

Initial Plan

Author - Ellie Robins C1618392

Supervisor – Chris Jones

CM3203

One semester individual project – 40 credits

Project Title

“Recognising and disambiguating place names in text documents”

Project description

To index documents with regards to geographic space it is required to recognise and geocode place names in the textual content. My project is an information retrieval project that is concerned with developing machine learning methods to perform this task. Georeferencing is an important and challenging task due to the richness and ambiguity of natural language. Geo-references are a common place task in GIS that involves associating information with same location in physical space which can take different forms of postal addresses, place names, post codes and coordinates. This location should only refer to one place which is typically is the case with given reference. In a document identifying references to location helps understand its geographical context. Geoparsing will include identifying geo-references which will use named entity recognition which is the process of assigning words or groups of words to a set of predefined classes or categories such as locations, person, organisation etc. Most named entity recognition algorithms exploit lists of known locations. These lists are called gazetteers. The task can be challenging because of the difficulty of distinguishing genuine place names from other terms, such as the names of people and organisations, and because some place names (such as Newport) are ambiguous due to different places having the same name. There are various approaches to geoparsing which include simple list lookup which is simple fast and language independent, knowledge or rule-based methods which uses surrounding context to capture geographical context. Machine learning approaches to geoparsing employ types of evidence that a name is a place name based on whether it occurs in a gazetteer or if it is preceded by spatial relations such as ‘near’ or ‘within’ and whether it is a particular instance of a vernacular place name such as ‘Big Apple’ is associated to New York City, USA.

Project Aims and tasks:

Aim 1 – *Detect place name in a text document:*

Task 1: Become familiar with machine learning and techniques that will be used to apply algorithms, packages, tools and libraries.

Task 2: Research named entity recognition algorithms, understand how they work and how they can be applied. This includes downloading all relevant tools, packages, libraries and source codes.

Task 3: Research gazetteers and understand how they work and how they can be applied.

Task 4: Be familiar with using and applying named entity recognition algorithms and understand how applying them can detect place names.

Task 5: Be able to use gazetteers to give all instances of a name it recognises.

Task 6: Be able to use the technologies, techniques, algorithms, APIs, tools, libraries and gazetteers that will be able to detect a word is a place name in a text document.

Task 7: Be able to use and understand how StanfordNLP, NLTK, SpaCy work.

Task 8: Be able to use and understand how API's from geonames and OpenStreetMaps work.

Task 9: Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.

Aim 2 – *Add coordinates and geographic identifier to a place name in a text document:*

Task 1: Be familiar of machine learning techniques that apply disambiguating algorithms.

Task 2: Research and understand how techniques can disambiguate place names and how they can be applied. This includes downloading all relevant libraries, tools, source code and packages.

Task 3: Become familiar with applying and using techniques that disambiguate place names with respect to a gazetteer.

Task 4: Apply and use the techniques to disambiguate place name in a text document.

Task 5: Be able to use and understand how techniques such as Edinburgh geoparser, CLAVIN, geolocate and google places API's work.

Task 6: Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.

Aim 3 – *Evaluate the techniques used to detect and disambiguate place names in a text document:*

Task 1: Research what and how the techniques should perform like.

Task 2: Familiarise how to evaluate the techniques such as setting up timers for them.

Task 3: Test the techniques using the relevant research and evaluate them as such if they work, how can they be improved.

Task 4: Give areas of how techniques can be improved.

Task 5: Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.

Aim 4 – *Improve techniques that have been applied to detect and disambiguate place names in a text document*

Task 1: Be familiar with machine learning, training machine learning mechanism and train and test data.

Task 2: Add to techniques if they do not work.

Task 3: Adapt techniques or bring other techniques together to improve performance or any other areas that need improving.

Task 4: Add to techniques to detect non-gazetteer place names.

Task 5: Be able to write my own techniques, APIs or algorithms to achieve aims 1 and 2 of detecting and disambiguating place names in text documents.

Task 6: Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.

Aim 5 – Further gazetteers to detect biological specimens in places to determine place names

Task 1: Be familiarised with biological specimen data.

Task 2: Add to gazetteer to detect biological specimens to determine place names.

Task 3: Apply techniques and use them to detect and disambiguate place names in a text document by detecting biological specimens.

Task 4: Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.

Work Plan:

I will be following the university weeks in the semester as a guide to follow a work plan.

Priority rating from 1 – 4, very important – least important. I have given the main aims and their tasks a week in which I am for them to be completed and a priority rating. I have given research and familiarisation tasks a second priority as while attempting to use the techniques I will be familiarising myself with them. Research is still a high priority as I want to be able to gain a full understanding of the techniques and the function that will add to my project and help me successfully complete my tasks. As I have very little machine learning knowledge, I have given myself up to week 6 to complete the first two aims as this will be a challenging task becoming familiar with the algorithms and programs. This will give me a longer length of time compared to the other aims as I want to ensure I can complete the initial project title aims and learn new machine learning techniques and approaches to applying the APIs. I believe giving myself the time to learn and understand these techniques thoroughly I will benefit from a further understanding so the last 3 aims will become easier to achieve.

Aim	Task	Week	Priority
1 – Detect place name in a text document	1 - Become familiar with machine learning and techniques that will be used to apply algorithms, packages, tools and libraries.	2	2
	2 - Research named entity recognition algorithms, understand how they work and how they can be applied. This includes downloading all relevant tools, packages, libraries and source codes.	2	2
	3 - Research gazetteers and understand how they work and how they can be applied.	2	2
	4 - Be familiar with using and applying named entity recognition algorithms and understand how applying them can detect place names.	2	2
	5 - Be able to use gazetteers to give all instances of a name it recognises.	3	2

	6 - Be able to use the technologies, techniques, algorithms, APIs, tools, libraries and gazetteers that will be able to detect a word is a place name in a text document.	3	1
	7 - Be able to use and understand how StanfordNLP, NLTK, SpaCy work.	4	1
	8 - Be able to use and understand how API's from geonames and OpenStreetMaps work.	4	1
	9 - Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.	4	1
2 – Add coordinates or geographic identifier to disambiguate place name in text document	1 - Be familiar of machine learning techniques that apply disambiguating algorithms.	4	2
	2 - Research and understand how techniques can disambiguate place names and how they can be applied. This includes downloading all relevant libraries, tools, source code and packages.	4	2
	3 - Become familiar with applying and using techniques that disambiguate place names with respect to a gazetteer.	5	1
	4 - Apply and use the techniques to disambiguate place name in a text document.	5	1
	5 - Be able to use and understand how techniques such as Edinburgh geoparser, CLAVIN, geolocate and google places API's work.	6	1
	6 - Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.	6	1
3 – Evaluate techniques used to detect and disambiguate place names in a text document	1 - Research what and how the techniques should perform like.	6	3
	2 - Familiarise how to evaluate the techniques such as setting up timers for them.	7	3
	3 - Test the techniques using the relevant research and evaluate them as such if they work, how can they be improved.	8	3
	4 - Give areas of how techniques can be improved.	Easter Break	3
	5 - Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.	Easter Break	1
4 - Improving techniques and tools used to detect and disambiguate place names in a text document	1 - Be familiar with machine learning, training machine learning mechanism and train and test data.	Easter Break	3
	2 - Add to techniques if they do not work.	Easter Break	3
	3 - Adapt techniques or bring other techniques together to improve	Easter Break	3

	performance or any other areas that need improving.		
	4 - Add to techniques to detect non-gazetteer place names.	9	3
	5 - Be able to write my own techniques, APIs or algorithms to achieve aims 1 and 2 of detecting and disambiguating place names in text documents.	9	3
	6 - Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.	9	1
5 – Further gazetteers to detect biological specimens to detect place names in text documents	1 - Be familiarised with biological specimen data.	9	3
	2 - Add to gazetteer to detect biological specimens to determine place names.	10	3
	3 - Apply techniques and use them to detect and disambiguate place names in a text document by detecting biological specimens.	11	4
	4 - Write report on how and if I achieved this aim, what went well and how it could be improved if it did it again.	11	1

References:

- I. R.S. Purves, P. Clough, C.B. Jones, M.H. Hall and V. Murdock. Geographic Information Retrieval: Progress and challenges in spatial search of text. Foundations and TrendsR in Information Retrieval, vol. XX, no. XX, pp. 1–161, 2018.
- II. Peng Qi, Timothy Dozat, Yuhao Zhang and Christopher D. Manning. 2018. Universal Dependency Parsing from Scratch In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 160-170. [StanfordNLP]
- III. <https://www.geonames.org/>
- IV. © OpenStreetMap contributors - <https://www.openstreetmap.org/copyright>
- V. SpaCey - <https://explosion.ai/legal>
- VI. Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc. [NLTK]
- VII. [Edingurh geoparser]
 - A. Beatrice Alex, Clare Llewellyn, Claire Grover, Jon Oberlander and Richard Tobin. Homing in on Twitter users: Evaluating an Enhanced Geoparser for User Profile Locations. 2016. In the Proceedings of the 10th Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož, Slovenia. [pdf]
 - B. Beatrice Alex, Kate Byrne, Claire Grover and Richard Tobin. 2015. Adapting the Edinburgh Geoparser for Historical Georeferencing. International Journal for Humanities and Arts Computing, 9(1), pp. 15-35, March 2015. [pdf]

- C. Beatrice Alex, Kate Byrne, Claire Grover and Richard Tobin. 2014. A Web-based Geo-resolution Annotation and Evaluation Tool. In Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII), COLING 2014, Dublin, Ireland. [pdf]
- D. Presentation at Pelagious workshop, March 2011.
- E. Bea Alex and Claire Grover. 2010. Labelling and spatio-temporal grounding of news events. In Proceedings of the workshop on Computational Linguistics in a World of Social Media at NAACL 2010, Los Angeles, USA. [paper]
- F. Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010b. Use of the Edinburgh Geoparser for georeferencing digitised historical collections. *Philosophical Transactions of the Royal Society A*, 368(1925):3875-3889. [paper]

VIII. [CLAVIN 4/6/2015]

- A. CLAVIN Installation and Usage Guide — Built using Maven Site, this collection pages provides details on Installation and Usage.
- B. CLAVIN-REST Installation and Usage Guide — Built using Maven Site, this collection pages provides details on Installation and Usage.
- C. CLAVIN, CLAVIN-REST, CLAVIN-NERD Javadocs — Typical javadoc output for the three primary components within the CLAVIN ecosystem.
- D. CLAVIN Product Brochure — The Berico Technologies product brochure on CLAVIN, and our professional support services to tailor CLAVIN to your use case.
- E. AWS Support and EULA — Our Amazon Web Services Support and End User License Agreement.

IX. [Google places] https://cloud.google.com/maps-platform/places/?utm_source=google&utm_medium=cpc&utm_campaign=FY18-Q2-global-demandgen-paidsearchonnetworkhouseads-cs-maps_contactsal_saf&utm_content=text-ad-none-none-DEV_c-CRE_342707335086-ADGP_Hybrid+%7C+AW+SEM+%7C+BKWS+~+Google+Maps+Places+API+EXA-KWID_43700042842848036-kwd-22859391737-userloc_9045648&utm_term=KW_google%20places%20api-ST_google+places+api&gclid=EAIaIQobChMI44LFsMqk5wIVFeDtCh2TcAoDEAAYASAAEgIQOvD_BwE

- X. Nelson E. Rios, Lead Principal Investigator, email: nelson.rios@yale.edu <https://www.geo-locate.org/standalone/default.html>
- XI. Crown copyright ©. All data and other material produced by Land Information New Zealand (LINZ) constitutes Crown copyright administered by LINZ. <https://gazetteer.linz.govt.nz/>