A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service

Ian Densham
Edinburgh University Data Library
Main Library Building
George Square
Edinburgh EH8 9LJ
Scotland, UK
i.densham@ed.ac.uk

James Reid
Edinburgh University Data Library
Main Library Building
George Square
Edinburgh EH8 9LJ
Scotland, UK
james.reid@inf.ed.ac.uk

Abstract

We describe a basic Geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. The development of a geo-parser comes from the need to explicitly georeference large resource collections such as the Statistical Accounts of Scotland which currently only contain implicit georeferences in the form of placennames thus making such collections inherently geographically searchable.

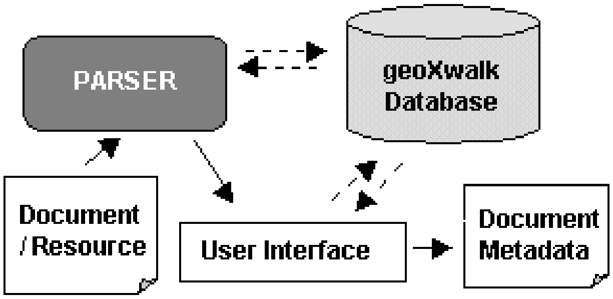

Figure 1: The geo-coding process

1 Introduction

The project is being undertaken by the Edinburgh University Data Library (Edina) as a part of the larger geoXwalk (www.geoXwalk.ac.uk) project which aims to develop a protocol based (ADL, OGC and Z39.50) UK digital gazetteer service. The geo-parser uses the geoXwalk server as the name authority for identified placename candidates.

2 The geo-coding process

In its current implementation, the service consists of two main components, the geo-parser and the geoXwalk gazetteer, with a generic demonstrator interface. The term geo-parsing refers to the identification of placenames in a document/resource, where geo-coding refers to the tagging of the candidate and consequently the resource with a geographic footprint. Figure 1 shows the basic geo-coding flowline.

A resource is submitted to the geo-parser, which identifies a series of potential placenames. Each placename is displayed along with the number of occurrences in the text, and the number of matching gazetteer candidates. For each placename, a link to the gazetteer records is displayed and a highlight option is available for identification in the original text which is displayed beneath the table. Various sorting functions are also available for records of the table. County and feature type are the default attributes for disambiguation, although more are available through the geoXwalk feature specification. Currently multiple gazetteer entries can be attached to a single placename, enabling output of different instances of the same name in the text. Geo-coding output is available in an application specific xml schema, csv, or html, and contains parser and editor matadata. Outputted placenames can be viewed on a map. Clearly the degree of human interaction is high duri! ng the review stage, with the process currently limited to individual resources. As geo-parser development continues, user interaction at this stage of the process will become less, although the potential for 'post process' queries will rise, as the parser is more closely integrated with the geoXwalk database. As geo-parser development progresses the interface will need to accommodate a more flexible approach to the geo-coding process, as clearly interface requirements are determined by users with associated collections of specific document types, and output requirements. A range of functionality is required at various levels between a fully automated batch processing mode and a more interactive analytical approach to individual documents. Further investigation is required on the integration of geo-coding output into existing document metadata.

3 The geo-parser

The current architecture of the geo-parser is conceptually based on several passes across the text at varying levels of abstraction. Documents are split into blocks, blocks into tokens. Tokens are re-constituted into sentences, and the sentences run through a place name finders to identify candidate place names. The current parser implementation uses two techniques. The first applies approximately 300 different regular expressions at the token level based on patterns from training data (The Statistical Accounts of Scotland (http://edina.ac.uk/statacc). Once all the patterns have been run on the document then a second pass is made to find likely placenames in conjunctions / disjunctions with other placenames. Other patterns are also used to attempt to remove false positives such as the names of people, while others are based on the proximity of placename-like words (‘shire’, ‘river’ etc.). The second approach uses the Brill tagger (Brill, 1994) to mark each token with a p! art-of-speech tag, enabling rules to be applied to the text surrounding proper nouns to select likely placenames. Candidate placenames are then cross-referenced with the geoXwalk gazetteer, and a marked up version of the original document and a summary XML version of results returned. The need for large quantities experimental data in order to develop identification and disambiguation further is recognised.

4 GeoXwalk

GeoXwalk is more than just a simple lookup facility for the geo-parser as every geographic feature stored in the gazetteer has its detailed geometry stored with it. This clearly enables more complex searching. The ability to derive the relationships between features implicitly by geometric computation is significant and provides more accurate results than can be ascertained by simple lookups based on hierarchical thesauri methods as in traditional gazetteers. When candidates are referenced against the gazetteer, geoXwalk provides a means to access its ‘alternate’ geographies (of which there are many in the UK) as well as a standard footprint. For example a candidate placename ‘Knowsley’ could be resolved as parish code ‘BX003’ as well as grid reference 340900, 392300 - 347217, 397660. The result is that more powerful geographical based search strategies can be applied e.g. ‘find me all documents about Gaelic songs that do not reference the Western Isles’.

5 Conclusions

Issues encountered during the ongoing development of a document geo-coding tool are on the one hand concerned with the identification and disambiguation of placenames in the text, and on the other, the use of a sophisticated multi-purpose gazetteer service against which candidates are referenced. Interface flexibility is required to accommodate the range of possible approaches to the application.

Acknowledgments. The authors would like to thank all those involved with the GeoXwalk project, David Medyckyj-Scott, Andy Corbett, James Crone , Andrew Robson, Robert Atkey.

References

James Allen. 1995. *Natural Language Understanding*. Benjamin Cummings, Redwood City, CA.

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. In *National Conference on Artificial Intelligence*, pages 722–727.

Claire Grover and Alex Lascarides. 2001. XML-based data preparation for robust deep parsing. In *Proc. Joint EACL-ACL Meeting*, pages 252–259, Toulouse.

Andrei Mikheev. 1999. A knowledge-free method for capitalised word disambiguation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 159–166.

Nina Wacholder Yael Ravin and Misook Choi. 1997. Disambiguation of proper names in text. In *Proceedings of the 17th Annual ACM-SIGIR Conference*, pages 202–208.