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

Paul Rayson, Alistair Baron, Andrew Hardie
Lancaster University

It is already possible to apply simple gazetteer-based named-entity recognition (NER) techniques to textual sources in order to extract candidate place names. Combining NER techniques with Geographical Information Systems (GIS) allows the digital humanities researcher to overlay the information on a map and visualise the result. This allows us to ask questions such as "what place is this corpus talking about?", "what is being said about different places?" and "how has the way that places are represented in the corpus changed over time?". Together, these techniques have spawned the development of Spatial Humanities (Bodenhamer et al. 2010). However, as the application of these techniques widens and the scale of full-text datasets grows (e.g. EEBO-TCP), we need to refine the techniques to improve their accuracy with both identification of candidate place names and linking to the correct location on a map. For example, *Lancaster* (a city), needs to be extracted whereas *Lancaster bomber* (a plane or type of beer), *Stuart Lancaster* (the England rugby coach) and *Duke of Lancaster* (a nobleman or pub) do not. In addition, linking occurrences of *Lancaster* from a textual source to the correct location in the north-west of England rather than other locations in Australia, Canada and the United States would be vital. In this paper, we argue that techniques from corpus linguistics and natural language processing are essential when tackling both these issues, e.g. concordancing to examine the co-text, part-of-speech tagging to identify proper nouns, frequency analysis and keyness statistics to highlight significant differences in place names between texts. Techniques such as edit distance, letter-replacement rules and phonetic matching also permit spelling variation issues in place names to be overcome when extracting mentions (Baron and Rayson, 2008). Collocation between place names and semantic tags on the surrounding context has already proven useful in identifying topics associated with particular places (Gregory and Hardie, 2011). In addition, we argue that combining collocations between place names and distance information from GIS will provide strong evidence for the disambiguation of location.

References

Baron, A. and Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. In *proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, 22nd May 2008.

Bodenhamer D.J., Corrigan J. and Harris T.M. (2010, eds.) *The Spatial Humanities: GIS and the future of humanities scholarship*. Indiana University Press: Bloomington.

Gregory, I. and Hardie, A. (2011). Visual GISting: bringing together corpus linguistics and Geographical Information Systems. *Literary and linguistic computing*, 26 (3), 297-314.