Building a Graph Database of Place Descriptions

201374125

Introduction

This project aims to provide a structured graph database of location descriptions for use in geographic natural language processing tasks. Using a graph database for the store of this information allows for semantic associations between locations to be inferred based on their proximity within the database. This provides an alternative to the purely coordinate based and euclidean associations between geographic locations, and attempts to capture platial connections, as opposed to purely spatial ones. This considers previous work addressing the development of hierarchical geo-ontologies (Sun *et al.*, 2019), and attempts to describe the hierarchical nature of definable geographic concepts in a computer interpretable way, demonstrated on **Fig. 1**.

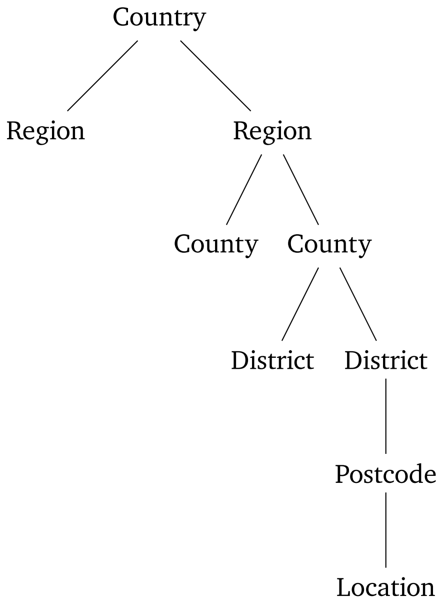

Fig. 1: *The hierarchical representation of geographical concepts*

Purely geographic implementations of similar work already exist, for example Geonames, and OpenStreetMap, which both have comprehensive coverage of places for much of the UK. However, these do not provide any descriptive language associated with the geographic locations. DBpedia is an example of a more general knowledge base, and given it is built from data extracted from Wikipedia, it provides descriptions for many of the items contained within it.

Labelled data is often a primary concern when performing many analytical tasks, and is considered a particular issue in geographic natural language processing (Gritta, 2019; Stock *et al.*, 2013). Much of the existing work has relied on time consuming, manual labelling of data which leads to smaller datasets (Middleton & Krivcovs, 2016; Wallgrün *et al.*, 2018; Gey *et al.*, 2006), and many which are not made freely available (Leidner & Lieberman, 2011; Leidner, 2007; Andogah, 2010; Weissenbacher *et al.*, 2019). The lack of large, high quality, labelled geographic natural language data is well noted by many authors in this subject area

(Tobin *et al.*, 2010; Speriosu, 2013; Weissenbacher *et al.*, 2015, 2019; Gritta *et al.*, 2018; Karimzadeh *et al.*, 2019).

The production of a dataset containing descriptive information, in addition to hierarchical geo-information regarding locations may assist with a variety of geographic natural language tasks. For example, in toponym disambiguation, contextual information provided alongside the identified toponym is often used as a method for correctly resolving to a single toponym (Tobin *et al.*, 2010; Roberts, 2010; Speriosu, 2013), this may include topics associated with particular toponyms (Speriosu, 2013; Adams & McKenzie, 2013; Ju *et al.*, 2016), and metadata associated with the toponyms, including geotags (Zhang & Gelernter, 2014), and other structured information (Weissenbacher *et al.*, 2015). Additionally, this dataset acts as labelled descriptive information regarding a specific, known (geocoded) location, useful for recent developments in fine-grained localisation research (Al-Olimat *et al.*, 2019; Chen *et al.*, 2018a, 2018b), and point of interest identification (Moncla *et al.*, 2014; Li & Sun, 2014).

This project therefore aims to bring forward the most complete corpus of labelled geographic natural language available through automatic extraction of Wikipedia place summaries, providing both contextual information associated with toponyms and hierarchical geo-information.

Methodology

DBpedia

DBpedia is a crowd sourced collection of information extracted from Wikimedia projects, presented in a structured format resembling an open knowledge graph (OKG). This provides linked data in a machine-readable format, accessible through a SPARQL querying API. The data follows the Resource Description Framework (RDF) as defined by the World Wide Web Consortium (W3C) specifications, providing an alternative web linking structure. RDF models for data exchange use URIs to name a relationship between things, as well as information regarding the two ends of each link, generally known as a triple.

As of this report, the DBpedia knowledge base describes 4.58 million things, including persons, places, creative works, and organisations. The data is available under the Creative Commons Attribution-ShareAlike 3.0 Licence and the GNU Free Documentation Licence which allows for copying, redistribution and adaptation of the data, including for commercial use.

Building a SPARQL Query

SPARQL is an RDF query language which allows for the use of namespace prefixes to query URI triples from the DBpedia RDF database. These prefixes include DBpedia defined ontologies, resources, or properties, and additional prefixes, including those defined by the W3C. The DBpedia prefixes are perhaps the most useful, as they provide consistent class definitions for types of thing contained in the database. For example anything given the class `Place` is likely of interest for this research. This class `Place` then provides various subclasses which may be used to perform specific queries, `Architectural Structure` and Celestial `Body` are examples of `Place` subclasses. These subclasses also provide their own subclasses, e.g. `Building`, `Pyramid.`

However, from inspection of specific URIs for places within the United Kingdom, it appears that these RDF links are often incomplete, or provide mismatched information. It is rare for a place that is not classified as a `Building` to have `Building` level class granularity, and many places in the UK only contain the `Place` class. Additionally, the `Country` relation for many places returns primarily the *United Kingdom*, but includes counties e.g. *Dorset*, and occasionally returns *England*, but never *Scotland* or *Wales*. The same is true with `District`, which occasionally returns cities e.g. *Burthwaite* returns *City of Carlisle*. Some naming is inconsistent e.g. *Lewes (district)* vs *Scarborough (borough)* and *South Lakeland* are all labels given to `Place` in the United Kingdom, chiefly when place names are not unique they contain additional information in parentheses. Therefore, it is unlikely that comprehensive coverage may be achieved if attempting to query places using their more granular classes. It is also important for the inclusion of all places that, instead of just querying using the *United Kingdom*, the query includes the constituent countries. Coordinate information is also present for many, but not all, DBpedia places.

Considering the limitations of the database in relation to the United Kingdom, a query was built to obtain English Wikipedia abstracts for each `Place` in any country within Great Britain, using the DBpedia SPARQL endpoint. Given the above constraints, no further information was extracted. This query is given below:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX res: <http://dbpedia.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?lab ?abs
WHERE {
        { ?uri dbo:country res:England } UNION
        { ?uri dbo:country res:United_Kingdom } UNION
        { ?uri dbo:country res:Scotland } UNION
        { ?uri dbo:country res:Wales } UNION
        { ?uri dbo:location res:England } UNION
        { ?uri dbo:location res:United_Kingdom } UNION
        { ?uri dbo:location res:Scotland } UNION
        { ?uri dbo:location res:Wales } .

        { ?uri rdf:type dbo:Place } UNION
        { ?uri rdf:type dbo:Organisation } .

          ?uri rdfs:lab ?lab . FILTER (lang(?lab) = 'en')
          ?uri dbo:abstract ?abs . FILTER (lang(?abs) = 'en')
} LIMIT 10000
```

To overcome the 10,000 query limit, the query was looped over several times with a 10,000 query offset to obtain a full set of results.

Including Metadata

To obtain the further geographic information required for building the hierarchical links between the extracted locations, the Ordnance Survey Linked Data was considered as it was accessible through a SPARQL endpoint. However, the data itself did not contain the hierarchical links required, and given such a large amount of data was being extracted, it was not an efficient way to gather the data. Alternatively, Ordnance Survey provides access to two core datasets through the Edina Digimap service under the Educational User Licence, allowing for free unlimited access of the data for *Educational Use*. However, as this project is undertaken in collaboration with Ordnance Survey, it is likely that the licence will be more flexible.

The two datasets accessed through Digimap were Ordnance Survey Points of Interest (POI), and OS Open Names. The POI data contains various information regarding certain locations in the United Kingdom classified as POI. A POI is generally defined as a location that a person may find useful or interesting. In the case of this data, POI include locations such as *All Saints Church Hall*, the word *Church*, or the names of shops e.g. *The Co-operative*. While much of this data does not provide useful locational information for this project, for any unique named location it may provide additional meta information. The POI dataset was linked with the names of DBpedia locations and provided the additional metadata for the feature geometries `X`, `Y`, `admin_boundary`, `geographic_county`, and `postcode`.

The OS Open Names dataset provided additional information for the majority of DBpedia locations, linking was made particularly easy given the presence of a variable in this dataset called `SAME_AS_DBPEDIA` which enabled accurate linking with the DBpedia dataset for the majority of locations. Additional locations were linked by name as with the POI information. The OS Open Names data provides the additional metadata TYPE; including *hydrography*, *populatedPlace*, *transportNetwork*, etc. With a more granular LOCAL_TYPE. Additionally feature_easting and feature_northing geometries, POSTCODE, BOROUGH, COUNTY and COUNTRY, provided the in depth hierarchical structure as outlined in **Figure 1**.

Building the Database

Neo4j is the most popular graph database management system and has been used in previous work relating to the construction of databases for use in geographic natural language processing to semantically link associations between places (Chen *et al.*, 2018b, 2018a; Kim *et al.*, 2017).

First constraints were created to ensure no duplicates are created, and to improve the efficiency of the database construction.

```
CREATE CONSTRAINT ON (pc:Postcode) ASSERT pc.name IS UNIQUE;
CREATE CONSTRAINT ON (b:Borough) ASSERT b.name IS UNIQUE;
CREATE CONSTRAINT ON (c:County) ASSERT c.name IS UNIQUE;
CREATE CONSTRAINT ON (cy:Country) ASSERT cy.name IS UNIQUE;
```

Following this, the complete data including DBpedia descriptions, and OS metadata was read in periodically 10,000 rows at a time to ensure memory constraints were followed. This code snippet shows the creation of the *Location* node, which includes the name, type, X, Y, and abstract variables. This was repeated for each of the hierarchical locations contained within the data.

```
USING PERIODIC COMMIT 10000
LOAD CSV WITH HEADERS FROM "file:///wiki.csv" AS row
WITH row WHERE NOT row.label IS null
MERGE (l:Location {name: row.label,
                   type: row.TYPE,
                   X: row.GEOMETRY_X,
                   Y: row.GEOMETRY_Y,
                   abstract: row.abs})
RETURN count(l);

..
```

Creation of the relations between locations and their hierarchical relations is given below, in this case postcodes containing locations:

```
USING PERIODIC COMMIT 10000
LOAD CSV WITH HEADERS FROM "file:///wiki.csv" AS row
WITH row.label as lname, row.POSTCODE_DISTRICT AS pcname
MATCH (l:Location {name: lname})
MATCH (pc:Postcode {name: pcname})
MERGE (pc)-[rel:CONTAINS]->(l)
RETURN count(rel);

..
```

Results

Of the 49,923 *Place* and *Organisation* labelled results extracted from the United Kingdom (excluding Northern Ireland), 23,020 were linked successfully with the OS POI, and OS Open Places data.

The Ordnance Survey POI data contains a total of 4,320,574 results, of which, 70,959 were initially joined based on the DBpedia label. Of these, only 6,646 provided unique abstracts, with analysis revealing that DBpedia places with many repeated entries in the POI inventory included general concepts such as *Guide Post*, and company names like *Premier Inn*. Any ambiguous entry was therefore removed to ensure the abstract was correctly associated with the POI. Following this, the POI inventory provided a total of 5,534 unambiguous results. OS Open Data provided 2,927,487 places, joining using the place name first provided 29,367 results, and following the removal of duplicates, this was reduced down to 13,881. It is noted that place names from DBpedia which include additional information e.g. *Place Name, (County)* are unlikely to be included through this method. Using the `SAME_AS_DBPEDIA` column enabled the linking of a further 15,122 rows, assumed to be accurate.

Discussion and Further Work

Previous studies have made use of the various classes DBpedia provides, for example Gao *et al.* (2013) used the `dbpedia-owl:nearestCity` relation in the `City` class to obtain *platial buffers* for city boundaries. They also note the ability to perform *platial joins* to obtain total populations of all towns in the Californian County of Santa Barbara by using the `partOf` predicate in relation to the `County` class. Such examples however rely on both the comprehensive coverage of classes and relations, and do not provide the granularity of relations this study was hoping to achieve. It should be noted that it appears that the United States often has more complete and consistent representations on Wikipedia when compared with the United Kingdom. Additionally, Rizzo & Cano (2015) utilised the ontologies DBpedia provides for entity classification, however, entity recognition work generally considers entities using widely used entity tags, such as those present in the OntoNotes Release 5.0.

Scheider & Purves (2013) discuss the ability to combine Semantic Web reasoning with techniques associated with Geographic Information Retrieval to localise places based on both spatial and semantic relationships found in place descriptions. Namely the relationships found between the place and other places, objects, or activities. Building on this, the work presented in this report could be used for the construction of a semantic gazetteer, first described by Montello *et al.* (2003), with the goal of providing relevant geo-information, extracted from descriptions, in a structured format. Early spatial cognition research identified the semantic associations with place, Lynch (1960) for example describes the cognitive concepts that are apparent from sketches of a city, while Scheider & Purves (2013) note that these concepts are also present within verbal place descriptions, together with the spatial prepositions which link them.

The additional information provided through a semantic gazetteer is considered essential for the ability to both extract and geocode spatial expressions in natural language and to further improve toponym disambiguation techniques. Chen *et al.* (2018b) demonstrate the use of graph databases to provide semantic links between spatial relations from place descriptions, where places are nodes, and spatial relationships are edges. Chen *et al.* (2018a) note however that by constructing place graphs in such a way, the triples formed do not contain the majority of the additional context. Particularly they note that these place graph models do not consider the additional semantics or related human activities which would prove useful for further spatial reasoning tasks.

Wolter & Yousaf (2018) describe in detail the ability to derive information from place descriptions. When presented with a river, humans may perceive the ability for it to be followed, or with a hill, a person may describe going up or down, information which may be present in descriptive language. Spatial relations within descriptions rely on the reference frame of the place, e.g. *"in front of"* is interpreted differently for a building than for then end of a route, *nearness* also relies on a understanding of the scale of the place being considered. Human cognitive principles also shape how place is described, however as often is the case, Wolter & Yousaf (2018) note that place description and the place itself are likely all that is known. Finally, Wolter & Yousaf (2018) note that the language used within a description may provide an indication of the granularity being considered, and further entities being described will also provide further information regarding the place.

Chen *et al.* (2018a) take the principals as described by Wolter & Yousaf (2018), and use them to further extend the place graph database model, taking all the information provided by place descriptions to form detailed nodes containing detailed semantic information regarding a place. Future work may mirror and build on these concepts, enabled through the data provided in this report, and build towards the construction of a semantic gazetteer.

References

Adams, B. & McKenzie, G. (2013) 'Inferring thematic places from spatially referenced natural language descriptions', *Crowdsourcing geographic knowledge*. Springer, pp. 201–221.

Al-Olimat, H.S., Shalin, V.L., Thirunarayan, K. & Sain, J.P. (2019) 'Towards Geocoding Spatial Expressions (Vision Paper)', *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '19*. Chicago, IL, USA: ACM Press, 2019, pp. 75–78.

Andogah, G. (2010) *Geographically Constrained Information Retrieval*. p. 205.

Chen, H., Vasardani, M., Winter, S. & Tomko, M. (2018a) A Graph Database Model for Knowledge Extracted from Place Descriptions. *ISPRS International Journal of Geo-Information*. 7 (6). p. 221.

Chen, H., Winter, S. & Vasardani, M. (2018b) Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*. (17). pp. 31–62.

Gao, S., Janowicz, K., McKenzie, G. & Li, L. (2013) *Towards Platial Joins and Buffers in Place-Based GIS*. p. 8.

Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D. & Rocha, P. (2006) *GeoCLEF 2006: The CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview*. p. 20.

Gritta, M. (2019) *Where are you talking about?* p. 159.

Gritta, M., Pilehvar, M.T. & Collier, N. (2018) 'Which Melbourne? Augmenting Geocoding with Maps', *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1285–1296.

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B. & McKenzie, G. (2016) 'Things and strings: Improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling', *European Knowledge Acquisition Workshop*. Springer, 2016, pp. 353–367.

Karimzadeh, M., Pezanowski, S., MacEachren, A.M. & Wallgrün, J.O. (2019) GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*. 23 (1). pp. 118–136.

Kim, J., Vasardani, M. & Winter, S. (2017) Landmark Extraction from Web-Harvested Place Descriptions. *KI - Künstliche Intelligenz*. 31 (2). pp. 151–159.

Leidner, J.L. (2007) Toponym resolution in text: Annotation, evaluation and applications of spatial grounding. *ACM SIGIR Forum*. 41 (2). pp. 124–126.

Leidner, J.L. & Lieberman, M.D. (2011) Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*. 3 (2). pp. 5–11.

Li, C. & Sun, A. (2014) 'Fine-grained location extraction from tweets with temporal awareness', *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*. Gold Coast, Queensland, Australia: ACM Press, 2014, pp. 43–52.

Lynch, K. (1960) OCLC: 255254857 *The image of the city*. Publication of the Joint Center for Urban Studies Nachdr. Cambridge, Mass.: MIT PRESS.

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. *ACM Transactions on Information Systems*. p. 27.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J. & Gaio, M. (2014) 'Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus', *Proceedings of the 22nd ACM SIGSPATIAL Interna-*

tional Conference on Advances in Geographic Information Systems - SIGSPATIAL '14. Dallas, Texas: ACM Press, 2014, pp. 183–192.

Montello, D.R., Goodchild, M.F., Gottsegen, J. & Fohl, P. (**2003**) *Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries*. p. 20.

Rizzo, G. & Cano, A.E. (**2015**) *Making Sense of Microposts (#Microposts2015) Named Entity rEcognition & Linking Challenge*. p. 10.

Roberts, K. (**2010**) *Toponym Disambiguation Using Events*. p. 6.

Scheider, S. & Purves, R. (**2013**) 'Semantic place localization from narratives', *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP '13*. Orlando FL, USA: ACM Press, 2013, pp. 16–19.

Speriosu, M. (**2013**) *Methods and Applications of Text-Driven Toponym Resolution with Indirect Supervision*. p. 173.

Stock, K., Pasley, R.C., Gardner, Z., Brindley, P., Morley, J. & Cialone, C. (**2013**) 'Creating a corpus of geospatial natural language', *International Conference on Spatial Information Theory*. Springer, 2013, pp. 279–298.

Sun, K., Zhu, Y., Pan, P., Hou, Z., Wang, D., Li, W. & Song, J. (**2019**) Geospatial data ontology: The semantic foundation of geospatial data integration and sharing. *Big Earth Data*. 3 (3). pp. 269–296.

Tobin, R., Grover, C., Byrne, K., Reid, J. & Walsh, J. (**2010**) 'Evaluation of georeferencing', *Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10*. Zurich, Switzerland: ACM Press, 2010, p. 1.

Wallgrün, J.O., Karimzadeh, M., MacEachren, A.M. & Pezanowski, S. (**2018**) GeoCorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*. 32 (1). pp. 1–29.

Weissenbacher, D., Magge, A., O'Connor, K., Scotch, M. & Gonzalez-Hernandez, G. (**2019**) 'SemEval-2019 Task 12: Toponym Resolution in Scientific Papers', *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 907–916.

Weissenbacher, D., Tahsin, T., Beard, R., Figaro, M., Rivera, R., Scotch, M. & Gonzalez, G. (**2015**) Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. *Bioinformatics*. 31 (12). pp. i348–i356.

Wolter, D. & Yousaf, M. (**2018**) 'Context and Vagueness in Automated Interpretation of Place Description: A Computational Model', P. Fogliaroni, A. Ballatore, & E. Clementini (eds.). *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*. Cham: Springer International Publishing, pp. 137–142.

Zhang, W. & Gelernter, J. (**2014**) Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*. (9). pp. 37–70.