Which ‘Lancaster’ do you mean?
Disambiguation challenges in extracting place names for Spatial Humanities

Paul Rayson, Alistair Baron, Andrew Hardie
Lancaster University
Motivation

• For Spatial Humanities, we want to be able to ask:
  • “what place is this corpus talking about?”
  • “what is being said about different places?”
  • “how has the way that places are represented in the corpus changed over time?”

• We can already combine simple gazetteer-based lookup and a Geographical Information System (GIS) to extract candidate place names and overlay the information on a map

• But …
  • scale of full-text datasets is increasing
  • accuracy needs to be evaluated and improved
  • need to overcome the ‘disambiguation challenge’
Challenges for effective place name extraction

• Is it a place name or not?
  • Lancaster (a city)
  • Lancaster bomber (a plane or type of beer)
  • Stuart Lancaster (the England rugby coach)
  • Duke of Lancaster (a nobleman or pub)

• Which place is it?
  • Lancaster (north-west of England)
  • Lancaster (Australia)
  • Lancaster (Canada)
  • Lancaster (United States)

• Noisy historical data
  • Spelling variation in place names.
  • Places changing name over time, boundary changes, etc.
Data and existing place name extraction

- 200,000 pages of data, 12,783,888 words.
- Place names automatically marked up with Edinburgh geoparser.
- Evaluation of place name detection: 82.1% precision / 80.8% recall. Resolution to correct place name: 75.3%–82.0% accuracy.
- 153,840 place names marked up which have been resolved to a gazetteer entry.
- Texts (with place name markup and metadata) put into CQPweb for analysis: http://cqpweb.lancs.ac.uk/
Methods for place name recognition

• Evaluate existing place name recognition via Edinburgh geoparser.

• Begin to explore methods for improving place name extraction and resolution.
  • Part-of-speech tagging
  • Semantic tagging
  • Gazetteer lookup

• To feed into machine learning method for full Named-entity recognition (NER) system.
  • Other contextual features
  • Fuzzy gazetteer matching
  • Metadata aware (where, when, etc.)
Part-of-speech tag: $NP[12]$

- CLAWS part-of-speech tagger
  - Hybrid rule-based and statistical methods
  - Trained and tested on 100 million-word British National Corpus
  - 97-98% accuracy across a variety of text types
  - [http://ucrel.lancs.ac.uk/claws/](http://ucrel.lancs.ac.uk/claws/)

- Tag $NP1$ (singular proper noun) and $NP2$ (plural proper noun) are closest match to place names, but will also match person names, organisations, etc.

- CQPweb used to search for all occurrences of $NP[12]$, evaluated against single words marked as place names.
Part-of-speech tag: \textit{NP[12]}

- 122,181 (99\%) of single word place names are found to be \textit{NP1/NP2}. 3,261 place names types (95.7\%).

- 1,262 (389 types) that are not \textit{NP1/NP2}:
  - Most (764) are tagged as \textit{NN1} (Singular common noun), e.g. “Stone”, “Moray”, “County”, “Connaught”, “shire”.
  - Other tags include \textit{VVO} (213: e.g. “Renfrew”, “Stronvar”, “Connaught”), \textit{JJ} (85: e.g. “Loth”, “British”, “Norman”) and \textit{NN2} (70: e.g. “Ferns”, “Slaters”, “Broadstairs”).

- 413,832 words are tagged as \textit{NP1/NP2} in whole of HistPop.
  - 29.5\% of all \textit{NP1/NP2} are marked as single word place names.
  - 37,914 are part of multi-word place names.
  - Many of the remaining words appear to be missed place names. e.g., London (4,471), Edinburgh (2,742), Glasgow (2,012), etc.
  - Although some person names: George (700), William (639), John (604).
Semantic tag: Z2

- USAS (UCREL Semantic Analysis System)
  - Rule-based and knowledge-based system
  - Trained and tested on wide variety of corpus types and domains
  - 91% accurate on ‘general’ language
  - Tagset based on Tom McArthur’s Longman Lexicon
  - [http://ucrel.lancs.ac.uk/usas/](http://ucrel.lancs.ac.uk/usas/)

- Semantic tag Z2 used which marks “Geographical names”.

- CQPweb used to find all occurrences of Z2, which are then compared to the marked up single word place names.
Semantic tag: Z2

- 99,291 (80.4%) of all single word place names are Z2.
- Only 869 types (25.5%), indicating mainly frequently occurring place names are tagged as Z2.
- Of the 24,152 (2,538 types) that are not Z2:
  - 23,006 are Part-of-speech tag NP1/NP2.
  - 17,736 are Semantic tag Z99 (Unmatched), e.g. “Nairn”, “Lanark”, “Connaught”, “Kinross”, “Leinster”.
  - 4,102 are Z1 (Personal names), e.g. “Leith”, “Renfrew”, “Bedford”, “Selkirk”, “Stafford”
- 275,544 words are tagged as Z2 in whole of HistPop.
  - 36% of all Z2 are marked up as single word place names.
  - 30,760 are part of multi-word place names.
  - Many seem to be missed place names: London (4,173), Edinburgh (2,712), Glasgow (2,003), etc.
  - But many others not place names: British (3,726), Irish (3,447), Church (3,421), County (2,935), Boroughs (2,276).
Gazetteer lookup

• Gazetteers provide a list of place names, simply searching this list for each word in a text allows for crude place name recognition.

• GeoNames used for experimentation.  [www.geonames.org](http://www.geonames.org)
  • Contains over 10 million global geographical names.
  • Available as a web service or can be downloaded, both free of charge.
  • Includes populations, alternate names, as well as longitude and latitudes.

• Need to restrict database:
  • Some place names are likely to match words which are not place names (e.g. other proper nouns).
  • No need to include obscure place names which are unlikely to be mentioned in corpus.
  • Can focus on area of the world most applicable to corpus.
Gazetteer size

Minimum population for global place names

- All: Recall
- No Alts: Recall
- No GB: Recall
- All: Precision
- No Alts: Precision
- No GB: Precision
Filtered Gazetteer

- Filtered Gazetteer to include:
  - All of Great Britain places.
  - All country and capital city names.
  - All places worldwide with population over 100,000.
  - All alternate names for places already included.

- 120,359 (97.5%) of all detected single word place names are found in the filtered gazetteer. 81.1% of place name types.

- Most frequent place names not found: “Kingsway”, Tyne”, “Kilkenny”, “Hodges”, “Stronvar”.

- 332,234 (first letter uppercase) words in the corpus are found in the gazetteer.
  - 36.2% of these words are detected single-word place names.
  - 31,856 are part of multi-word place names.
  - But some odd things in the gazetteer: “Church”, “Law”, “I”, “II”, “Man”, etc.
Combining methods

Recall

Gazetteer
97.5%
POS
96.8%
99.0%
Semtag
79.9%
80.3%
80.4%

Precision

Gazetteer
36.2%
POS
45.1%
29.8%
Semtag
49.4%
47.2%
39.3%
## Multi-word place names

<table>
<thead>
<tr>
<th>No. words</th>
<th>Tokens</th>
<th>Types</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 words</td>
<td>18,913</td>
<td>3,232</td>
<td>West Riding, United States, West Ham, St. Albans</td>
</tr>
<tr>
<td>3 words</td>
<td>10,205</td>
<td>1,409</td>
<td>Counties of Scotland, Isle of Man, City of Glasgow</td>
</tr>
<tr>
<td>4 words</td>
<td>929</td>
<td>163</td>
<td>Administrative County of London, Stow on the Wold</td>
</tr>
<tr>
<td>5 word</td>
<td>345</td>
<td>24</td>
<td>County and City of Cork, City and Suburbs of Dublin</td>
</tr>
<tr>
<td>6+ words</td>
<td>5</td>
<td>4</td>
<td>Administrative County of the Soke of Peterborough</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>153,840</strong></td>
<td><strong>8,238</strong></td>
<td></td>
</tr>
</tbody>
</table>
Multi-word place names

Recall

Precision (3-words only)
Gazetteer ambiguity

- Out of all 153,840 marked place names in corpus, 130,031 (84.5%) are found in (filtered) gazetteer.
- 3,389 different place names.
- 2,567 (75.7%) have only one entry in gazetteer.
- 509 (15%) have 2 options.
- “Sutton” has 20 entries in gazetteer.
- Average of 1.47 entries per place name.
Gazetteer ambiguity

Ambiguous place names found with different sizes of gazetteer.
Which Lancaster?

- 255 occurrences of “Lancaster” in HistPop. All are POS tag NP1.
- 248 are semantic tag Z2:
  - Z1mf: 4 (2=“Joseph Lancaster”, 2=County) (all 4 not marked as place names).
  - Z2c: 2 (1 “Lancaster County”, 1 “Lancaster Borough”) (both marked as place names).
  - Z3c: 1 (“Lancaster Castle”) (not marked as place name).
- 170 occurrences are marked as place names, all mapped to Lancaster City.
  - 104 “Lancaster” (82: County, 20: City, 1: “Lancaster M.B.”, 1 “House of Lancaster”)
- Leaving 85 occurrences not tagged as place names.
  - 38 should be County, 30 should be City.
  - 8 refer to a person (2 “Joseph Lancaster”, 4 mentions of Joseph as “Lancaster”, 2 in lists of last names).
Which Lancaster?

- 4 different Lancaster’s in filtered gazetteer.
  - 1 is the actual City of Lancaster, North England.
  - 1 is modern-day Lancashire, UK.
  - 1 is “Lancaster and it’s surroundings” (North England)
  - 1 is in California, USA.

- 10 if include all global place names with a minimum population of 5,000.

- 47 in whole of GeoNames.
Future work

- Investigate further features for place name recognition.
  - i.e. frames ("County/District/Borough of …"), surrounding context (other place names, "in"), gazetteer entry details (population, location, etc).

- Improve techniques for finding multi-word place names.

- Disambiguation between places with the same name, i.e. using textual metadata and surrounding context.
  - If text concerns UK matters, then more likely to be Lancaster, England.
  - If Lancaster appears in a list of counties, then likely to be modern-day Lancashire.
  - How close (spatially) are other place names mentioned to each candidate place?

- Fuzzy matching against gazetteer.
  - e.g. to deal with problems of spelling variation.

- Apply to other data sets, particularly the Lake District data currently being transcribed.
Acknowledgments

• The research leading to these results has received funding from the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013) / ERC grant “Spatial Humanities: Texts, GIS, places” (agreement number 283850).

• http://www.lancs.ac.uk/spatialhum/

• Thank you to Claire Grover at Edinburgh University for providing the HistPop texts marked up with place names.