as the Twitter User Profile based locations. We show that content-based geo-location detection provides an effective way of geo-localizing trending news topics, geo-political entities and Hashtags.

Keywords: Geo-location, twitter, social media

1 Introduction

Micro-blogging services such as Twitter have become a very popular communication tool among Internet users, being employed for a wide range of purposes including marketing, expressing opinions, broadcasting events or simply conversing with friends. Each day, more than 200 million active users publish more than 400 million tweets per day in the social network, sharing significant events in their daily lives [1]. Additionally, Twitter allows researchers unprecedented access to digital trails of data as users share information and communicate online. This is helpful to parties seeking to understand trends and patterns ranging from customer feedback to the mapping of health pandemics [2]. As explained by T. Sakaki et al. [3], every Twitter user can be described as a sensor that can provide spatiotemporal information capable of detecting major news events such as earthquakes or hurricanes.

Location is a crucial attribute to understanding the ways in which online flow of information might reveal underlying economic, social, political, and environmental trends. Localization facilitates temporal analyses of trending news topics from a geo-spatial perspective, which is often useful in further analysis. Studies such as [4] and [5] have addressed the capability to track emergency events and how they evolve, as people usually first post news on Twitter, and are later broadcast by traditional media corporations [6]. One of the biggest challenges is identifying the location where events are taking place.

Twitter supports per-tweet geo-tagging feature which provides extremely fine-tuned Twitter user tracking by associating each tweet with latitude and longitude coordinate points. In our sampling of 20 million tweets, less than 0.70% of all tweets actually use the geospatial tagging functionality. When this feature is enabled, it generally functions automatically when a tweet is published with the coordinate data coming either from user's device itself via GPS, or from detecting the location of the user's Internet (IP) address. Additionally, these features do not provide location estimates based on the content of the user-posted tweet messages.

Effective geo-location of tweets based purely on their textual content is a difficult task, and although Twitter provides vast amounts of data, it introduces several natural language processing (NLP) challenges:

- Multilingual posts and code-switching [7] between languages makes it harder to develop language models and often needs Machine Translation (MT).
- With the limitation of 140 characters per-tweet, Twitter users often use shorthand and non-standard vocabulary which makes named-entity detection and geo-location via gazetteer more challenging.
- Twitter content tends to be very volatile, and pieces of content become popular and fade away within a matter of hours.

In this paper, we present a content-based, geo-location detection approach that is capable of geo-locating multilingual tweets, within a time window, by exclusively using the textual content of these tweets. Our premise is that tweets encode geospatial location-specific content; either specific place names or named-entities. Additionally, our intuition is that within a time window, Twitter users tweet specific to their current location or specific to localized trending events which are of interest.

The rest of the paper is organized as follows: in Section 2, we review the related work on content-based geo-location detection. In Section 3, we describe the dataset, and the evaluation metrics we used to benchmark our geo-location detection performance. In Section 4, we explain our approach to content-based geo-location detection for placing tweets on a map. In Section 5, we present results from performance evaluation of our geo-location detection algorithm, followed by examples of using geo-location detection for placing tweets pertaining to trending news on map in Section 6.

2 Related Work

Recently, content-based geo-location detection techniques have been explored, some focusing on supervised and language model based approaches, while others focusing on location name-based approaches. Applications vary from providing relevant advertisements, to public health awareness, user modeling and tracking trending news events. Cheng et al. [8] propose a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of that user's tweets.

Roller et al. [12] present a supervised, text-based geo-location using language models on an adaptive grid. Given training documents labeled with latitude and longitude coordinates, *pseudo-documents* are constructed by concatenating the documents within a grid cell overlaid on Earth; then a location for a test document is chosen based on the most similar pseudo-document. Paradesi [13] explores research that identifies the locations referenced in a tweet and show relevant tweets to a user based on that user's location. For example, a user traveling to a new place would not necessarily know all the events happening in that place unless they appear in the mainstream media. The proposed system, called TwitterTagger geo-tags tweets in near real-time and shows tweets related to surrounding areas.

The contributions of this paper towards content-based geo-location research are:

- Novel approach for collating multilingual tweets into a temporal document using user-tweeting-frequency based time window;
- Named-entity detection and geospatial points-based clustering;
- Location-specific feature set calculation and document scoring for best-match content-based location identification for a document.

3 Dataset, Evaluation Setup and Metrics

We developed a process to identify Social Media users across the Middle East region who are influential contributors on the Twitter social media platform. Through this process, we created a list of Twitter users, culled from mainstream journalism feeds, diplomatic circles, and political circles having wide Arabic regional appeal. We collected tweets over a period of 3 months (January 2013 to March 2013) using the Twitter Spritzer streaming API with a filter for our selected users of interest. Using this setup, we collected approximately 17 million multilingual tweets distributed into 85% Arabic, and 15% English from 2.6 million Twitter users.

To evaluate the performance of our tweet geo-location detection algorithm, the first metric we consider is the ***Error Distance***, which quantifies the distance in miles between the actual geo-location of the tweet $l_{act}(t)$ and the estimated geo-location $l_{est}(t)$ [8]. The Error Distance for tweet *t* is defined in equation (1) as –

$$ErrDist(t) = d\big(l_{act}(t),\ l_{est}(t)\big) \tag{1}$$

The overall performance of the content-based tweet geo-location detector can further be measured using the Average Error Distance across all the geo-located tweets T using equation (2) –

$$AvgErrDist(T) = \frac{\sum_{t \in T} ErrDist(t)}{|T|} \tag{2}$$

A low Average Error Distance indicates that the detector may geo-locate tweets close to their geo-location on average as provided by the user profile or user device. This metric does not provide more insight into the distribution of the geo-location detection errors. We apply maximum allowed distance in miles thresholding at three points; 100 miles, 500 miles and 1000 miles and calculate the next metric, Accuracy100, Accuracy500 and Accuracy1000 using equation (3) –

$$Accuracy_K(T) = \frac{|t|t \in ErrDist(t) \leq K|}{|T|}\ where\ K\ is\ distance\ in\ miles \tag{3}$$

4 Technical Approach

In this section, we present our approach for content-based geo-location detection as outlined by the processing flowchart in Fig. 1.

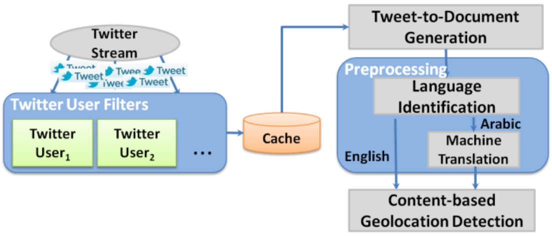

Fig. 1. Processing flowchart for content-based geo-location showing all the stages, starting from tweets collection, document conversion, preprocessing and geo-location detection

4.1 Tweets-to-Documents Generation

We developed an approach for generating cohesive documents from tweets that can be used as a subject of analysis by Information Extraction (IE) algorithms. The motivation for defining *document* was two-fold; (1) a single tweet is limited to 140 characters and may not have sufficient content for estimating location that corresponds to a specific topic, and, (2) most Twitter users post tweets on specific trending topics and move on to other topics within a certain temporal window [10]. This approach is formulated in Algorithm 1.

Algorithm 1 Tweet-to-Document Conversion

Input: *tweets:* List of n tweets from m Twitter users in time window t
minWindowSize: The minimum size of the time window in hours
maxWindowSize: The maximum size of the time window in hours
minTweetsInWindow: The minimum number of tweets per-user in a time window
maxTweetsInDocument: The maximum number of tweets allowed in a document
Output: *documentList*: List of documents in time window t
Notation: { } – List, [ ] - Array

```
startTime = currentTimeInHours
epochStartTime = startTime
While (True)
    timeSpan = startTime – epochStartTime
    windowSize =(timeSpan>=maxWindowSize)?maxWindowSize : minWindowSize
    endTime = currentTimeInHours + windowSize
    foreach userTweetTime in userTweetTimeTable
        if (created_at  >= startTime && created_at  < endTime)
            userTweetList.Insert(userTweet)
    foreach userTweet in userTweetList
        tweets = userTweetList[userTweet]
        foreach tweet in tweets
            Document.Add(tweet);
            if(Document.Size >= maxTweetsInDocument)
                DocumentList.Insert(Document)
    startTime = endTime
```

4.2 Preprocessing

Once all the tweets in a time-delineated window are converted into documents, such that each document contains multiple tweet posts from a specific user, we preprocess all the documents in preparation for content-based geo-location detection. First, we perform n-gram based language identification [10] to identify Arabic versus English tweets and translate Arabic tweets into English using the SDL Language Weaver Machine Translation (MT) system. The geo-location detection algorithm operates on source English tweets and the MT-English equivalent of the Arabic tweets.

4.3 Content-based Geo-location Detection

Our geo-location detection algorithm has three distinct phases as shown in Fig. 2. In the first phase, tweets that were grouped into a time-delineated content window via the document generation algorithm described in section 4.1 are submitted to a named entity detection algorithm [11].

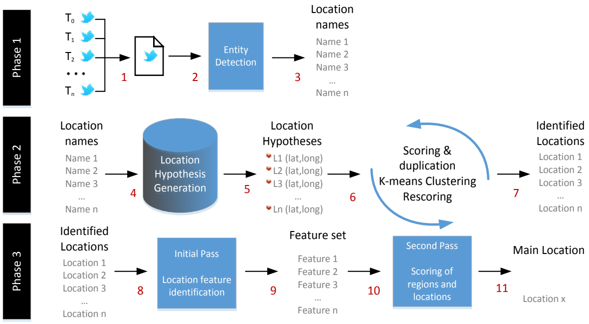

Fig. 2. Three phases of the Content-based Geo-location Clustering and Detection Algorithm

In phase two, the list of named entities which were discovered in phase one is now employed to select location records from several gazetteers. Each match is then given a preliminary score based on features both internal to the location record and features from external sources. Points are then duplicated proportionally to their scores to create a weighting scheme for k-means clustering. The randomly assigned points are then rescored based on how close they are to their cluster's center or centroid location. Finally, location identities are assigned to location names according to their membership in the cluster with the highest score containing that name.

The third phase is concerned with selecting the best overall location associated with the document. This phase begins by iterating through the locations identified in the previous step. During this initial pass, common features such as political administrative unit membership are identified, as well as other features such as order of occurrence. In a second pass, each location is scored by comparing it to the results of the first pass; certain features are biased and others receive an anti-bias. After each point is scored, the highest scoring location is returned as the estimated location.

5 Results and Discussion

A key point to be noted is that our geo-location detection evaluation is based solely on the location of the users where they were tweeting. While these results help us assess the performance of geo-location detector, we believe that creating a manually annotated set would allow use to demonstrate greater accuracy. This is due to the discrepancy between a user's physical location and the topic a user may be tweeting about. For example, a user from Boston, MA, USA might be traveling in Egypt, while tweeting about trending news in Syria.

5.1 Comparison against device-based geo-location provided by Twitter

To minimize outliers, we filtered tweets that are from potential spammers based on 2 criteria; (1) filter tweets that are not from our core selected users, and, (2) filter tweets that are auto-generated by advert spreading tools. After filtering, we had approximately 50K tweets with Twitter-provided device-based geospatial data in terms of latitude and longitude points. Table 1 shows the results of our content-based geo-location detection algorithm using metrics defined in section 3.2.

Table 1. Performance of our content-based geo-location detection

AvgErrDist (Miles)	Accuracy$_{100}$	Accuracy$_{500}$	Accuracy$_{1000}$
1881.98	0.122	0.321	0.497

We found that only 12% of the 50K tweets in the test set could be geo-located within 100 miles of their device-provided geospatial points and that the AvgErrDist across all 50K was 1,881 miles. The accuracy does improve close to 50% for tweets that could be geo-located within 1000 miles of their device-provided location.

5.2 Comparison of results after varying algorithmic parameters

For our baseline evaluation, we set the parameters *minWindowSize* and *maxWindowSize* of our Tweet-to-Document generation (Algorithm 1 described in section 4.1) to 4 hours and 8 hours respectively. These values were motivated by an initial assessment that users tweet on a specific topic for a short period and move on to other topics of interest that are trending on that specific day. The *maxWindowSize* parameter controls the maximum time window allowed for the user's tweets such that they are considered localized to specific topic or news story. In Table 2, we present some results with variation of these parameters.

Table 2. Impact of Tweet-to-Document Generation parameter adjustments on content-based tweet geo-location

Method	AvgErrDist(Miles)	Accuracy$_{100}$	Accuracy$_{500}$	Accuracy$_{1000}$
Base (min:4,max:8)	1881.98	0.122	0.321	0.497
Var1 (min:2,max:8)	773.43	0.313	0.392	0.574
Var2 (min:2,max:4)	**693.24**	**0.377**	**0.412**	**0.581**

In Variant 1, we changed *minWindowSize* parameter to 2 hours which reduced the contextual time window, leading to smaller length documents. We noticed that Accuracy$_{100}$ increased by 156% relative to our baseline parameters and the AvgErrDist also reduced to 773 miles from 1,881 miles. This indicates that, even though shorter time window leads to smaller length documents, the content is more localized to a specific region. In Variant 2, we changed both, the *minWindowSize* and *maxWindowSize* parameters to 2 hours and 4 hours respectively. This lead to a further improvement in Accuracy$_{100}$; 209% relative to baseline and 20% relative to Variant 1. This improvement indicates that a time window of 4 hours leads to a more optimal context for all tweets that pertain to topic or news story.

6 Twitter Trends on a Map

Our application of content-based geo-location detection is to segregate tweets pertaining to specific hashtags or trending news story and localize them on the global map. Such geo-location leads to detection of news or events that are trending in a specific city, country or region.

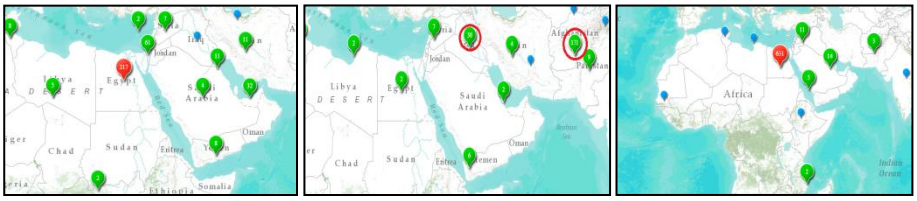

Fig. 3. Examples of news trends on a map displays the output of our geo-location detection system as clusters of geospatially distributed tweets matching a search query

As shown in Fig. 3 leftmost map, we searched our database of more than 20 million tweets using the keyword “muslim brotherhood” and displayed the top 1000 tweet results on the global map. As expected, the largest number of hits for this keyword query put the tweets on Egypt. The map in the middle shows an example of an event “roadside bomb” that was trending in and around countries in Middle East on July 3, 2013 and Google News reported roadside bombs in Baghdad, Afghanistan and southern Thailand. The majority of tweets are distributed around Afghanistan and Iraq with a few outliers that mention the keyword “roadside bomb” and are geo-located in India and Yemen. Finally, the rightmost map shows an example of Hashtag #30June that was trending during July 3, 2013 and pertained to trending event “protests in Egypt” that happened on June 30, 2013.

7 Conclusion

In this paper, we present an approach that incorporates two novel algorithms; (1) user-tweeting-frequency based time window to collate multilingual tweets into a document, and, (2) location-entity clustering and disambiguation, for content-based geo-location detection. We compare our geo-location detection with Twitter-provided device-based and user-profile-based geospatial coordinates and show that we are able geo-locate 58% of the tweets in the test set within 1000 miles with algorithmic parameter adjustments. Furthermore, our content-based geo-location algorithm operates not only on native English but also on machine-translated English tweets, thereby, enabling multilingual tweet geo-location. We demonstrate an application of content-based geo-location of tweets through examples of country-localized trending keyword and geo-political entity.

8 References

1. K. Lerman, R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM), 2010.
2. Zook, M.; Graham, M.; Shelton, T.; and Gorman, S. 2010. Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. World Medical & Health Policy 2(2):7–33.
3. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. WWW ’10 Proceedings of the 19th international conference on World Wide Web, pages 851–860, 2010.
4. F. Abel, C. Hau, G. J. Houben, R. Stronkman, and K. Tao. Semantics+filtering+ search=twitcident. Exploring information in social web streams. In Proceedings of the 23rd ACM conference on Hypertext and social media, page 285294, 2012.
5. S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazardsevents: what twitter may contribute to situational awareness? In Proceedings of the 28th international conference on Human factors in computing systems, 2010.
6. A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management, 6(3):248260, 2009.
7. Sebba, Mark, Shahrzad Mahootian, and Carla Jonsson. Language Mixing and Code-Switching in Writing: Approaches to Mixed-Language Written Discourse. Routledge Critical Studies in Multilingualism. Routledge, Taylor & Francis Group.
8. Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10).
9. Fabian Abel, Qi Gao, Geert-Jan Houben, Ke Tao. Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web In Proceedings of ACM WebSci '11, 3rd International Conference on Web Science, Koblenz, Germany (June 2011)
10. William B. Cavnar, John M. Trenkle. N-Gram-Based Text Categorization (1994). In Proceedings of SDAIR-94.
11. S. Miller, J. Guinness, A. Zamanian. Name tagging with word clusters and discriminative training In Proceedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting, Vol. 4 (2004).
12. Roller, Stephen, Speriosu, Michael, Rallapalli, Sarat, Wing, Benjamin and Baldridge, Jason. "Supervised Text-based Geolocation Using Language Models on an Adaptive Grid." Paper presented at the meeting of the EMNLP-CoNLL, 2012.
13. Paradesi, Sharon Myrtle. "Geotagging Tweets Using Their Content." In FLAIRS Conference. 2011.