Reverse Geocoding for Photo Captioning with a Meta-Gazetteer

Vlad Tanasescu
TanasescuV@cardiff.ac.uk

Philip D. Smart
SmartP@cardiff.ac.uk

Christopher B. Jones
c.b.jones@cs.cardiff.ac.uk

Cardiff University
Cardiff CF24 3AA, United Kingdom

ABSTRACT

Gazetteers play an essential role in GIS in translating between place name and coordinate-based descriptions of location. The proliferation of location-aware social media applications has led to new sources of gazetteer data, many of which are crowd-sourced. They complement the conventional authoritative resources that are typically linked to published map products. We illustrate the variation in performance of several, mostly social media based, gazetteer resources for a reverse-geocoding photo captioning task and demonstrate the advantage of a meta-gazetteer service that integrates multiple individual gazetteer resources and employs several toponym ranking methods.

Categories and Subject Descriptors

H2.8 [**Database Applications**]: Spatial databases and GIS; H.3.3 [**Information Search and Retrieval**]: Information filtering; H.3.5 [**Online Information Services**]: Web-based services

General Terms

Experimentation, Design

Keywords

Gazetteer, Photo captioning, Data integration, Ranking, Location API, Reverse geocoding, Crowd-sourced data

1. INTRODUCTION

Gazetteers play a key role in geographical information systems by providing the means to translate between the qualitative description of location with place names and their quantitative representation with coordinates. Core information for each place or geofeature recorded in a gazetteer is the toponym (place name), the geographical coordinates, the type of place, and parent places in a geographical hierarchy [4]. A major application of gazetteers is that of reverse geocoding which is concerned with generating one or more toponyms that correspond to a given location specified with coordinates. With the proliferation of geo-tagged media, arising from the increasingly pervasive use of GPS-enabled mobile devices, reverse-geocoding has become an essential method of automatically generating natural language tags. The importance of place names in tagging social media is reflected in the fact that several of these applications, such as Google Maps and Foursquare, have become repositories of user-supplied place names and are associated with gazetteer services that provide access to these names. Such services complement commercial, national mapping-agency and user-contributed sources.

Because so many mobile phones are equipped with GPS and include a simple camera it has become increasingly important to develop methods to help people keep track of where their photos were taken using simple natural language descriptions or tags. In gazetteer based methods for such captioning procedures reverse geocoding services are used to generate place names referring to topographic features such as settlements, lakes, mountains and rivers and to other finer grained points of interest such as cafes, churches and museums. Clearly the effectiveness of the reverse geocoding procedures of gazetteers depends upon their geographic coverage and the level of completeness within a given area. An indication of the variation in geographic coverage of several gazetteers sources was presented in [11].

In an alternative approach to photo captioning, locational tags are obtained from the captions of existing photos that are taken at the same or a similar location as the photo to be tagged [10]. This method is dependent upon other people having taken photos at the location and providing useful locational tags.

In this paper we test the effectiveness of several gazetteer services for the purpose of generating toponyms to use in photo captions and we show that combining multiple gazetteers into a meta-gazetteer service results in significantly improved performance. The performance of the gazetteers is evaluated by attempting to match the toponyms that are used in the captions of photos that have been uploaded to the Geograph web site. Geograph is a project that aims to create a photographic record of every square kilometer of the British Isles. It differs radically from sites such as Flickr and Google Picasa, in that much more care is invested in creating a useful caption for each photo that is uploaded. Typical photos on Geograph employ several toponyms to describe both the subject of the photo and the local context. We treat these

SIGSPATIAL '14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666492

captions as a form of gold standard for purposes of evaluation.

Reverse geocoding procedures will typically generate multiple toponyms for any given location. In order to generate a preferred name or set of names for a given photo location, a means of ranking is required. When testing the individual services this ranking is provided for us by the respective service. The methods used are not well documented but we can expect that distance from the target location is likely to be a primary consideration. For our meta–gazetteer service we take the names generated by the individual services and re–rank them using several methods, based on distance from the target location, web popularity of the name and the frequency of use of the name if it is present in existing captions of Flickr photos taken in the vicinity of the target photo location.

2. RELATED WORK

There have been several efforts to construct gazetteers from one or more social media resources. Kennedy et al [5] exploited the content of geo-tagged Flickr captions to extract place names as well as events. Popescu et al [9] extracted place names, place types and coordinates from georeferenced Wikipedia articles and from the Panoramio photo sharing site, using Panoramio to refine the coordinates of toponyms identified in Wikipedia. The relative importance of names for a given location was determined using a combination of the rank data item from Panoramio and by measuring the popularity of a toponym when used in a search engine query. Here we also experiment with ranking methods that includes web search popularity, but we focus on integrating multiple existing structured gazetteer resources, rather than extracting place name knowledge from text sources.

A gazetteer service that integrates multiple sources of structured geo-referenced toponym data was described by Manguinhas et al [7]. Using ETL (extract transformation and loading) methods, they implemented wrappers to convert to a common XML format based on the Alexandria Digital Library schema, accessing various data sources including the Geonames gazetteer and the GeoNetPT OWL ontology (with Portugese toponyms). For purposes of matching equivalent instances they applied a thresholded string matching metric in combination with a test for candidate pairs either being within some distance threshold or having equivalent feature types. Gazetteer integration was also addressed by Brauner et al [3] who presented a method to use inferred equivalence of gazetteer instances, based on similarity of coordinates, to integrate the corresponding feature types (from the Alexandria Digital Library feature type thesaurus (FTT) and the GeoNET thesaurus). Our approach to gazetteer integration is similar to the meta-gazetteer methods of [11] which adopted a three layered architecture consisting of a foundation layer with, mostly API-based, interfaces to multiple gazetteer resources, such as Geonames and OpenStreetMap, a mediation layer to select resources and to merge retrieved data using entity resolution methods, and an application layer to implement methods for reverse-geocoding and geocoding.

Automated reverse geocoding was integral to the PhotoCompas system [8]. It organised collections of photos, that have geographical coordinates, into coherent hierarchical groups based on space (map coordinates) and time. The spatio-temporal clusters of photos were labelled with place names, that were categorized as either regional containment names or nearby places. A single (un-named) source of US geographic data with polygonal boundaries was used for generating the place names. Nearby cities were ranked according to a combination of distance from the photo cluster, population and Google search count when searching with the city name and its parent state. To decide on the salience of selected names in the present work, we have included the use of the web search engine count method, as previously mentioned, in combination with two other measures based on distance from the given photo location and on use within captions of Flickr photos taken at the same location.

3. META-GAZETTEER SERVICE

For the purpose of comparing individual gazetteers we selected seven resources, summarized as follows:

Google Places. These names are obtained from the Google reverse geocoding API that returns places containing or near to the input location. It includes settlements, street names and points of interest.

Foursquare. The names are provided by the Foursquare reverse geocoding API and are particularly rich in commercial venues.

Geonames. This is a user-contributed resource dedicated to recording place names. It has a wide range of feature types and extensive geographical coverage and includes settlements and urban and topographic landmarks.

OpenStreetMap Nodes. This is a user-contributed resource that includes streets, buildings and points of interest. It records many urban commercial premises. Only entities tagged with 'amenity' are retrieved here with the *MapQuest* Open API service

Yahoo! PlaceFinder. Names are obtained from the reverse geocoding service that returns places located at the input query point coordinate.

DPpedia. This is the semantic web version of Wikipedia and contains many geo-referenced places. It is accessed via the SPARQL endpoint and while not formally a gazetteer source it is used here as the named places can be regarded as having a relatively high salience and hence might be expected to figure in photo captions.

Ordnance Survey (UK) 1:50,000 gazetteer. The toponyms consist of settlements and topographic feature names that are derived from the OS 1:50,000 scale map series. The coordinates in the gazetteer are at 1Km resolution. In being a national mapping agency source it is characterised by systematic coverage of Great Britain.

The meta–gazetteer (MG) integrates the above sources with access predominantly based on live use of APIs, as opposed to the ETL style of methods that characterise other integration approaches such as DIGMAP [7] and aspects of [11]. The basic query methods consist of search for toponyms and points of interest centered on a point location, or by bounding box, or both. Typical query results are lists of geofeatures, in JSON or XML syntax.

Characteristics of each API can be described as data or terms of service related. Most though not all of the resources require authentication, the rationale for which is of-

ten to allow the provider to enforce invocation rates and conditions described in terms of service. The conditions of use of APIs include attribution of data provenance, display requirements for source logo and permission to cache results. An important parameter of the individual services with effect on data quality is whether some of the data are crowdsourced. Other characteristics of location APIs relevant to the functioning of the MG are the APIs' scope, bias as well as granularity. These describe, respectively, the geographic area coverage, their orientation towards a certain type of geofeature, and the scale of the geofeatures.

Once the first invocation returns, and results are retrieved, the MG performs an entity resolution (ER) process on the result set. ER, also known as deduplication, merge-purge, fusion or record linking, consists in identifying, grouping, and merging records determined to represent the same real-world entity [2]. As data are received by the MG, they first go through a schema-level integration phase, in which name, location, and type are extracted according to their original schema, before being mapped to a common data model.

This is followed by an entity identification phase which consists in processing pairwise similarity tests between attributes of the abstracted entities. Unless entities are crowd-sourced which could result in duplicates in the same result set, entities are only compared to entities originating in a different data store. When comparing entities only spatial and textual similarity measures between the coordinates and name are used. This is motivated by the need to be able to integrate a variety of resources that differ considerably in the types of data item that are recorded for each geofeature. The textual similarity is determined with a combination of the Levenshtein edit distance [6], Soundex [1] and text normalisation, using the same method as described in [11]

3.1 Ranking methods

Here we describe briefly the three ranking methods used to order results retrieved by the meta–gazetteer. They are based on distance from the query point, web popularity of the toponym and use of the toponym in Flickr photo captions.

For the distance rank function each toponym is assigned a score in $[0, 1]$ using a linear scale from the closest toponym (given a score of 1) to the furthest away (scored 0).

The Web popularity rank function performs a Yahoo web search engine query for each toponym candidate, in combination with its parent country, and allocates a score in $[0, 1]$ based on the number of web pages returned, normalized relative to all candidate toponyms that are to be ranked. It may be noted that this is used here largely for purposes of comparison, as it was used in some previous studies (PhotoCompas, DIGmap). It cannot be regarded as particularly effective as many toponym words have an alternative non-geographic sense or may be associated with some commercial or other popular feature that will bias the results.

The Flickr popularity ranking function is based on the assumption that places that are photographed are likely to be more salient than those that are not and that there should be some increase in the ranking according to the number of photos that use a toponym in a caption. The ranking employs a count of the number of Flickr photos within 100m of the query location that contain the target toponym in the title, description or tags. The count is normalised by the total number of photos in the query region. Clearly the use of this function depends upon the presence of Flickr photos at the given location.

4. EVALUATION WITH GEOGRAPH

For the evaluation we selected a geographically randomised sample of 400 captions and their associated coordinates from the Geograph website of manually generated photo captions. From each caption we retrieved all toponyms, using named entity recognition to identify proper nouns that were treated as candidate toponyms that we then filtered manually. Note that within Geograph captions the majority of proper nouns represent place names. The toponyms were then normalised to remove punctuation and all stop words.

4.1 Comparison of individual gazetteers

Each of the gazetteers listed in Section 3 was queried with each of the Geograph photo locations and for each location the first five toponyms returned by the respective service were retained. The number of five was chosen as being a reasonable maximum number of toponyms that might be included in any caption. When treated as a ranked list they also provide a measure of how well each gazetteer service is able to predict a toponym that might be used in a real world caption. Each toponym was then subject to the same process of normalisation and stop word removal that was applied to the Geograph caption toponyms. The ordering of each individual gazetteer is that provided by the API, with the exception of the Ordnance Survey Gazetteer and DBpedia in which ranking was based on distance from the query location.

In order to assess the relevance of a gazetteer toponym to the Geograph caption we implemented a simple scoring procedure to measure, for each photo location, the degree of match between each retrieved gazetteer toponym and each of the Geograph toponyms. If the set of tokens in the gazetteer source T_G matched exactly the set of tokens of a candidate Geograph toponym T_g a score of 2 was awarded. If a subset of the tokens matched, then a score of 1 was applied and if there were no matching tokens between the compared pair of toponyms then the score was 0.

These scores were then used to compute the Normalised Discounted Cumulative Gain (NDCG) [12] as a measure of the effectiveness of the ranking of toponyms returned by each gazetteer and of the meta–gazetteer.

The results for the seven gazetteers are presented in Table 1. A Full Match is one in which the Geograph toponym tokens are equivalent to or a subset of those retrieved by the gazetteer, while Any Match refers to the situation in which the retrieved gazetteer toponym tokens either include a subset of those in the Geograph caption or there is a full match, i.e. it includes the Full Match results.

The best performing gazetteer both with respect to ranking and the number of matches within the 5 retrieved geofeatures is the UK Ordnance Survey gazetteer. Notably this

Table 1: Comparison of of individual gazetteers

Resource	NDCG	Full Match%	Any Match%
OSGazetteer	0.44	27.00	55.50
Foursquare	0.31	21.75	48.00
Geonames	0.30	14.50	30.25
Yahoo	0.23	10.75	22.75
GooglePlaces	0.20	13.00	26.50
OSMnodes	0.18	12.00	27.00
DBpedia	0.13	7.50	13.75

Table 2: Performance of the Metagazetteer

Ranking method	NDCG	Full Match%	Any Match%
flickr	0.57	29.25	78.00
distance	0.54	29.25	78.00
web popularity	0.43	29.25	78.00

gazetteer is the only national mapping agency resource and it has a uniform coverage across Great Britain. The next best resource in both respects is Foursquare. It is apparent that there is considerable variation in the performance of the set of gazetteers. It is perhaps not surprising that DBpedia is last in the list on all criteria as it is not in itself a mapping or gazetteer service.

4.2 Performance of the meta-gazetteer service

Table 2 illustrates the results of running the meta-gazetteer service to perform the same task as described above for the individual gazetteers. The three rows of the table distinguish between the results of using each of the three ranking methods summarised in Section 3.1

These results demonstrate the very clear advantage of the meta-gazetteer over all of the individual sources with an approximately 50% advantage over the best performing individual gazetteer. With regard to the ranking methods, it is clear that the web popularity based method performs poorly compared to the other two methods. The Flickr method has a small performance advantage over the distance based method, but is computationally more expensive as it requires a call to the Flickr API. The computational cost can be mitigated somewhat by performing this call in parallel with all other API calls of the meta–gazetteer.

5. CONCLUSIONS

In this paper we have evaluated the performance of several gazetteer services for a reverse geocoding task to generate place names for geo-located photos and demonstrated the benefits of using a meta-gazetteer service that integrates multiple sources. In the evaluation of the individual sources the single country-specific national mapping agency gazetteer proved the most effective in generating toponyms to match those used in manually generated captions from the Geograph site that records photos for the whole of the British Isles. The other social media and user-generated gazetteers varied considerably in their performance, with Foursquare being the best performing of those. The meta-gazetteer service has been described and evaluated with regard to the reverse geocoding task and specifically with respect to the use of three ranking methods based on distance from the query point, web popularity of the names and level of use of the toponym in Flickr. The Flickr method gave the highest performance but was only marginally superior to distance weighting.

6. REFERENCES

[1] The soundex indexing system. In *http://www.archives.gov/research/census/soundex.html*.

[2] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. Whang, and J. Widom. Swoosh: a generic approach to entity resolution. *The VLDB Journal*, 18:255–276, 2009. 10.1007/s00778-008-0098-x.

[3] D. Brauner, M. A. Casanova, and R. L. Milidiú. Towards gazetteer integration through an instance–based thesauri mapping approach. In *Proceedings of the 8th Brazilian Symposium on GeoInformatics*. 2006.

[4] M. Goodchild and L. Hill. Introduction to digital gazetteer research. *International Journal of Geographic Information Science*, 22(10):1039–104, 2008.

[5] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: Context and content in community-contributed media collections. In *Proceedings of ACM Multimedia*. ACM, 2007.

[6] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1965.

[7] H. Manguinhas, B. Martins, and J. Borbinha. A geo-temporal web gazetteer integrating data from multiple sources. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 146 –153, nov. 2008.

[8] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Fourth ACM/IEEE-CS Joint Conference on Digital Libraries*, 2004.

[9] A. Popescu, G. Grefenstette, and P.-A. Moëllic. Gazetiki: Automatic creation of a geographical gazetteer. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 85–93. 2008.

[10] B. Sigurbjornsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. World Wide Web Conf.*, pages 327–336, 2008.

[11] P. Smart, C.B.Jones, and F.A.Twaroch. Multi-source toponym data integration and mediation for a meta-gazetteer service. In *Proceedings of GIScience 2010*, volume 6292 of *Lecture Notes In Computer Science*, pages 234–248, 2010.

[12] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu. A theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*. 2013.