Tweet Location Detection

Bahareh Rahmanzadeh Heravi
Insight Centre for Data Analytics
National University of Ireland
Galway, Ireland
Bahareh.Heravi@insight-centre.org

Ihab Salawdeh
Insight Centre for Data Analytics
National University of Ireland
Galway, Ireland
Ihab.Salawdeh@insight-centre.org

ABSTRACT

Mapping Twitter conversations on maps over time has become a popular way of visualising conversations around events on Twitter. Large events have been the subject of most of these types of visualisations, where the rate of geo-tagged tweets is high enough to make interesting visualisations over the selected time period. However, in the case of smaller events, or smaller countries where the frequency of tweets generated for events is lower, we are naturally faced with a low number of geo-tagged tweets, which makes it uninteresting to use these data for mapping and visualisations. This paper demonstrates application of Twiloc - a tweet location detection system - for mapping the conversation around an EU Qualifiers match between Ireland and Scotland. The paper further presents a small comparison between the results obtained from Twiloc and CartoDB Twitter Maps for Dublin Marathon tweet dataset. Twiloc uses various features for determining the location of every single tweet it receives, resulting in a significantly higher rate of tweets with associated location information, and hence enables tweet location analysis and visualisation for smaller events.

Categories and Subject Descriptors

J.5 [**Computer Applications**]: Arts and Humanities

General Terms

Algorithms, Design, Experimentation

Keywords

Data Journalism, Computational Journalism, Twitter, Location Detection, Geo Referencing, User Generated Content, Data Visualisation, Natural Language Processing, Social Semantic Journalism.

1. INTRODUCTION

Visualisation of tweets over time for various events has become popular in the past number of years. Examples are the animated Sunrise on Twitter map, the Super Bowl 2014 tweet visualisation, Mapping “Happy NewYears” 2014 around the world, FA Cup Final 2014 tweet visualisation, the 2014 Indian Election on Twitter and Geography of Hate in the US. They are eye catching and informing maps, normally loved by audiences, and shared widely. However, the main question here for journalists is: “How to determine the location?” There are two rather straightforward and user friendly approaches for journalists to do so: (1) use only geo-tagged tweets, or (2) use services such as CartoDB Tweet Map [1] for tweet collection, geo referencing and mapping. These are briefly explained in the following:

(1) Geo-tagged tweets: Depending on the type and location of events, only around 1% of tweets are normally geo-tagged [7]. This means if an event is large enough, such as the examples above, one is likely to get a decent number of geo-tagged tweets for visualisation. For example if one collected 2 million tweets for a 2 hour event, there would be approximately 20,000 geo-tagged tweets, which makes it possible to create a flaring visualisation during that period, i.e. ~167 geo-tagged tweets per minute on average. However, if the event is not as popular or widespread, or if it is a local or national event in a small country, there will likely not be as many geo-tagged tweets found in the collection of tweets gathered for the event. For example, we collected tweets for the Euro 2016 Qualifiers football match between Ireland and Scotland, and it resulted in around 20,000 tweets in total within 2 hours. Using the 1% rule of thumb, we would have only 200 tweets to use for visualisation, which means 0.3 geo-tagged tweets per minute on average, which is not nearly enough for an interesting visualisation.

(2) Using CartoDB: CartoDB maps and in particular Torque is an interesting and easy to use method for visualising tweets on a map and over time. CartoDB can be used for mapping tweets with geo-tagged information. In addition to this, CartoDB provides its own tweet collection and geo-referencing services, called Twitter Maps. This solution geo-references tweets using proprietary algorithms,, and thus results in more geo referenced tweets than only the ones with GPS information. How CartoDB geo references tweets is a black box and not much information can be found on their algorithm and processes. Looking at the data, the authors suspect they may make use of GNIP geo referencing, which considers more factors for location detection, including GPS, Profile Location and Mentioned Location [2, 3]. CartoDB Twitter Maps is a seamless and easy to use service for collecting and mapping tweets, but costly for users, and also is only offered as an add-on to Enterprise customers. Enterprise plans price starts from $9,900 a year (at the time of writing the paper) [4], and for Twitter maps there is an extra cost, which requires discussion with the CartoDB sales team. This makes this solution not accessible for most journalists and newsrooms.

In this paper we focus on making Twitter conversation visualisations possible for smaller events, while identifying the highest possible number of associated locations in a dataset. For this, we developed a set of algorithms for tweet location detection. This work is part of a larger set of work on Social Semantic Journalism [5] and Location Based Event Detection [6].

We have further presented the steps taken for visualisation of twitter conversation during Ireland-Scotland Euro Qualifiers football match, which included tweet collection, geo referencing, data cleaning, data correction and data visualisation.

In the final part of the paper we have made a small comparison between the results from Twiloc's geo referencing and CartoDB Twitter Map geo referencing. Further evaluation of the quality of results, however, are yet to be conducted.

2. TWEET LOCATION DETECTION

Tweets may include multiple locations within its text and metadata; the place where the tweet was tweeted from, places mentioned within the tweet text, user profile and user network information. This paper proposes a method for identifying location information in tweets, which the use of the following features for Tweet location identification: *(1) GPS information*, *(2) User profile metadata, (3) Entity Extraction and Natural Language Processing techniques on tweet text and user bio information,* and *(4) Social Network Analysis*.

GPS information

This can be used for tweets which are tagged with GPS coordinates. It is a simple and straightforward location identification approach, and can give the exact location on a map where the tweet was published. However, only a small fraction of tweets (~1%) include GPS coordinates [7].

User profile metadata

This information users include in their profile, including language, time-zone and profile location. Most users would provide relevant information in these fields. This information can be used to identify locations associated with the user, and not specifically the tweet itself.

Entity extraction and NLP techniques for Tweet text and user bio information

These techniques are to extract relevant location information from: (a) tweet text and (b) user-specified profile information and location – in their bio, explained in the following:

Tweet Text

Tweet text contains the relevant information describing the event. It contains up to 140 characters and may contain links to images, videos, sound, etc. The potential locations and places mentioned within the text of a tweet are likely to be about the tweet/event under discussion, and could provide relevant location information if extracted and disambiguated appropriately.

User specified bio information

This is the information users include in their profile bio, which often includes bio text and bio location. Similar to user profile information, users normally provide relevant information in their profile bio and location. However, as they are free fields and twitter does not validate them, they may include information such as 'Mars' or 'home'.

In order to identify relevant information that describe places within the user profile metadata and tweet text, entity extraction and Natural Language Processing (NLP) techniques are used. NLP techniques assist in observing events and sentiments, extraction information such as variety of entities and tagging them. In order to extract the location for an event from the user-generated content, the textual data is processed through NLP techniques to determine the entities and their context with respect to parts of speech (POS). Named Entity Recognition (NER) as part of Information Extraction aims to identify and classify text into multiple predefined categories, such as persons, organisations and places [8]. The importance of NLP techniques to identify named entities from Twitter stream data has increased. Multiple works [9, 10] are applying NLP techniques to identify named entities and determine the event location along with user location.

In this work the Stanford Named Entity Recogniser -- part of the Stanford CoreNLP Natural Language Processing Toolkit -- is used to identify entities that describe people, places and organisations [11]. In order to disambiguate locations and to get more detailed information about the extracted entities such as country, city, and the geo-coordinates, the extracted entities are linked to multiple knowledge bases such as DBpedia[12] and GeoNames[13].

User Social Network Analysis

A user's social network plays an important role in determining the user's location. Often when the content-based approaches (geo-tagged data, user profile location) fail to determine the location of a user, it is the user's social network that can help in understanding from where the user is posting the content. This method leverages a user's social relationships and the spatial distribution of locations in her/his network for identification of potential locations [6].

Using social networks to identify user location is implemented as part of Insight News Lab's work on tweet Location Detection, however, after experimenting with various datasets and factors we decided to leave this feature out for the purpose of basic Twitter location detection, and for the work presented in this paper. This approach is slower than the other approaches and gives indication of the network of the user, as opposed to the location of the tweet and the user. There is a likelihood that the location with highest frequency might be the same location as the user's location, but the computational overhead this approach adds to the system makes it less suitable for the location detection from the Twitter stream in near real-time, as for each single tweet the network of the sender would need to be computed. This feature is however tremendously useful when authenticity of a user for posting about events in a specific location is under question.

The next section introduces Twiloc and the proposed framework that leverages the aforementioned techniques for inferring the location of a tweet.

3. TWILOC

Twiloc is a Tweet Location Detection engine, designed and developed at the Insight News Lab. It employs the approaches explained in Section 2 and results in a high degree of location identification for Twitter datasets, which ultimately enables tweet mapping and visualisation for smaller events. Figure 1 depicts the flowchart for Insight News Lab's tweet location detector – Twiloc. This process is repeated for each tweet.

For each tweet multiple locations could be found within the tweet text, GPS information and user profile. When choosing a location for mapping, Twiloc gives the highest priority for the place where the tweet is posted from, then the location mentioned in the tweet text, and then finally locations related to the user profile. Each tweet is annotated with a new field that details location information and geo-coordinates for tweet mapping. If the tweet has GPS information then that information is used as location information since the GPS information are the most accurate for tweet location. The second most important piece of information in our scenario is a location mentioned in the text of a tweet. If we did not find this information, we then proceed to user profile location as a best alternate indicator for tweet location, after GPS and location mentioned in the tweet text.

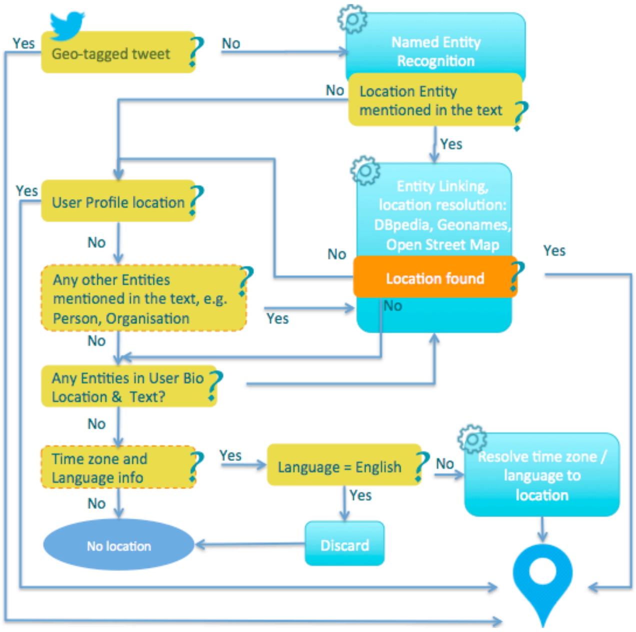
Yes
Geo-tagged tweet ?
No
Named Entity Recognition
No
Location Entity mentioned in the text ?
Yes
Yes
User Profile location ?
No
Entity Linking, location resolution: DBpedia, Geonames, Open Street Map
No
Yes
Any other Entities mentioned in the text, e.g. Person, Organisation ?
Location found ?
Yes
No
No
Any Entities in User Bio Location & Text?
No
Time zone and Language info ?
Yes
Language = English ?
No
Resolve time zone / language to location
Yes
No
No location
Discard

Figure 1. Twiloc: A Tweet Location Detection Framework

To determine the location mention in the tweet text and free fields in user profile we use Natural Language Processing techniques. For this we extract the Named Entities from the tweet text. These include: Location, Organisation and Person mentions. Location entities are the most straightforward and most important in our scenario. However, if we did not find a location entity in the text, followed up by lack of relevant user profile location information, we may decide to turn into Organisation and People type entities. For these we use further knowledgebase lookups, such as DBpedia, to find the location of an organisation, or a location associated to a Person. We consider this step as optional, since it adds to processing time and in some scenarios may not be considered as relevant enough. In the experiment presented in this paper this step was excluded.

4. METHOD AND RESULTS

This paper presents the tweet collection, annotation and visualisation process used for visualising the Ireland – Scotland EU 2016 Qualifiers football match.

For tweet collection we used the following hashtags, most of which were trending during the game in Ireland, Scotland or both:

#coybig, #scoirl, #scovroi, #ScotlandvIreland, #wearescotland, #comeonscotland, #tartanarmy, #irevsco

For data collection we used tweet collection tools developed at the Insight News Lab[1]. Twiloc was used for tweet geo referencing.

CartoDB was used for visualisation of the tweets on a map. As mentioned earlier CartoDB also provides tweet collection and location annotation tools – Twitter Maps, which is only available as part of their Enterprise plan and is costly. At the end of this section we have made a brief comparison between our results and CartoDB location tagging results.

The total number of tweets collected during the Ireland-Scotland match, including 5 minutes before and after (19:40 – 21:44 on the 14th Nov 2014) was 22,957 tweets. Out of these only 1,055 tweets were geo-tagged by users (users who had their location information turned on when sharing a tweet), which give us a 4.5% of all collected tweets – higher than the usual 1% suggested in the literature. Using Insight News Lab's Twiloc, we geo-tagged 16,008 tweets out of the 22,957, which means 70% of our tweets were geo-tagged. Table 1 present a summary of results for Ireland-Scotland EU Qualifiers match.

Table 1. Ireland vs Scotland EU Qualifiers football match tweet stats

Total tweets collected	**Including GPS coordinates**	**Geo-tagged with Twiloc**
22,957	1,055 4.5%	16,008 70%

Figure 2 depict the moment of the only goal in the match for only user geo-tagged tweets (fig. 2 .A) and tweets geo-tagged by Twiloc. The figure depict the moment of the only goal in the match, which presented a burst in tweets posted and the highest number of tweets present in a small time period. These figures show how different the two maps could and would look like at any moment, depending on the number of tweets geo-referenced, and particularly in the least and most exciting moments of the match. You can visit and compare the two interactive maps from the URLs provided.

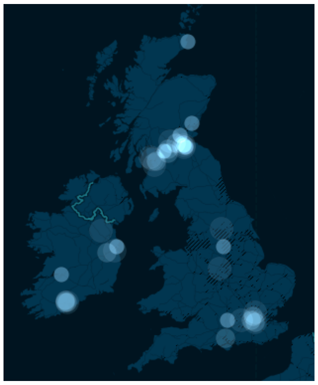

Figure 2. A. Visualisation of tweets - only user geo-tagged - the moment of the only goal of the game (21:19 14 Von 14) http://bit.ly/1Jjh0Lb

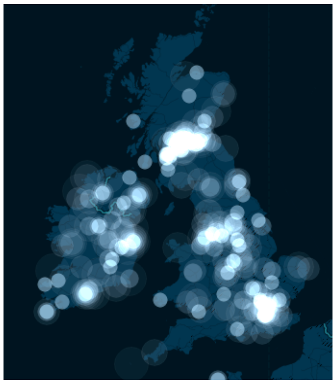

Figure 2. B. Visualisation of tweets – Geo tagged by Twiloc - the moment of the only goal of the game (21:19 14 Von 14) http://bit.ly/1IOzUGm

As mentioned above we used CartoDB for data visualisation. For this we first used lat-long feature in CartoDB for tweets, which we had an exact location information for, i.e. the ones with GPS coordinates. The second round was to run CartoDB geo-tagging based on our extracted city and county names. CatroDB was not able to resolve some of the locations in our dataset. Examples of such locations are Irish or Scottish counties, where the name of the county is different to the county city/town name.

[1] Newslab.insight-centre.org

To remedy this, a set of rules was defined to replace the county names with their county city/town name. We used OpenRefine for these data transformations. Another transformation we needed to perform was to transform the locations, with only country name, to an associated city in the country. This is because otherwise the centre point of the country would have been considered for the location, which in many cases is not the best representative of where the tweets might have been sent. In this case we replaced the geo coordinates of the centre point of the countries with geo-coordinates of their associated capital city. This may not be the best representative of the tweets with only country information extracted from them, as they might have been sent from other parts of the country, but we believed capital is a much better representative than the centre point of the country, which in many cases (at least in Ireland) may be in the middle of some fields, which may not even be close to any village, town or city. There is higher chance that the tweets are sent from the capital city, than from the middle of a field or on the motorway.

After we had the data geo-referenced and cleaned up, we used the free version of CartoDB for our data visualisation. The visualisation can be found on http://bit.ly/1IOzUGm and a story on this published in the *Irish Times* on http://bit.ly/1DQJhX6.

EU Qualifiers football matches are an example of a small but significant event. We further investigated a relatively small local event by collecting tweets for the Dublin Marathon, which is a popular event in Dublin. We used the Dublin Marathon dataset to compare our results with results from geo-referencing with the CartoDB tweet map. CartoDB provides a trial of 10,000 tweet collection if you contact them to ask about the Tweet Map service. We used this to collect data the Dublin Marathon, knowing that the tweet count would likely be manageable within CartoDB Tweet Map's trial allowance we were given.

4.1 Comparison with CartoDB

Using CartoDB trial version of Tweet Maps, we collected just over 8,000 tweets for the Dublin Marathon between 26 to 28 October 2014. Out of these only 164 (2%) tweets were Geo-tagged, i.e. had GPS information, which was a lower rate when compared with the Ireland-Scotland football match, but closer to the figures suggested in the literature. We initially used CartoDB Tweet Map service for geo-tagging and visualisation of the Dublin Marathon event. CartoDB, using data from GNIP, geo-tagged 4,814 of tweets (60%). We then geo referenced the same data with Insight Insight News Lab's Twiloc. Twiloc geo-tagged 5,320 of tweets (66.5%). This allowed us to compare Twiloc geo referencing with CartoDB's. The summary of data and results for Dublin Marathon 2014 are presented in Table 2.

Table 2. Dublin Marathon tweet stats, Twiloc compared with CartoDB

Total tweets collected	Including GPS coordinates	Geo-tagged with CartDB	Geo-tagged with Twiloc
8,007	164 2%	4,814 60%	5,320 66.5%

5. CONCLUSIONS

Mapping Twitter conversation on a map over time has become a popular way of visualising conversations around events on Twitter. A straightforward approach for visualising tweets on a map is using GPS location information from the geo-tagged tweets. This information, however, is only present in around 1-5% of tweets, which makes for not so interesting visualisations of smaller events, or events in smaller countries where there are fewer people to tweet. This paper presents a Tweet Location Detection approach, which uses various features in a tweet for detecting the best possible location for a tweet. We used this as a part of data journalism work for mapping the twitter conversation around the Ireland-Scotland Euro Qualifiers game. Twiloc resulted in 70% location geo referencing for this dataset, as opposed to the 4.5% originally geo-tagged tweets (by users). We further used Twiloc for geo-tagging Dublin Marathon tweets and compared our results with the results we got from using CartoDB Tweet Maps for the same event. Twiloc resulted in slightly higher geo referenced tweets in comparison to CartoDB, 66.5% vs 60% respectively.

Overall Twiloc shows promising results for location detection and geo tagging tweets on the datasets presented in this paper. However, further testing and evaluation of results for determining the quality of detected locations is required in the next stages of our work.

6. ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Numbers 12/TIDA/I2389, SFI/12/RC/2289 and 13/TIDA/I2743. We would like to acknowledge the work of other researchers at the Insight News Lab in development of Twitter Location Detector: Ravindra Harige and Prashant Khare.

7. REFERENCES

[1] CartoDB Tweet Map. Retrieved August 14, 2015, from https://cartodb.com/solutions/twitter-maps

[2] Moffit, J. 2014. Twitter Geo-Referencing: An Example Use-case. (March 21, 2014). Retrieved August 14, 2015, from http://support.gnip.com/articles/twitter-geo-referencing.html

[3] Cairns, I. 2014. Get More Twitter Geodata From Gnip With Our New Profile Geo Enrichment. (Apr 29, 2014). Retrieved August 14, 2015, from https://blog.gnip.com/twitter-geo-data-enrichment/

[4] CartoDB Enterprise Plan. Retrieved August 14, 2015, from https://cartodb.com/pricing/#enterprise.

[5] Heravi, B. R., McGinnis, J. Introducing Social Semantic Journalism, *The Journal of Media Innovations: Special issue on Innovations in the Newsroom*. 2015.

[6] Heravi B. R., Morrison, D., Khare, P., Marchand-Maillet, S. Where is the News Breaking? Towards a Locaiton-based Event Detection Framework for Journalists, *Lecture Notes in Computer Science, 8326*, 2014, 192-204.

[7] Jurgens, D. 2013. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. In *Seventh International AAAI Conference on Weblogs and Social Media 2013*

[8] Finkel, J. R., Grenager, T., & Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics,* (2005) , Association for Computational Linguistics, 363-370

[9] Ritter, A., Clark, S., & Etzioni, O. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2011*. Association for Computational Linguistics, 1524-1534

[10] Unankard, S., Li, X., & Sharaf, M. A. Location-Based Emerging Event Detection in Social Networks. *Web Technologies and Applications*, Springer, Berlin Heidelberg, 2013, 280-291

[11] Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55-60.

[12] Pablo N. Mendes, Max Jakob and Christian Bizer. 2012. DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. In *Proceedings of the International Conference on Language Resources and Evaluation* (Istanbul, Turkey, May 2012). LREC 2012. 21-27

[13] GeoNames. Retrieved August 14, 2015, from http://geonames.org/