Homing in on Twitter Users: Evaluating an Enhanced Geoparser for User Profile Locations

Beatrice Alex, Clare Llewellyn, Claire Grover, Jon Oberlander and Richard Tobin
School of Informatics, University of Edinburgh, UK
{balex|llewellyn|grover|jon|richard}@inf.ed.ac.uk

Abstract

Twitter-related studies often need to geo-locate Tweets or Twitter users, identifying their real-world geographic locations. As tweet-level geotagging remains rare, most prior work exploited tweet content, timezone and network information to inform geolocation, or else relied on off-the-shelf tools to geolocate users from location information in their user profiles. However, such user location metadata is not consistently structured, causing such tools to fail regularly, especially if a string contains multiple locations, or if locations are very fine-grained. We argue that user profile location (UPL) and tweet location need to be treated as distinct types of information from which differing inferences can be drawn. Here, we apply geoparsing to UPLs, and demonstrate how task performance can be improved by adapting our Edinburgh Geoparser, which was originally developed for processing English text. We present a detailed evaluation method and results, including inter-coder agreement. We demonstrate that the optimised geoparser can effectively extract and geo-reference multiple locations at different levels of granularity with an F1-score of around 0.90. We also illustrate how geoparsed UPLs can be exploited for international information trade studies and country-level sentiment analysis.

Keywords: Geo-parsing, Twitter, Social Media Analytics

1. Introduction

Studies taking advantage of Twitter data often need to geolocate Tweets or Twitter users, identifying their real-world geographic locations. Geolocation underpins analyses ranging from sentiment analysis towards particular topics or events or the tracking of disasters and epidemics, through to the mapping of breaking news stories. When analysing tweets one might want to know where users tweeted from (useful for disaster management) or where they usually live or originated from (useful for comparing groups of people's attitudes and beliefs). Tweet-level geotag locations are encoded as latitude/longitude (lat/long) coordinates for a very small percentage of tweets made available by Twitter (see Table 1). Many existing geography-based Twitter visualisations are therefore limited to this highly biased subset of data. However, tweet location is not a good proxy for a person's home location and can significantly distort the results of any study or visualisation which tries to capture information for different countries or regions. By contrast, UPL is theoretically a much better approximation for home location and is much more frequently specified. In this paper, we focus on the geoparsing of UPLs to enable studies of regional differences between Twitter users and their attitudes.

UPL is a relatively static piece of information which can be optionally specified by a user when creating their account by typing into a free form text field. The Twitter documentation states that this field is: *The user-defined location for this account's profile. Not necessarily a location nor parseable.* The data is entered in textual form, and places are not geolocated. Hecht et al. (2011) present an analysis of Twitter UPLs. They show that in their data (collected in 2010), after having performed automatic language identification to identify English tweets, 66% of 10,000 randomly sampled profiles contained a genuine UPL, mostly at the city level. 18% of users left this field blank and 16% specified non-geographic information. Moreover, users are not required to keep this field up-to-date. Still, while it has its limitations, the percentage of users who do specify valid locations in their profile is much larger than the proportion who reveal their individual tweet locations. We therefore argue that correctly geoparsed UPLs yield a much more useful approximation of users' home locations than their individual tweet locations. In Section 3.4 we demonstrate this by distance measures.

Previous work has geolocated UPLs by using off-the-shelf tools (see Section 2). These methods can work fairly reliably on one-word tokens which exactly match a location name, but tend not to perform reliably for more complex notations. We describe how we adapted an existing Edinburgh Geoparser, originally developed for processing running English text, to geoparse UPLs (Section 3). As no gold standard existed, we performed the evaluation manually and determined inter-coder agreement. Our evaluation method considers multiple locations per UPL at different levels of granularity (if specified). Finally, we present information trade visualisations and sentiment analysis (Section 4) in which we treat the geoparsed UPL output as a proxy for home location(s) of users.

2. Background and Related Work

A review of Twitter geolocation algorithms and evaluation methods can be found in Ajao et al. (2015). Most related work attempts to geolocate users based on tweet-level information (Cheng et al., 2010; Eisenstein et al., 2010; Hecht et al., 2011; Kinsella et al., 2011; Chang et al., 2012; Han et al., 2014; Mahmud et al., 2014) or using network analysis (Java et al., 2007; Takhteyev et al., 2012; Ryoo and Moon, 2014; Compton et al., 2014). The ultimate goal is to locate all users based on information present in their tweets or via connections to other users. The main motivation is that only a small percentage of tweets are geolocated and while a majority of users do specify their UPL that information is often not consistently structured. Han et al. (2014) show that user profile meta data (particularly location but

also timezone) can significantly improve the geolocation of users. Their UPL-based feature is computed by taking 4-grams of the specified string. As some user profiles contain ambiguous and/or multiple locations, we would expect that an additional feature based on high-accuracy geo-parsing of UPL information could improve their models even further. Osborne et al. (2014) also exploit UPL for their supervised machine learning method of geolocating individual tweets.

Java et al. (2007) report that of the 76,000 analysed user profiles in their dataset, only 39,000 could be resolved to lat/long coordinates using the Yahoo! Geocoding API. They did not report accuracy. Kulshrestha et al. (2012) evaluated country-level geolocation of UPLs by exploiting Yahoo! PlaceFinder, Bing and timezone information stored in user profiles. They report an accuracy of 94.7% for correctly geolocating the country for a subset of 1,000 UPLs for which two of the three geolocation methods agreed. They ignored UPLs for which all three methods disagreed which happened in 2.2% of specified UPLs.

Graham et al. (2014) applied Yahoo's PlaceFinder and Google's Geocoder to 4,000 UPLs of people tweeting from four cities (Cairo, Montreal, San Diego, Tokyo). Geolocation is considered correct if it falls within the bounding boxes of the cities. Overall Yahoo! PlaceFinder was able to geolocate more UPLs than Google's Geocoder but the latter geolocated more correctly with an accuracy of 54.5% for UPLs which were not left blank.

Methods for creating a gold standard vary considerably. Some studies use the first or most frequent lat/long pair of a user's geolocated tweets (Eisenstein et al., 2010; Kinsella et al., 2011). Han et al. (2014) use a city-based gold standard by extracting cities from GeoNames[1] and identifying a city per user as the one with the largest number of geolocated tweets. Tweets that do not occur in close proximity to a city are ignored. Mahmud et al. (2014) collected Geonames bounding boxes for 100 US cities and use the geotag information for tweets within them as the ground truth. Cheng et al. (2010) and Ryoo and Moon (2014) evaluate their method against a subset of users with lat/long coordinates specified in the UPL. They place 56.7% of Korean Twitter users within 10 km of their main location. Others evaluate against automatically generated geolocation of UPLs, using Yahoo, Bing or exact match against Wikipedia titles with associated lat/long pairs (Hecht et al., 2011) which are rarely evaluated for accuracy. Kinsella et al. (2011) use Yahoo! Placemaker geolocation of UPLs as a baseline to compare their language modelling approach to user location prediction at different levels of granularity. This baseline performs best for states (47.1% acc.) and similarly to their best method for cities (31.4% acc.).

Most of this work is evaluated using distance-based accuracy (within 100 miles from gold is often considered correct) which is to account for commuters. However, a lot of people also tweet when they are travelling further afield, in which case it would be very difficult to identify their primary "home location" correctly. A distance of up to 100 miles (160.9km) is fairly large. Two users automatically geolocated 321.8km apart from each other could therefore be considered correctly geolocated to the location exactly in-between them. For certain regional geographical studies one would like to be confident that user-level geolocation is accurate at much smaller distances, for example when trying to determine sentiment of inhabitants of Glasgow and Edinburgh towards Scottish independence.

[1] http://www.geonames.org/

3. Geoparsing User Profile Locations

3.1. Twitter Data

All of our datasets are based on the 1% Twitter API stream and limited to English tweets. As the work presented in this paper was for the UK CONNECTIVITY project to identify influence of the UK and other countries as well as attitudes of inhabitants of counties towards particular subjects, we created subsets for four topics or events (UKRAINE, SYRIA, GLASGOW 2014 and CITIES) and different time-periods. Exact details of the method used for creating the datasets can be found in Llewellyn et al. (2015) [2] and a summary description is provided here:

UKRAINE: The Ukraine dataset contains Twitter data gathered across three time periods: 6-12 Mar 2014, when the EU held an emergency Heads of State meeting in response to events in Ukraine; 20-27 Jun 2014, when there was troop build up in Ukraine, and a peace plan was put in place; and 17-23 Jul 2014, when flight MH17 was destroyed in Ukrainian airspace. This data was created by doing an initial case-insensitive grep for 'Ukraine' or 'Ukrain' in all tweets collected in the time periods, pulling out the hashtags mentioned within them and expanding the set by adding the most frequent hashtags that appeared to be relevant to the Ukraine to the grep expression.

SYRIA: For the Syria dataset we selected two contrasting weeks of Twitter activity. The first week, 1-8 Mar 2012, was a relatively non-eventful week, although there was a UK-related event in Syria with the closing of the UK embassy in Damascus. The second week, however, 29 Aug - 4 Sep 2013, was a week of intense debate on Syria as the UK Parliament voted not to authorise military action over chemical weapons. We followed a similar approach as for creating the Ukraine data by case-insensitive grepping for 'Syria' and then expanding the grep using most frequently related hashtags. To check that the manual step of selecting related hashtags is something that can be done reliably, we asked two coders to go through the most frequent hashtags (208 for both weeks in total). Their agreement on deciding if a hashtag is related to Syria was perfect (Kappa 1.0). Both the Ukraine and Syria data were provided to us by the REDITES project group (Osborne et al., 2014).

GLASGOW 2014: This dataset is centred around the Glasgow 2014 Commonwealth Games. It was created by selecting English tweets collected between 21-25 Jul 2014, the two days leading up to the day of the opening ceremony and the two days following it. We then searched case-insensitively for tweets containing hashtags starting with

[2] All data, including the gold standard, are available **http://groups.inf.ed.ac.uk/UKConnect/publications.html**.

Dataset	Total tweets	Geotagged tweets	Non-empty UPLs	Geo-resolved UPLs
CITIES	467,893	10,371 (2.2%)	321,540	256,633 (79.8%)
UKRAINE	79,374	1,063 (1.3%)	55,700	43,781 (78.6%)
GLASGOW 2014	5,853	191 (3.3%)	4,513	4,041 (89.5%)
SYRIA	26,393	156 (0.6%)	19,315	15,685 (81.2%)
ALL	579,513	18,781 (3.2%)	401,068	320,140 (79.8%)

Table 1: Number of all tweets, geotagged tweets, non-empty UPLs and UPLs which were geo-resolved to at least one location using version 2 of the Edinburgh Geoparser. Percentages relate to the previous column.

at least one of the following three strings: #bbcglasgow, #glasgow or #commonwealth.

CITIES: This dataset was created by monitoring Twitter activity over 12 weeks (19 May - 10 Aug 2014) with approximately 1.4 millions tweets per day and limiting it to tweets containing the names of a series of UK cities (Belfast, Birmingham, Cambridge, Edinburgh, Glasgow, Liverpool, London, Manchester and Oxford) as well as Paris.

Table 1 lists various counts for each dataset which only contain 3.2% of geotagged tweets in total. Many users specified a UPL, a large majority of which we were able to geo-reference to one or more locations. While we created each of these datasets based on tweet content, we focus only on the UPL provided with each tweet (if specified) for the geoparsing experiments.

3.2. Geoparser

The Edinburgh Geoparser was developed to geolocate place names found in regular running text (Grover et al., 2010). The standard implementation[3] is not designed to process Twitter location fields; therefore we had to adapt it to geolocate Twitter users. After assessing some of the shortcomings of the first Twitter-adapted version, we created a second improved version. Both versions are described below but it is helpful first to explain briefly how the regular Edinburgh Geoparser works. Its standard geolocation process involves the following five steps:

1. **Identify place names using named entity recognition (NER).** Our NER system is rule-based and uses lexicons of known place names in combination with contextual features to mark up place name mentions in text. If there is a clear indication in the text of two place names being in a containment relation (e.g. "They moved to London in Ontario."), 'contains' and 'contained-by' attributes are added to the markup to aid disambiguation between possible interpretations.

2. **Create a set of place name queries.** The identified place names are converted into gazetteer queries.

3. **Query the gazetteer and store the results as a set of records per place name.** The Edinburgh Geoparser can access a number of different gazetteers, but the relevant one for this work is GeoNames, as this has world coverage.

4. **Rank records.** Where there are multiple records for a place name, rank them to discover the most plausible interpretation in context.

5. **Add to markup.** Add the geolocation information from the highest ranked record into the place name markup of the text file.

The ranking algorithm creates a score per record using weights for certain pieces of information. The main parameters for this are:

POPULATION: a place with a larger population is weighted more highly.

TYPE: each record has an associated type and certain types are weighted more highly than others. For example, type="country" and type="civila" (a civil administrative unit) are the most highly weighted, followed by type="ppl" (populated place), while type="fac" (facility) and type="road" are not highly weighted.

CLUSTERINESS: we assume that a textual document generally has a degree of geographic coherence to it so that, for example, the interpretation of a mention of "Paris" in a text will be influenced by whether the text also mentions "Texas" or "France". We model this by assigning a clusteriness score to each candidate reflecting how close it is to all the other place name interpretations in the text.

CONTAINMENT: if the NER markup supplied 'contains' and 'contained-by' attributes, extra weighting will be given to records in close proximity for place names with these attributes.

LOCALITY: if the user knows what geographic area the text is about, this information can be supplied to weight places within that area more highly.

3.2.1. Twitter Geoparser: Version 1

As UPL strings are not running text, the NER step (step 1 above) would perform unreliably when processing Twitter UPLs. In the first adaptation to Twitter data, we therefore ran the Edinburgh Geoparser with UPLs directly as gazetteer queries but removing full stops. We discarded any UPL without alphabetic characters and treated each UPL without commas as one query (see Figure 1). UPLs with commas were split into two queries with attributes to indicate that the first is contained by the second (see Figure 2). If there are multiple commas, we only create two queries from the first two parts of the split.

Step 3 from the regular Edinburgh Geoparser was unchanged. We looked up the place name queries in a local copy of GeoNames. Despite the preprocessing, a number of

[3]It is available as a download and demo at: `https://www.ltg.ed.ac.uk/software/geoparser/`

```
location: +254                is discarded
location: 86000, Kluang #MY   becomes <placename name="86000, Kluang #MY"/>
location: Au/NZ/US            becomes <placename name="Au/NZ/US"/>
location: London &amp; Global becomes <placename name="London &amp; Global"/>
location: Melbourne           becomes <placename name="Melbourne"/>
location: Drumchapel/Glasgow  becomes <placename name="Drumchapel/Glasgow"/>
location: Scotland... DUH     becomes <placename name="Scotland DUH"/>
location: At home.            becomes <placename name="At home"/>
```

Figure 1: Place name queries extracted from UPLs for version 1 of the Edinburgh Geoparser.

```
From location: North Cornwall, Britain...  we create:
<placename name="North Cornwall" contained-by="22" id="21"/>
<placename name="Britain" contains="21" id="22"/>

From location: The Cage, Hell  we create:
<placename name="The Cage" contained-by="510" id="509"/>
<placename name="Hell" contains="509" id="510"/>

From location: Perth/Adelaide, Australia  we create:
<placename name="Perth/Adelaide" contained-by="291" id="290"/>
<placename name="Australia" contains="290" id="291"/>
```

Figure 2: Place name queries for UPLs containing commas (step 2 of version 1 of the Edinburgh Geoparser).

queries containing genuine place names did not match any gazetteer entry due to other punctuation being present in the UPLs. For example, lookup failed for queries like "86000, Kluang #MY", "Au/NZ/US", "Perth/Adelaide" but succeeded for "Melbourne", "North Cornwall" or "Britain". For UPLs which are probably not meant as actual place names, lookup usually failed (e.g. for "Worldwide" or "Behind you"), but sometimes it succeeded (e.g. for "Hell", "The Shire" or "Saturn").

In some cases vernacular terms, shorthands, hashtags, or witticisms are used to describe locations; for example, "Brum", referring to Birmingham, matched a place in Serbia but lookup failed for "Sheffieldish", probably referring to a location in or around Sheffield. Successful lookup does not guarantee that any of the gazetteer records returned are actually the correct record for that place name. However, the geo-resolution stage will always choose one of the records, so for example "Saturn" will be geolocated when it should not be and "Brum" will be incorrectly resolved.

The resolution step of version 1 (step 4) works the same as for the standard geoparser with two small changes. Each Twitter UPL can only be resolved in isolation, so measures pertaining to document coherence must be avoided. We therefore excluded the geographic coherence and locality parameters from the ranking scoring which means that country, population and type are the most relevant parameters considered. Information about containment is also used effectively. We saw above that a comma is taken to indicate a containment relation, and the contains/contained-by weightings in combination with population and type ensure better disambiguation.

The final step (a modification of step 5) adds the highest ranking results back into the input file. Only one resolved place name was used per UPL. In the case of multiple resolved place names, the first one was used (see Table 2).

3.2.2. Twitter Geoparser: Version 2

After performing evaluation (see Section 3.3) and error analysis, we created version 2 of the adapted Edinburgh Geoparser to address some of the shortcomings of version 1. For example, there are cases where commas are used for something other than containment ("US, Canada and UK"), where other punctuation is used to indicate containment ("Dalston - London - Earth") and where punctuation between places is omitted ("Runcorn Cheshire UK"). We therefore attempted to geolocate each distinct place name in the UPL.

To identify multiple place names we included a cut-down version of the regular NER component looking up strings in the NER location lexicons without relying on context rules. Before lookup we first split each UPL into subparts using punctuation (split at [,;:/!]). For example, in "Glasgow, UK via Denmark" we mark up the entities "Glasgow", "UK" and "Denmark" as well as the putative entity "UK via Denmark".

The gazetteer querying and resolution steps are the same as in version 1. In the final step, information from the top ranked interpretation of each resolved place name is added back into the UPL. While version 1 of the geoparser identifies only one location per tweet, version 2 was set up to identify up to six possible locations within one UPL field. This results in correct interpretations of the three genuine place names for the example above (see Figure 3).

UPL	Geoparser output [lat,lng; country]
Melbourne	Melbourne[-37.81,144.96; AU]
Brum	~~Brum[43.65,21.20; RS]~~
Glasgow, Scotland	Glasgow[55.87,-4.26; GB]
Perth/Adelaide, Australia	Australia[-25.0,135.0; AU]

Table 2: Geoparser (version 1) output for different types of UPLs. Incorrect resolutions are crossed out.

```
<location>
 <geoinfo country="GB">55.8651500,-4.2576300,GB,Glasgow</geoinfo>
 <geoinfo country="GB">54.7584400,-2.6953100,GB,UK</geoinfo>
 <geoinfo country="DK">56.0000000,10.0000000,DK,Denmark</geoinfo>
 <loctext>Glasgow, UK via Denmark</loctext>
</location>
```

Figure 3: Top-ranked place name interpretations.

3.3. Evaluation Data and Method

The overall aims of this evaluation were to determine: (i) if the adapted Edinburgh Geoparser can correctly identify locations within UPLs; and (ii) if it can accurately resolve them to lat/long coordinates and country codes.

The evaluation was conducted in two phases using the two different versions of the Edinburgh Geoparser. Our gold standard dataset is made up of UPLs along with one explicitly geotagged tweet per user, i.e. a tweet with automatically created lat/long coordinates. While this location is often distinct from the UPL, it may give a reasonable indication of the user's home location and helped during manual evaluation. This gold dataset may be subject to a sampling bias as users who permit automatic geolocation of their tweets may be less likely to obfuscate their UPL. We therefore conducted an additional manual evaluation on a small random sample (see Section 3.5).

Our data selection method resulted in the sets of UPLs, each associated with a different topic of interest. We randomly

selected a subpart of each data set for manual evaluation (see Table 3) with a user only being included once per set.[4] These gold sets were manually coded by a single coder and 1% of the larger dataset was double coded to assess inter-coder reliability.

Dataset	UPLs in set	... and in gold
CITIES	7,494	825
UKRAINE	711	71
GLASGOW 2014	156	14
SYRIA	122	12
TOTAL	8,483	922

Table 3: Number of UPLs in each subset. Subsets are collected by only considering unique users (in individual sets) tweeting with a UPL and geo-tagged tweet coordinates.

Our evaluation of the performance of the Edinburgh Geoparser aims to determine the accuracy of multiple criteria of its output. Manual inter-coding and geoparsing evaluation adhered to the same set of criteria in the following order taking a value of 0 (false) or 1 (true):

is place: It is possible for Twitter users to specify multiple locations in their UPL. For example, the evaluation data contains up to four geo-referenced locations per UPL. So firstly, the evaluators assigned binary 'is place' scores to each potential place name in the UPL.

is formal place: Some users entered slang or colloquial terms as their UPL, for example "Brisneyland" for Brisbane. Such vernacular place names are more difficult to identify than official ones as they are unlikely to be contained in gazetteers. To be able to make this distinction, the evaluators specified whether the place name was a formal location with defined boundaries.

is resolved: They also determined whether the location was resolved by the Edinburgh Geoparser.

is resolved correctly: Finally, they checked if the geo-resolution was correct.

If a UPL did not contain an actual place name or if the Edinburgh Geoparser resolved a string which is not a place name, then the values for "is place" and "is formal place" were both set to 0. In such cases, the values for "is resolved" or "is resolved correctly" depended on whether the Edinburgh Geoparser had incorrectly attempted geo-resolution.

To provide some coding examples, Table 4 presents three UPLs ("North Belfast, Ireland", "Brent Cross, London" and "The Wall") and version 1 geoparsed output. "North Belfast" and "London" are not geo-referenced, but "Ireland" and "Brent Cross" are. Therefore, "North Belfast" was recorded as a formal place not geo-referenced (coded as 1,1,0,0), "Ireland" as a formal place geo-referenced correctly (1,1,1,1), "Brent Cross" as a formal place geo-referenced correctly (1,1,1,1) and "London" as a formal place not geo-referenced (1,1,0,0). In the third example, the user specified "The Wall" as their profile location, which the Geoparser geogrounded incorrectly to an actual place with that name in the mountains of New Mexico, USA (0,0,1,0).

To establish accuracy of the geo-resolution, the coders first checked if the lat/long coordinates were very similar to the automatically assigned tweet coordinates. If this was the case and information in the tweet text did not contradict this assumption, then the geo-resolution was marked as correct. If there was a mismatch, then the evaluators used a mapping service such as Google Maps to check the distance between the locations and determined if a location of the same name was in closer proximity to the provided location. The evaluators also used the tweet text as a guide, because people often explicitly state when they are away from home. In the case of vernacular place names, the evaluators consulted the Urban Dictionary to identify to corresponding formal name.[5]

3.4. Results

We compare the performance of both versions of the geoparser against a baseline which was created by running the Google Maps Geocoder[6] over the same data and selecting the top-ranked result. Table 5 presents the number of decisions made for each criterion per system output in the singly coded datasets. Our gold data comprising of 922 UPLs contains a total of 1,245 place names of which 1,202 are formal ones (96.5%).

Table 6 shows the corresponding performance measures in precision, recall and F-score. These evaluation metrics provide better understanding of erroneous and spurious output which has been lacking in previous work where the chosen evaluation metric was mostly distance-based accuracy.

The scores for formal places are slightly higher than those for all places. The Google Maps Geocoder baseline yields a 10% higher recall but a 11% lower precision than version 1 of the geoparser. Performance improves significantly between the two versions of the geoparser. For version 1, scores are very high in precision (0.95) but recall is low (0.46) because this version tends to query the gazetteer for the entire UPL string, which typically fails as soon as there is no exact match. Recall increases substantially for version 2 of the geoparser at a slight loss of precision. The number of correctly geo-referenced locations almost doubles for version 2. The improved geoparser performs with an F-score of 0.90 for all place names and 0.92 for formal place names. Our results illustrate that it is worth spending some time to adapt a geoparser to a new type of data.

The inter-coder agreement Kappa scores (Table 7) are almost perfect for most of the criteria except for "is formal place" for which agreement was substantial. This is mostly due to disagreements on the definition of formal place and questions on locations having defined boundaries. For the

[4] We selected around 10% per data set but slightly more for the CITIES data, slightly less for the GLASGOW data and 10.9% overall.

[5] During the coding process, the evaluators also identified gold geo-location information (identifier, lat/long pair, country code and name) for each place name in the UPLs in the gold standard. This data is available at: `http://groups.inf.ed.ac.uk/UKConnect/publications.html`

[6] `https://developers.google.com/maps/documentation/geocoding`

User profile location	Geoparser output	Coding
North Belfast,Ireland	Ireland[53,-8]; IE	North Belfast: 1,1,0,0; Ireland: 1,1,1,1
Brent Cross, London	Brent Cross[51.57715,-0.22433]; GB	Brent Cross: 1,1,1,1; London: 1,1,0,0
The Wall	The Wall[36.9419700,-105.1630600], US	The Wall: 0,0,1,0

Table 4: Geoparser version 1 output examples and coding.

	Gold standard		**Baseline**		**Geoparser: Version 1**		**Geoparser: Version 2**	
Dataset	Places	Formal places	Resolved	Res. correctly	Resolved	Res. correctly	Resolved	Res. correctly
CITIES	1,124	1,081	743	618	534	507	1,045	973
UKRAINE	82	81	55	52	48	46	78	74
GLASGOW 2014	20	20	13	10	9	9	18	18
SYRIA	19	17	10	10	6	6	15	15
ALL	1,245	1,202	821	690	597	568	1,156	1,080

Table 5: Frequencies of counts for all places and formal places per data set and overall as well as counts for the number of resolved and correctly resolved places for the baseline and both versions of the Edinburgh Geoparser.

All Places	**Baseline**			**Version 1**			**Version 2**		
Dataset	P	R	F1	P	R	F1	P	R	F1
CITIES	0.83	0.55	0.66	0.95	0.46	0.62	0.93	0.87	0.90
UKRAINE	0.95	0.63	0.76	0.96	0.56	0.71	0.95	0.90	0.93
GLASGOW 2014	0.77	0.50	0.61	1.00	0.45	0.62	1.00	0.90	0.95
SYRIA	1.00	0.53	0.69	1.00	0.32	0.48	1.00	0.79	0.88
ALL	0.84	0.56	0.67	0.95	0.46	0.62	0.93	0.87	0.90
Formal Places	**Baseline**			**Version 1**			**Version 2**		
Dataset	P	R	F1	P	R	F1	P	R	F1
CITIES	0.83	0. 57	0.68	0.95	0.47	0.63	0.93	0.90	0.91
UKRAINE	0.95	0.66	0.78	0.96	0.57	0.71	0.95	0.94	0.94
GLASGOW 2014	0.77	0.5	0.61	1.00	0.45	0.62	1.00	0.90	0.95
SYRIA	1.00	0.59	0.74	1.00	0.35	0.52	1.00	0.88	0.94
ALL	0.84	0.58	0.68	0.95	0.47	0.63	0.93	0.90	0.92

Table 6: Geoparsing performance measured in precision (P), recall (R) and balanced F-score (F1) for all and formal places.

double coding of the output of version 2, there was perfect agreement for "is resolved correctly".

Criterion evaluated	**V1**	**V2**
NUMBER OF LOCATIONS	0.9583	0.9581
IS PLACE	0.9483	0.9483
IS FORMAL PLACE	0.7697	0.7428
IS RESOLVED	0.8571	0.8426
IS RESOLVED CORRECTLY	0.9100	1.0000

Table 7: Inter-coder agreement Kappa scores for output version 1 and 2 of the Edinburgh Geoparser output.

We were also interested in determining how far the geo-referenced gold locations within the UPLs are from the geo-coordinates of their geotagged tweets. Table 8 presents the number of tweets and their mean and median Vincenty distances in kilometres to the geo-referenced gold location (choosing the closest if multiple were identified by the evaluators) in the UPL from very small to larger distance ranges. It is clear to us that distances would be increased

	GOLD STANDARD		
0-9	354 (42.9%)	3.7	3.1
10-99	215 (26.1%)	34.8	26.5
100-999	140 (17.0%)	345.3	314.4
1,000-9,999	108 (13.1%)	4,405.8	4,287.7
>= 10,000	7 (0.9%)	12,028	12,298.5
All	824 (100%)	749.0	17.2

Table 8: Mean and median Vincenty distances for different ranges (in km) between geotagged tweet coordinates and the nearest gold geo-location in the UPL of their authors.

for all users who have not kept their profile location up-to-date and moved elsewhere. Nevertheless, these figures show that tweet locations are often in close proximity to locations specified in the user profile but in 31.0% of cases there are distances of $\geq$100 km between them. This helps to supports our claim that tweet location should not be used as a proxy for the home location of a user.

3.5. Correction for Sample Bias

So far, our evaluation only included UPLs which also had an automatically assigned tweet location. This introduced potential bias into the sample, meaning that the gold standard may not be representative of the entire corpus. Furthermore, each of the topic or event specific datasets could be biased as a result of the data selection methods used.
We therefore repeated the evaluation using a small random sample of 100 non-empty UPLs selected from the 1% Twitter stream (collected between 12 May and 1 Oct 2014) and geoparsed them using version 2. The manual coding of this set was more difficult and time consuming. The evaluator used the same coding scheme as described earlier and identified 95 actual places of which 91 are formal as well as 87 geo-referenced locations of which 79 were correctly resolved. 56 of all UPLs contained at least one correctly geo-referenced place name, 38 contained no place name and 7 either contained a location which was geo-referenced incorrectly or contained no location but the geoparser identified and geo-referenced one.

Version 2	TP	FP	FN	P	R	F1
All Places	78	8	16	0.90	0.83	0.87
Formal Places	78	8	12	0.90	0.87	0.89

Table 9: Geoparsing performance for all and formal places when correcting for sample bias.

Table 9 shows that the geoparser performance is only slightly lower for the random sample than for the topic-specific datasets. However, the random sample could be increased in size to provide more robust scores, a shortcoming which we will address in future work.

4. Geo-specific Twitter Analysis

As noted earlier, geolocation underpins a number of different types of social media analyses which were investigated as part of the UK CONNECTIVITY project.

4.1. Information Trade

To illustrate the utility of the geoparser in context, consider the concept of cross-border information flow: extracting locations from user profile information enabled the study of international information trade. We define information trade as occurring when a tweet written by someone from country A is re-tweeted by someone from country B. In such a case, country A has exported information and country B had imported it. Other research has considered this phenomenon (Kulshrestha et al., 2012), but there it was assumed that a data importer is anyone who *follows* the person from country A. While that approach captures more importers (and therefore more data to study), it does not guarantee that the data has actually been traded. By contrast, viewing retweeting of the data as active information trade ensures that a piece of information produced in one country has actually been consumed and passed on in another.
The assumption here is that the more information a country exports to another the more influence it has over that country. Thus, the important geolocation information is which country the Twitter users were from: we needed the location that the users consider themselves to either be living in, or to be their origin, instead of the location they happened to occupy when their tweet or retweet was sent. We also assume that if users are willing to state their location, they are most likely to define this in the UPL field.
We used the Ukraine and Syria data to study information trade and in particular the portion of it for which we obtained geoparser output of the UPLs. This allowed us to assign each tweet in the set to a Twitter user from a specific country. When a tweet is retweeted, the tweet metadata contains the UPL information from the original tweeter, and this allows us to geo-locate the original tweeter as well as the retweeter. Once we have determined the importer and exporter of the tweet, we can aggregate the data and chart information flows across national boundaries.

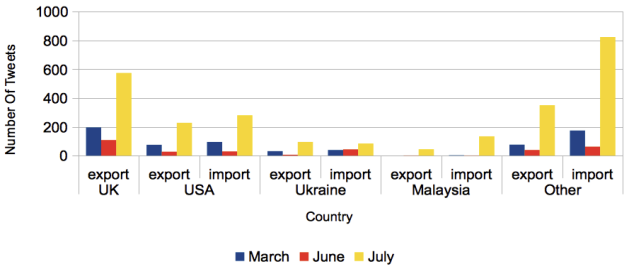

Figure 4: Domestic and international information trade of the UK for the Ukraine data.

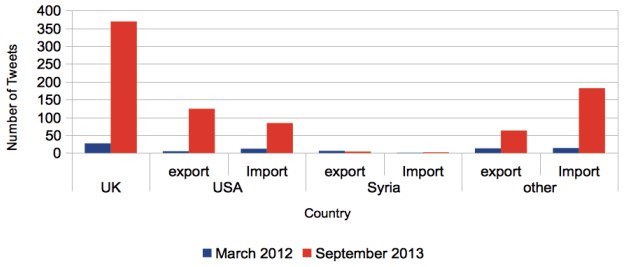

Figure 5: Information trade of the UK for the Syria data.

4.1.1. Ukraine

The Ukraine dataset contained data gathered across three weeks (see Section 3.1). The final week showed a dramatic increase in Twitter traffic on this topic mostly due to the MH17 plane crash. Figure 4 compares domestic information trade compared to the import and export activity between the UK and three selected countries (the USA, Ukraine, Malaysia) as well as all other countries. The import values are tweets imported from those countries to the UK, and the export values are tweets exported from the UK to those countries. The figure shows that tweets from the UK on the Ukraine topic were more retweeted in other countries than tweets from these countries were retweeted in the UK, suggesting a high level of influence of the UK on the rest of the world. This is particularly striking in the case of the USA as the official language of both countries is English and this study was restricted to English tweets. In order get a better understanding of influence of countries via Twitter information trade, this study needs to be broadened to include other languages as well.

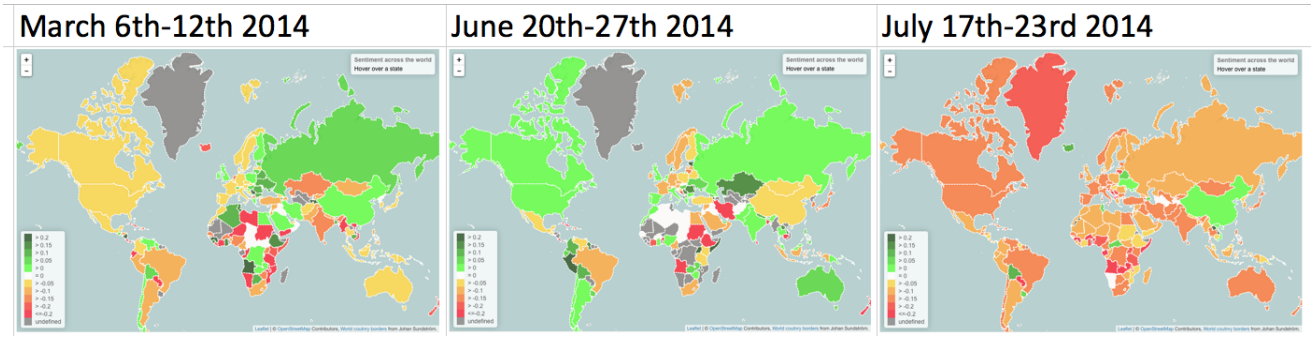

Figure 6: Sentiment towards Ukraine.

4.1.2. Syria

The method described above was also applied to the Syria dataset. Figure 5 shows the import and export activity between the UK and the USA, Syria, and all other countries. The UK was a net exporter of tweets during both periods, but the trade gap increased dramatically in the second week. Figures 4 and 5 show that our information trade case study involved very small numbers of tweets and re-tweets due to the topic and time limitation of the data collected. We believe that information trade patterns will be more meaningful for larger datasets.

In summary, Twitter can tell us about how well connected a country is through social media, and how that connectivity can be traced in the ways in which information flows between people and countries. This analysis requires accurate country-level geolocation of users around the world and illustrates the value of an effective geoparser. Here, it lets us identify when a tweet from the UK gains retweets outside the UK, and vice versa.

4.2. Sentiment Analysis

Analysis of conversations in Twitter can give an indication of what topics are under discussion worldwide. Extracting user locations from user profile information combined with sentiment analysis then enables the study of attitudes towards a topic country by country.

We used the UKRAINE, SYRIA, CITIES and GLASGOW 2014 datasets for this study but due to space restrictions we only present the visualisation for the Ukraine set. The locations for Twitter users were identified using version 2 of the Edinburgh Geoparser. The sentiment for each tweet was determined using VADER,[7] a sentiment analysis tool developed for social media data (Hutto and Gilbert, 2014). This tool give a positive, negative and neutral score for sentences from tweets. These scores can then be aggregated according to the tweet author's country.

Since the Ukraine dataset is made up of tweets from three weeks in 2014, we can compare the attitudes on the topic of Ukraine across time (see Figure 6). Initially the USA was mildly negative (yellow), then positive (green), and then very negative (deep orange) on the topic of the Ukraine. This reflects the initial political issues, the hope for a peaceful solution, and the reaction to the MH17 plane crash.

[7] https://pypi.python.org/pypi/vaderSentiment

5. Conclusion

We argued that Twitter UPL and tweet location are different types of information, and that for a range of studies, UPLs can be exploited to chart information origin or flow. The official Twitter documentation states that this record field is "[n]ot necessarily a location nor parseable". However, we showed that the Edinburgh Geoparser can be adapted to carry out effective geolocation on user profiles. Version 2 achieves both high recall and precision and an F-score of around 0.9. The main drawback of the current work is that the evaluation relies on a new dataset; however, we are releasing this data to the community along with this publication to enable its use as a proper gold standard.

We illustrated the utility of geoparsing UPLs with a brief description of results emerging from a study of international information trade, in which retweets are taken as explicit marks of cross-border influence. We also show an example of how the Edinburgh Geoparser output can be used in combination with sentiment analysis to reveal differences in attitudes between different countries. Clearly, this only scratches the surface of what is possible with improved geoparsing of Twitter UPL information, and in future work, we will continue to refine the geoparser (including adapting it for languages other than English) and explore its use on questions of interest in international relations.

6. Acknowledgements

This work was conducted as part of the UK Connectivity project funded by the British Council. We thank our funder, as well as Stuart MacDonald of the Centre for Cultural Relations at the University of Edinburgh, and Prof Juliet Kaarbo of the School of Social and Political Science, who were our collaborators on this project.

7. Bibliographical References

Ajao, O., Hong, J., and Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, pages 1–10.

Chang, H.-W., Lee, D., Eltaher, M., and Lee, J. (2012). @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*, pages 111–118. IEEE Computer Society.

Cheng, Z., Caverlee, J., and Lee, K. (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*, pages 759–768.

Compton, R., Jurgens, D., and Allen, D. (2014). Geotagging one hundred million Twitter accounts with total variation minimization. In *IEEE International Conference on Big Data*, pages 393–401.

Eisenstein, J., O'Connor, B., Smith, N., and Xing, E. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287.

Graham, M., Hale, S., and Gaffney, D. (2014). Where in the World are you? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., and Ball, J. (2010). Use of the Edinburgh Geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49(1):451–500.

Hecht, B., Hong, L., Suh, B., and Chi, E. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246.

Hutto, C. J. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of International AAAI Conference on Weblogs and Social Media*.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we Twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st Workshop on Web Mining and Social Network Analysis*, pages 56–65.

Kinsella, S., Murdock, V., and O'Hare, N. (2011). "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pages 61–68.

Kulshrestha, J., Kooti, F., Nikravesh, A., and Gummadi, P. (2012). Geographic dissection of the Twitter network. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*.

Llewellyn, C., Grover, C., Alex, B., Oberlander, J., and Tobin, R. (2015). Extracting a topic specific dataset from a Twitter archive. In S. Kapidakis, et al., editors, *Research and Advanced Technology for Digital Libraries*, volume 9316 of *Lecture Notes in Computer Science*, pages 364–367. Springer International Publishing.

Mahmud, J., Nichols, J., and Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology*, 5(3):47:1–47:21.

Osborne, M., Moran, S., McCreadie, R., Von Lunen, A., Sykora, M., Cano, E., Ireson, N., Macdonald, C., Ounis, I., He, Y., Jackson, T., Ciravegna, F., and O'Brien, A. (2014). Real-time detection, tracking, and monitoring of automatically discovered events in social media. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42.

Ryoo, K. and Moon, S. (2014). Inferring Twitter user locations with 10 km accuracy. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 643–648.

Takhteyev, Y., Gruzd, A., and Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, 34(1):73–81.