A proposal for comparative evaluation of automatic annotation for geo-referenced documents

Paul Clough
University of Sheffield
Western Bank
Sheffield, UK
+44 114 222 2664
p.d.clough@sheffield.ac.uk

Mark Sanderson
University of Sheffield
Western Bank
Sheffield, UK
+44 114 222 2648
m.sanderson@sheffield.ac.uk

1. INTRODUCTION

Organised evaluation campaigns such as the Text Retrieval Conference (TREC) for information retrieval [9], the Document Understanding Conference (DUC) for document summarization [7], and the Message Understanding Conference (MUC) for information extraction [3] have not only proven to be an important and effective stimulus for research, but also served to bring together members of the academic and industrial research communities. These campaigns have resulted in large-scale evaluations in which different approaches and techniques can be compared through the provision of common resources and evaluation strategies.

A core task for Geographical Information Systems (GIS) is identifying and disambiguating geographical references (see, e.g. the SPIRIT[1] and Geo-X-Walk[2] projects). Previous workshops such as the HLT/NAACL 2003 Workshop on Analysis of Geographic References[3] demonstrated a wide variety of different solutions to this problem, but methods were evaluated on different datasets of varying granularity in geographical reference, thereby making comparison difficult.

We propose an evaluation campaign similar to the MUC subtask of recognizing named entities. In MUC, the aim was to recognise entity types including `person`, `organization` and `location`. However in our task, the entities to be found will be more fine-grained than identifying a `location` entity, the only annotation assigned to geographic references. Therefore we propose one task to distinguish geographic references and contexts from other entities (such as organizations and people), and between entity types such as `city` and `postcode`. A further task will be the grounding of geographic references, i.e. assigning them spatial coordinates such as a point or polygon region.

Within such a comparative evaluation, tone can compare the effectiveness of different approaches for identifying and disambiguating geographic references. We will provide the necessary data, annotation scheme, training data and assessments, as well as co-ordinate the campaign. As partners in the SPIRIT project, we have access to the necessary geographical resources and knowledge to define this task. We also have experience of coordinating evaluation campaigns as we have been running a cross language image retrieval task called ImageCLEF [2] for two years. The SIGIR workshop would be an ideal opportunity for us to discuss with other members of the geographic community the details of such an evaluation and stimulate interest and hopefully help in running such a campaign to be run in late 2004 or early 2005.

2. PROPOSED EVALUATION

MUC provides a number of different subtasks for evaluating different aspects of an Information Extraction (or IE) system. For example, in MUC-7 these included Named Entity Recognition (NER), co-reference resolution, and a template relation task. In the NER task, 200 articles were selected from 158,000 newswire stories from the New York Times News Service using domain relevant terms. Named entities were annotated manually by human assessors and 100 articles were supplied as training data and 100 used for testing. Annotations were embedded in the newswire articles using a pre-defined SGML annotation scheme. For example, a location would be encoded within `<ENAMEX>` tags and identified as a location type using an attribute and value pair, e.g. `<ENAMEX TYPE= "LOCATION">Caribbean</ENAMEX>`.

The information extraction task in MUC-7 was realized as filling slots in a pre-defined template. An initial task was to identify Template Elements. This went further than identifying named entities to also extract entity attributes. Two main objects were identified: `ENTITY` and `LOCATION` and used selectively for a given scenario or relation. Attributes for the `LOCATION` object were used to classify the type of location: `CITY`, `PROVINCE`, `COUNTRY`, `REGION`, `WATER`, `AIRPORT` or `UNK` (for any other type of location)[4], the country or region of the location and attributes for further information such as comments. The MUC tasks provide a useful starting point for defining an evaluation and pertinent problems encountered when identifying named entities (e.g. that a body of water cannot be assigned to a country).

Although MUC data would provide a useful starting point (e.g. we could start with this data and geocode the locations), we believe that an evaluation campaign aimed specifically at dealing with the issues surrounding the identification and disambiguation of geographic references in which we collaborate with geographic experts and researchers will create a more useful and relevant resource. Although we will consider existing resources which

[1] http://www.geo-spirit.org/

[2] http://hds.essex.ac.uk/geo-X-walk/

[3] http://gunsight.metacarta.com/kornai/NAACL/WS9/

[4] For more information about the Template Element task, see: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ie_task.html

have been annotated with geographic information as suitable training and test sets, we believe that a more appealing and challenging type of text would be Web data which would enable methods based on analyzing the structure of Web documents to be tested and compared. In addition to this, we have in our possession a 1TByte collection of Web data indexed for text retrieval as part of the SPIRIT project. We would aim to select Web pages which offer various forms of geographical information, such as lists of names and "contact us" details. Because of the linked structure of Web pages may be a useful method of identifying and disambiguating geo-references, we would allow participants to gather these additional linked pages.

Web pages sampled for the training and test sets will be manually analysed for a range of geographic references. At present these are undefined, but might include entities such as: cities, rivers, mountains, islands and regions [8], province, cities world wide, and water regions [4], and more Web-specific geographic references and contexts such as address, post/zip code, phone number and email/web address [1][6][5]. In the SPIRIT project we have been analyzing Web pages to identify different kinds of geographic references which one might commonly encounter, but this would be a point to discuss with participants at the workshop. Like MUC, we will provide guidelines of what constitutes a valid geographic reference.

As well as identifying geographic entities (i.e. geoparsing), a further task will be to ground the entities (i.e. geocoding). Geographic references such as villages, towns, cities etc. can often be defined spatially using a point (e.g. longitude and latitude). However, geographic references which relate to a region (such as a body of water, or province) cannot be specified by a single point but require a set of points to bound them. Initially we may ignore locations which cannot be defined by a single point for simplicity and ease. We plan to use a range of resources and methods of encoding to eventually allow for either representation (e.g. a pointer into an existing geographic ontology). We will encode the training and test data using the Geography Markup Language (GML) to ensure consistency with current geographic annotations, and allow the use of existing mark-up tools based on GML.

We will use a similar evaluation to MUC for scoring systems on identifying and disambiguating geographic references. That is, the proportion of locations successfully found in the test set and whether they have been disambiguated correctly.

3. SUMMARY

We believe that a coordinated evaluation campaign specifically designed to test the effectiveness of identifying and disambiguating geographic references would not only create useful publicly-available resources for the geographic community, but also provide a standard framework in which methods could be compared on a common task/dataset. Results from these kinds of campaigns in other domains have shown them to be valuable and an excellent way of stimulating research. We believe this would help to bring together researchers designing geographic information systems.

4. REFERENCES

[1] Buyukokkten, O., Cho, J., Garcia-Molina, H., Gravano, L. and Shivakumar, N. (1999) Exploiting geographical location information of Web pages. In *Proceedings of Workshop on Web Databases (WebDB'99)* held in conjunction with ACM SIGMOD'99, June 1999.

[2] Clough, P.D. and Sanderson, M. (2003) The CLEF 2003 cross language image retrieval track, In: *Proceedings of Cross Language Evaluation Forum (CLEF) 2003*, Trondheim, Norway.

[3] Chinchor, N. (1998) Overview of MUC-7. In: *Messge Understanding Conference Proceedings*. http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html

[4] Manov, D., Kiryakov, A. and Popov, B. (2003) Experiments with geographic knowledge for information extraction. In: Kornai, A. and Sundheim, B. (Eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 1-9.

[5] McCurley, S.K. (2001) Geospatial mapping and navigation of the web. In *Proceedings of the Tenth International WWW Conference*, Hong Kong, 1-5 May, 221-229.

[6] Morimoto, Y., Aono, M., Houle, M. E. and McCurley, K. S. (2003) Extracting Spatial Knowledge from the Web, In Proceedings of 2003 Symposium on Applications and the Internet (SAINT 2003), 27-31 January 2003, Orlando, FL, USA, 326-333.

[7] Over, P. and Yen. J. (2003) An Introduction to DUC 2003. In: *Document Understanding Conference Proceedings*. http://www-nlpir.nist.gov/projects/duc/pubs.html

[8] Uryupina, O. (2003) Semi-supervised learning of geographical gazetteers from the internet. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 18-25.

[9] Voorhees, E. and Harman, D. (2001) Overview of TREC 2001. In: *Proceedings of TREC 2001*, NIST Special Publication 500-250.