INITIAL PLAN - MAPPING LOCATIONS IN TEXTS

CM3203 – Individual Project

Abstract

A short introduction to the project “Mapping Locations in Texts”, in which I will explore the use of Natural Language Processing methods for Named Entity Recognition in natural text

Supervised by Christopher Jones

Charlie Webb

Webbcj1@cardiff.ac.uk

Contents

Introduction

As technology progresses in the modern world, we often find that it can be difficult to combine the tracks of the human world and the technology world. This challenge highlights itself most in the task of linguistic input. While machines require precise, structured input, humans do not, and so when we speak in so-called "natural language", we find that the way we speak and communicate is unsuitable for input to machines.

It follows, then, that we need systems to process our natural language, in order to convert our loose, unstructured communication into meaningful text that a machine can understand.

Within the large domain of Natural Language Processing (NLP), we consider a subsection known as Named Entity Recognition (NER), in which a parsing system considers a chunk of natural language and tags entities within that chunk. These entities are tagged with a category, which allows a computer to gain an understanding of the subjects of a natural language caption.

Consider only the entities tagged as "location" by the NER method. Once we have those, what can we do with them? Problems such as geo-geo ambiguity suddenly arise – If the tagger returns "Raleigh", which of the 15 Raleigh's in the world is it referring to? We need a geocoder, an algorithm that will resolve these geo-geo ambiguity issues and work out the correct location to which the text refers.

Specifically in the world of ecological study and preservation, we find that we have massive sets of data stretching back to before we standardised the storage of data. Thousands of records of data pertaining to the location of species and specimens recorded in natural language that have yet to be processed and tagged, which could contain valuable data for those in the field. A challenge arises, then, that we must build systems to accurately tag this data and convert it from natural language to computerised data.

This idea is not new. There are systems in place and papers published (van Erp et al., 2015; Murphey et al., 2004) that have made attempts at tackling this NER task before. This project seeks to improve upon the work that has already been published and attempt to maximise the effectiveness of the NER and geocoding on the datasets given. In that regard, this project will not be considered a general georeferencing application, but instead a specialised application for use on ecological-based datasets.

It is worth mentioning that, as time has gone on, NER methods and geocoding methods have evolved from the old style of rule-based classification to use machine- and deep-learning based methods for classification (Melo and Martins, 2017). Notably, forefront NER tools such as SpaCy and the Stanford NE Recognizer use these machine-learning methods.

There are, already, existing geoparsing systems as well. (Gritta et al., 2017) performed an analysis of five of these geoparsers; Yahoo!PlaceSpotter, Clavin, Edinburgh Geoparser, TopoCluster, and GeoTxt.

The study concludes that, while the geoparsers are successful in tagging a proportion of the data, they still have limitations when it comes to the resolution of, e.g., fuzzy toponyms. In this regard, I feel motivated to attempt to create my own solution to this task.

Aims

This project aims to implement this NER, which will then be applied to a series of captions with references to locations within them. The locations will be disambiguated using a geocoding algorithm and then the disambiguated locations will be plotted on a map.

While the overall use of this project is not limited to any specific data, I will be collaborating with the National Museum of Wales, the Natural History Museum of England, and Kew Gardens (non-exhaustive). This is because the nature of the project lends itself well to use for ecological documentation and preservation. The aim is that the application will be used to convert rows of natural language data, each representing a caption describing the location of an object, into interactable points on a map which will show better the distribution of objects of interest to these parties.

The hope is that I will be able to involve these interested parties in the progress meetings along the course of this project. To that extent, the requirements listed below are, while certainly the core of what I feel I must implement, not exhaustive of what the final version will contain.

While the aim of this project is to "succeed", it is also worth noting that it aims to perform a service to the interested parties. The ability to automatically process and tag ecological data will provide a powerful tool to map the geographic spread of species over time, in order to infer more easily data pertaining to the general wellbeing and health of species. As well as converting textual data to spatial data, the project aims to be used on previously unreferenced data. Processing this unreferenced data may lead to valuable insights into data that was previously ignored, or data may be inferred that was previously obscured due to the textual nature of the input data.

Objectives and Requirements

1. Implement a suitable NER method for entity recognition within natural language.
    a. Optimise this NER method to maximise evaluative scores.
    b. Research into optimisation methods (training, statistical models, etc.)
2. Implement a suitable Geocoding library, likely GeoPy
    a. Optimise this Geocoding method to maximise evaluative scores.
    b. Research into optimisation methods (gazetteer choice, etc.)
3. Link the above to an intuitive and aesthetically pleasing GUI.
    a. Research GUI choice
4. Research and implement a database link to store processed data.
    a. Implement indexing.
5. Deploy application as a web application.
    a. Research methods for easy deployment
    b. CGI, Templating, Etc.
6. Provide a finished project that will perform a service in order to monitor the ecological health of the Earth.

Ethical Considerations

This project will not be storing any personal or sensitive data, and in that regard, no ethical considerations need to be made.

Work Plan

Week 2 – 8th February

- Begin project diary.
- Begin research on best NER libraries.
- Begin research on best practice for NER improvement methods.
- Create wireframe of front end.
- Supervisor meeting.

Week 3 – 15th February

- Implementation of basic un-improved NER working on manual input captions.
- Begin research on geocoding best practice and improvement methods.
- Implementation of basic front end unlinked to back end.
- Create medium fidelity mock-up of front end.
- Supervisor meeting.

Week 4 – 22nd February

- Begin improvements upon NER methods.
- Begin structuring for web application.
- Implement basic unimproved geocoding.
- Begin front end implementation.
- Supervisor meeting.

Week 5 – 1st March

- Concentrate on improving NER.
- Begin improving geocoding.
- Continue front end implementation.
- Begin research on DB storage and linking.
- Supervisor meeting.

Week 6 – 8th March

- Finalise DB functionality.
- Begin implementation of DB functionality.
- Polish front end.
- Supervisor meeting.

Week 7 – 15th March

- Continue DB implementation.
- Continue polishing front end.
- Supervisor meeting.
- (Possible) Interested party meeting for project evaluation.

Week 8 – 22nd March

- Finalise DB implementation.
- Finalise project.
- Supervisor meeting.

Week 9 – 19th April

- Begin project final report.
- Begin evaluating NER method.
- Begin evaluating geocoding method.

Week 10 – 26th April

- Continue writing final report.
- Produce graphs and metrics for NER accuracy.
- Produce graphs and metric for geocoding accuracy.

Week 11 – 3rd May

- Continue writing final report.
- Consider any final changes to code

Week 12 – 10th May

- Polish final report.
- Submit final report.
- Submit code.

References

Van Erp, M., et al., 2015, Georeferencing Animal Specimen Datasets, *Transactions in GIS* 19(4), 563 – 581

Murphey, P.C, et al., 2004, Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal DatabaseInformatics Initiative (Mapstedi), *PhyloInformatics* 3, 1-29

Melo, F. and Martins, B., 2017, Automatic Geocoding of Textual Documents: A Survey of Current Approaches, *Transactions in GIS* 21(1), 3-38

Gritta, M. et al. What's missing in geographical parsing?, *Lang Resources & Evaluation* 52, 603–623. https://doi.org/10.1007/s10579-017-9385-8