19th Jan 2016

Geoparse Benchmark Open Dataset from the University of Southampton IT Innovation Centre

In recent years there has been a growing trend for the use of publically available social media content (e.g. Twitter, YouTube, Facebook, Instagram) for analytics within the field of journalism. Often content such as eyewitness images and videos are uploaded by people on the scene and needs to be found in the context of its location.

Geoparsing is the process of extracting a location from text. Geoparsing locations [1] from textual descriptions is really useful to automatically geolocate media if an image or video does not already contain a geotag. Typically about 1.5% of tweets have a geotag for example. Geoparsing also helps where videos contain content shots at multiple locations or where the location where the content was uploaded (e.g. the authors home) is different from the location where the content was originally filmed (e.g. at the scene of a major terrorist incident). Manual inspection of many 1000's of images and videos is not very practical - automatic geoparsing is often the solution!

When developing and testing geoparsing and geolocation approaches the question 'how accurate is my approach compared to the others out there?' always arises. Standard benchmark datasets are the best scientific way to evaluate geoparsing techniques. By running approaches on a common dataset, and verifying the location labels against a gold standard ground truth, different approaches can be directly compared to previously published work and the strengths and weaknesses of each examined.

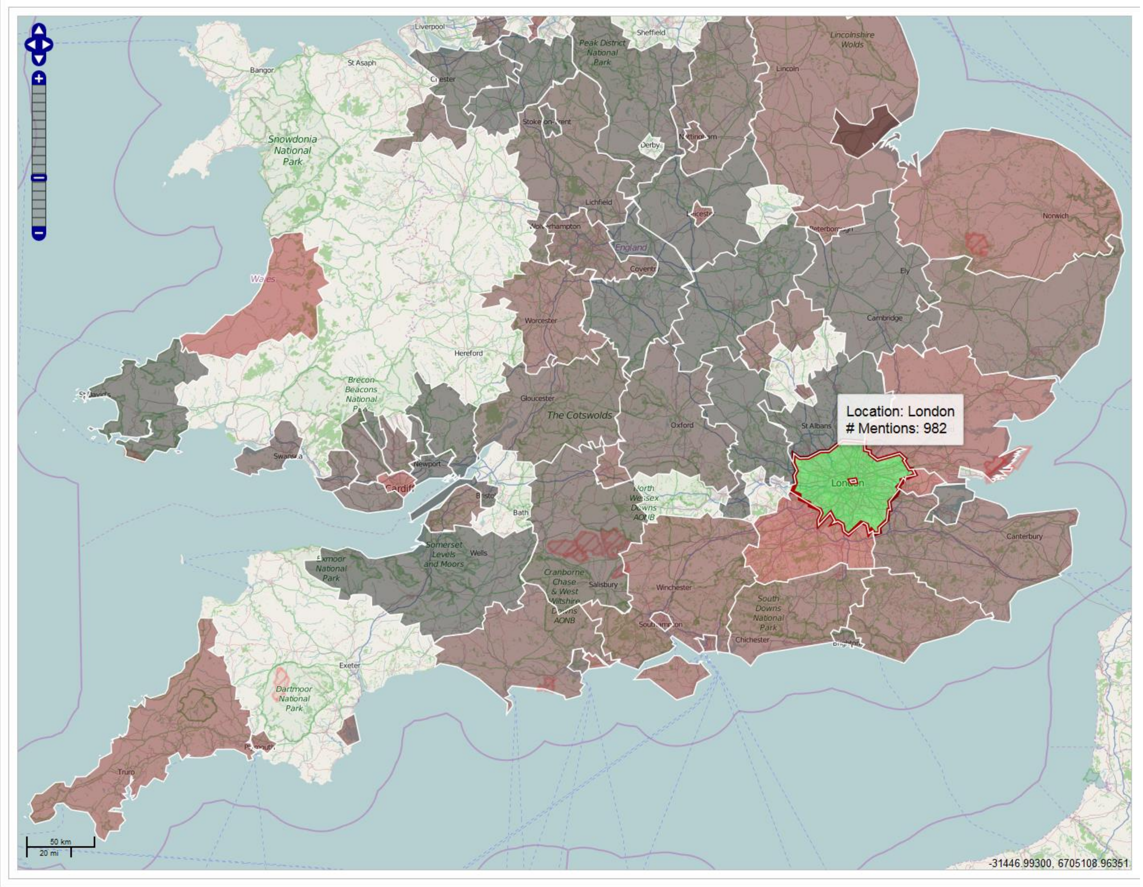

Figure 1. Geoparsing in action: screenshot from the live REVEAL ICT Lisbon 2015 demonstrator

The dataset

The University of Southampton IT Innovation Centre’s geoparsing benchmark dataset contains 1000's of tweets recorded during 4 different natural disasters. These events are Hurricane Sandy 2012, Milan Blackouts 2013, Turkish Earthquake 2012 and the Christchurch Earthquake 2012.

Each tweet in the dataset has been manually labelled with location entries at the building, street and region levels to provide a gold standard for evaluation work. The data consists of the full JSON serialized tweet metadata (i.e. including text) with an additional 'entities' field of type 'mentions' for the ground truth location annotations.

Access to this dataset is free and is hosted at the University of Southampton’s Web Observatory. Anyone can request access - search for ‘GEOPARSE TWITTER BENCHMARK DATASET’ to find it.

An example of previously published scientific results for state of the art geoparsing using this benchmark dataset can be found in [1]. Further reading can be found in [2] and [3] where geoparsing is used to support trust and veracity analysis of social media content (e.g. during breaking news stories).

Sharing to help the research community

This dataset is free to use and is provided with a 4-clause BSD license. Our aim in releasing this dataset is to support researchers who want to test their own innovative geoparsing approaches directly against previously published state of the art solutions.

Extract from the dataset

Below is an example Italian tweet from the Milan blackout event. Most of the fields are from the original tweet and should be faimilar to anyone who has handled Twitter JSON serialized metadata before. Notice the “mentions” field provides a gold standard location label that can be compared to the geoparse result from an algorithm.

```
{ "iso_language_code": "it",
  "entities": {
    "Mentions": [
      { "indices": [64, 69],
        "class": "Location",
        "subclass": "Region",
        "name": "Milano"
      }
    ],
    ...
  },
  "text": "Ma io stasera volevo cenare al lume di candela, non cucinarci! #milano #blackout",
  "id": 332905763536261120,
  "created_at": "Fri, 10 May 2013 17:11:33 +0000",
  ...
}
```

Acknowledgement

The work presented in this article is part of the research and development in the REVEAL project (grant agreement 610928), supported by the 7th Framework Program of the European Commission.

About the author

Stuart E. Middleton is a senior research engineer at the University of Southampton IT Innovation Centre. His main research interests are social media, sensor systems, data fusion and semantics. Stuart has a PhD in Computer Science from the University of Southampton.

@stuart_e_middle   @IT_Innov

http://www.it-innovation.soton.ac.uk   http://users.ecs.soton.ac.uk/sem/

http://web-001.ecs.soton.ac.uk/wo/dataset

REVEAL project, @RevealEU

http://revealproject.eu/

References

[1] Middleton, S.E. Middleton, L. Modafferi, S. 2014. *Real-Time Crisis Mapping of Natural Disasters Using Social Media. Intelligent Systems*, IEEE, vol.29, no.2, 9-17, DOI:10.1109/MIS.2013.126

[2] Middleton, S.E. 2015. *Extracting Attributed Verification and Debunking Reports from Social Media: MediaEval-2015 Trust and Credibility Analysis of Image and Video*. Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15

[3] Middleton, S.E. Krivcovs, V. 2016. *Geoparsing and Geosemantics for Social Media: Spatio-Temporal Grounding of Content Propagating Rumours to support Trust and Veracity Analysis during Breaking News*. To appear in ACM Transactions on Information Systems: Special Issue on Trust and Veracity, ACM