Name of the session: BioGeoRef2021: Georeferencing Biological Locality Descriptions

Type of session: Mini-Hackathon

Short description of the session:

The task of georeferencing text documents is well established within the broader field of geographical information retrieval and is essential to support spatial indexing of text resources. Most of the work conducted on the subject to date has focused on determining the geographic coordinates associated with a document based either on detecting and geocoding place names within the document, or on creating language models that associate the entire vocabulary of the document with geographic locations. In both cases little, or more commonly no, attention is paid to the specific aspects of descriptions of the locations that might be indicating that a referred location was some distance away from or had some other specific spatial relationship with a named place. For example “10km North of Tadcaster”. For many applications such distinctions might not be of much consequence, especially if the intention is to retrieve documents that relate to some local region. There are however a number of applications in which locational precision is very important, such as in descriptions of locations of an emergency (e.g. road accident), specific damage incidents caused by an earthquake, or in the case of this mini-hackathon, descriptions at which biological or other natural environment samples were obtained. In the latter case locational precision and accuracy could be of great importance in creating high quality maps of for example, the distribution of species.

This session is concerned with the task of predicting the locations referred to by brief but often quite specific descriptions of the locations at which biological samples were collected, as in “Hundalee Hills, stream gorge on west side of Highway 1 at rest area 4.9 km N of Conway River bridge.” The provided data comes from the environmental agency Manaaki Whenua - Landcare Research New Zealand, which is one of many such agencies that have accumulated thousand and sometimes millions of such descriptions. It would be of great value to these organisations to develop effective georeferencing methods for determining the coordinates of the locations referred to in the natural language descriptions. Doing so would enable mapping the distribution of particular biological species, in some cases over a few hundred years which would itself contribute to studies of biodiversity and climate change. Descriptions of this sort date back in New Zealand at least to the 18th century, to the time of Captain Cook’s voyages to the South Pacific. At present only a relatively small proportion of the records in such collections have been allocated useful geographic coordinates.

Several papers have been published on the subject of georeferencing detailed location descriptions a few of which are listed below. Some approaches investigate the use of acceptance regions that denote the possible area referred to a by a spatial relational term relative to a reference object. Notably the georeferencing software GeoLocate makes use of some of these methods. The published methods remain limited however in their ability to generate accurate coordinates for many types of description. Challenges that are faced include the use of multiple spatial relational terms (such as *on, near, beside, along from, next to, north of* etc) within a variety of language structures and often referring to reference objects that are generic names of things rather than formal toponyms. It is also the case that many of the toponyms used are ambiguous or absent from gazetteers, particularly for older descriptions.

About the Project

The data set belongs to the BioWhere database. The BioWhere project is funded under the New Zealand MBIE Endeavour Smart Ideas Funding Scheme. The project will develop methods to determine geographic coordinates (e.g. latitude and longitude) of text location descriptions to unlock huge amounts of biological data. Millions of records of species locations in biological collections, scientific reports, and journal papers are in textual form (e.g. 'South-east of Wellington, mouth of Orongorongo River, near coast'), lacking the coordinates needed to map species distribution. The project will use artificial intelligence to generate coordinates corresponding to descriptions of locations, in collaboration with both local (e.g. Te Papa Tongarewa) and international organisations (Kew Gardens, Natural History Museum, UK).

The Challenge

Participants will be provided with a collection of approximately 300K locality descriptions, and their coordinates, divided into training and test (about 10% of the total) sets. Each locality description describes the location at which a biological specimen from the collections of Manaaki Whenua – Landcare Research was collected. All of the locality descriptions describe locations in New Zealand (potentially including off-shore islands and NZ). Specimens may be for plants, fungi, microorganisms or invertebrates, but the species is not included in the data set that will be provided.

The challenge is to create a method to accurately georeference (predict the coordinates for) each locality string in the test set using any method. Approaches might include conventional use of named entity recognition and geocoding tools, out of the box methods (e.g. https://www.geo-locate.org/), machine learning regression or perhaps language modelling approaches as can be obtained using the BERT transformer models in regression mode.

The following is an example of the data that will be provided, possibly with additional attributes such as date, collector, region.

Locality	Lat	Long
Buller, Paparoa Mountains, north flank of Mt Euclid, c. 1-1.5km east of Morgan Tarn.	-41.9562	171.6032
Auckland Island, lower slopes about Musgrave Inlet	-50.6469	166.1533
Nelson, about l km SE of Lake Peel, in the track to Balloon Hut	-41.1316	172.6001
Marlborough, hills about Queen Charlotte Sound	-41.3859	173.7136
Lake Ellesmere Spit = Kaitorete Spit - About Midway along length.	-43.874	172.2679
Mokaikai Scenic Reserve, above Whareana Bay, North Cape	-34.4482	173.0002
Heaphy Track, slopes above Perry Saddle.	-40.8983	172.4001
Canterbury, Loburn area, White Rock, above Karetu River	-43.1546	172.4527
Lookout above Huia, Waitakere Ranges	-36.9982	174.5668
Canterbury, Banks Peninsula: above Breeze Bay, Lyttleton Harbour.	-43.5968	172.7831
Hunua, bush above Hunua Falls	-37.0676	175.0907

The winner will be judged as the entry with the lowest mean absolute error in geodetic distance in metres between the actual and predicted location calculated for 75% of the test data set (that is, the largest 25% errors may be excluded from the calculation).

Participants may plan or prepare their method beforehand if they wish, but the data set will not be made available until the beginning of the hackathon. The data is owned by Manaaki Whenua - Landcare Research (MW-LR), and when registering, participants must agree to delete any data after the event, or enter into negotiations with MW-LR to arrange and sign a data agreement.

Resources:

Participants will be provided with links to a range of different resources to get them started, which could include geographic data sets; gazetteers; tools such as NER tools, https://www.geo-locate.org/ etc, library functions to calculate geodetic distance between latitude and longitude (to assist with calculation of mean absolution error). They may also use any additional resources they choose.

Format:

The session will begin with a brief introduction (5 minutes), and then participants will have 60 minutes to complete their solution and prepare a single slide summarising their method and the results they achieved. Participants will then be invited to present/discuss their method and results (2-3 minute time limit per participant, depending on number of participants). The winner will be required to provide their code to the organisers for verification purposes after the event.

The winner will be receiving a certificate and NZ$250 cash prize after the event.

Expected participation (i.e., who would be interested in attending your session): Anyone who is interested in georeferencing textual descriptions, and willing to have a go at solving the challenge that we have set. Participants can compete either individually or as a team of up to 3 members.

Names and affiliations of team members that will lead the session:

- Azadeh Izadi, Massey University, New Zealand
- Kristin Stock, Massey University, New Zealand
- Chris Jones, Cardiff University, Wales
- Aaron Wilton, Manaaki Whenua – Landcare Research, New Zealand

References

Liu Y, Guo QH, Wieczorek J, Goodchild MF. Positioning localities based on spatial assertions. International Journal of Geographical Information Science. 2009, 23(11):1471–501.

Guralnick RP, Wieczorek J, Beaman R, Hijmans RJ, BioGeomancer Working Group. BioGeomancer: automated georeferencing to map the world's biodiversity data. PLoS Biol. 2006, 4(11):e381.

Doherty P, Guo Q, Liu Y, Wieczorek J, Doke J. Georeferencing Incidents from Locality Descriptions and its Applications: a Case Study from Yosemite National Park Search and Rescue. Transactions in GIS. 2011, 15(6):775–93.

Bloom D, Wieczorek, JR, Zermoglio P. Georeferencing Calculator Manual [Internet].Copenhagen: GBIF Secretariat; 2020. Available from: https://docs.gbifuat. org/georeferencing-calculator-manual/1.0/en/

Wieczorek J, Guo Q, Hijmans R. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science. 2004 Dec 1;18(8):745–67.

Guo Q, Liu Y, Wieczorek J. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. International Journal of Geographical Information Science. 2008, 22(10):1067–90.

van Erp M, Hensel R, Ceolin D, Meij M van der. Georeferencing Animal Specimen Datasets. Transactions in GIS. 2015, 19(4), 563–81.

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

Hall MM, Smart PD, Jones CB. Interpreting spatial language in image captions. Cognitive Processing 2011, 12(1):67–94.