NICTA i2d2 at GeoCLEF 2005

Baden Hughes
Victoria Laboratory
National ICT Australia
Department of Computer Science and Software Engineering
The University of Melbourne
Victoria, 3010, Australia
baden.hughes@nicta.com.au

Abstract

This paper describes the participation of the Interactive Information Discovery and Delivery (i2d2) project of National ICT Australia (NICTA) in the GeoCLEF track of the Cross Language Evaluation Forum 2005. We present some background information about NICTA i2d2 project to motivate our involvement; describing our systems and experimental interests. We review the design of our runs and the results of our submitted and subsequent experiments; and contribute a range of suggestions for future instantiations of a geospatial information retrieval track within a shared evaluation task framework.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

geospatial information retrieval, query expansion, geospatial gazetteer, geospatial grounding, information retrieval evaluation

Keywords

geospatial information retrieval, query expansion, geospatial gazetteers, geospatial grounding, information retrieval evaluation

1 Introduction

National ICT Australia (NICTA) is Australia's information technology and communications Centre of Excellence. The Interactive Information Discovery and Delivery (i2d2) project is based within NICTA's Victoria Laboratory, hosted in the Department of Computer Science and Software Engineering at the University of Melbourne. The i2d2 project is an interdisciplinary project located within a research cluster covering natural language processing, information retrieval, spatial databases, and constraint programming. The overall aims of the i2d2 project is to explore how intelligent linguistic and geospatial analysis of queries and content can enhance the ability of a basic search engine to fulfill a user's information need. This project aims to develop scalable natural language processing technologies for extracting, analysing and presenting information locked up in large bodies of text and speech data on the web. More specifically, this project is oriented towards Australian information content; will feature location-based query and visualisation; will employ scalable methods for linguistic and geospatial annotation; will support spatially-aware document clustering and multi-document summarisation and be deployed in a multimodal interface. The research agenda within i2d2 is aligned NICTA's Priority Challenge: From Data to Knowledge. The project involves the extraction, collation and analysis of high-level semantic relationships from massive quantities of text on the web.

Hence the i2d2 project is inherently interested in geographical information retrieval - finding information involving some kind of spatial context. In complement to the GeoCLEF motivation, i2d2 notes that existing evaluation campaigns such as TREC and NCITR do not explicitly evaluate geographical information retrieval relevance. This our involvement was initiated by the need for an generalized evaluation framework for geographical information retrieval systems, as well as a test case for the systems we are developing.

Our motivations for involvement in GeoCLEF 2005 were broadly twofold: to engage with the broader geospatial information retrieval community in a shared evaluation task and to test a variety of different system components and analytical approaches which are important to the higher level applications within the i2d2 context. As such our expectations about our system performance were in fact quite low, and largely we treated this exercise as a timely informal evaluation of our own progress in the area.

The structure of this paper is as follows: having outlined our motivation for involvement we consider the additional resources used in our experiments; describe our system; evaluate our results; and discuss future options for GeoCLEF.

2 Resources

The system that NICTA i2d2 used in GeoCLEF 2005 is in fact a loosely coupled aggregate of independent components, with the intermediate glue scripts, and some utility functions written from scratch.

For geospatial grounding, we used the Getty Thesaurus of Geographic Names[1], a standard broad coverage gazetteer collated from a wide range of different sources under the auspices of the J. Paul Getty Trust. The Getty Thesaurus of Geographic Names (TGN) contains around 1 million entries for geospatial entities including administrative political entites and physical features. The core of a TGN record is a place, each of which is identified with a unique numerical identifier. Place records are then linked to name variants (common, historical, linguistic); to a node in a hierarchy; to other relationship types; geographical coordinates; notes; data sources; and a place typology.

For named entity recognition, we used the Alias-I LingPipe system[2] in conjunction with the UIUC Cognitive Computation Group's Named Entity Tagger[3]. Neither of these systems was specifically trained for geospatial named entity recognition, but identified named entities generally, including geospatial named entities. In both cases, the benchmark performance for a non-specialised application of the software 'off the shelf' has been shown to approximate the state of the art for named entity recognition in general.

For indexing and retrieval, we used the RMIT Zettair system[4] (formerly known as Lucy). Zettair creates an inverted index of the document collections with native support for TREC formats, and provides support for simple, ranked (non Boolean) and phrase queries. The Zettair system is a robust and well tested system which has been used in a range of external evaluations including larger scale TREC tasks.

[1] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

[2] http://alias-i.com/lingpipe/

[3] http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=NE

[4] http://www.seg.rmit.edu.au/zettair/

3 System Description

In order to address the GeoCLEF tasks, analysis of both the topics and the collection was undertaken as a precursor to running our retrieval system proper.

3.1 Topic Level Analysis

Initially, the relevant parts of each topic was analysed using the two named entity recognition systems, and the geospatial entities indentified. Subsequently for each topic the geospatial entities were resolved via the Getty TGN, resulting in a geospatial and hierarchical grounding of the entity and a typological resolution. From this point the geospatial entities in a given topic can be expanded or reduced by hierarchy traversal in the gazetteer. The result of this analysis is a topic to geospatial scope mapping table, allowing for the expansion of an explicitly specified geospatial entity in a topic to a larger related set of geospatial entities. This table is then used to create the "modified" version of each topic by expansion (or annotation) of the geospatial content of each topic.

3.2 Collection Level Analysis

The LA Times and Glasgow Herald collections were also analysed using the named entity recognition systems, although only the document headlines were processed (rather than the whole document). Each geospatial entity in the headlines was then resolved via the Getty TGN, resulting in a geospatial and hierarchical grounding of the entity and typological resolution. From this point the geospatial entities could be expanded or reduced by hierarchy traversal in the gazetteer. The result of this analysis is a document headline to geospatial scope mapping table, allowing for greater accuracy in matching topic scope to document scope.

3.3 Collection Indexing

The LA Times and Glasgow Herald collections were combined into a single collection and subsequently indexed using the Zettair engine in TREC native mode. The combined collection was indexed in the raw (unannotated) form

The purpose of GeoCLEF is to experiment with and evaluate information retrieval techniques which are oriented towards geospatial entities which are in turn descriptive characteristics of documents in a collection. The basic hypothesis being tested is whether the addition of geospatial entities and locational operands will geographic places which are descriptive of documents. The main idea is to see if addition of geographic operators and geographic locations will improve the accuracy and specificity of retrieval of relevant documents.

In this year's GeoCLEF track, NICTA i2d2 participated in the English monolingual task, that is using English language topics to query an English language document collection. NICTA i2d2 submitted four runs for evaluation, including the two mandatory runs for each task (one run using only the topic title and topic description without using the topic concept tag or topic geographic tags or the topic narrative; and the other required run using both topic title and topic description (but not the topic narrative) and all geographic tags (operator and location) as well as the concept tag.) The runs that NICTA i2d2 submitted were fully automatic, with no human intervention in any part of the experiment process (eg via relevance feedback).

As will be seen in following sections, the main differences in our experiments submitted to GeoCLEF 2005 are in the inclusion of various parts of the topics and the level of geospatial entity expansion based on the topic to geospatial entity mapping tables described earlier. In their simplest forms, our experiments can be construed as principled query expansion of an existing topic with geospatial entities of relevance to create a larger bag of words for index query and retrieval.

4 Experiments

In the table below we show the types of topics and collection materials used in each run. Where the type is 'Raw', this label refers to the unexpanded and unannotated topic or document collection. Where the type is 'Modified', this label refers to the expanded and annotated topic or document collection.

Run	Topic Type	Topic Content
i2d2Run1	Raw	EN-title
i2d2Run2	Raw	EN-title
i2d2Run3	Modified	EN-title and EN-desc
i2d2Run4	Modified	EN-title and EN-desc

In i2d2Run1, we used the raw (ie unexpanded and unannotated) version of the GeoCLEF 2005 topics, with the EN-title element only, against the TREC indexed version of the LA Times and Glasgow Herald document collections.

In i2d2Run2, we used the raw (ie unexpanded and unannotated) version of the GeoCLEF 2005 topics, with the EN-title and EN-desc elements only, against the TREC indexed version of the LA Times and Glasgow Herald document collections.

In i2d2Run3, we used the "modified" (expanded and annotated) version of the GeoCLEF 2005 topics, with the EN-title element only, against the TREC indexed raw version of the LA Times and Glasgow Herald document collections.

In i2d2Run4, we used the "modified" (ie expanded and annotated) version of the GeoCLEF 2005 topics, with the EN-title and EN-desc elements, against the TREC indexed version of the LA Times and Glasgow Herald document collections.

5 Discussion

The overall performance of the i2d2 systems can be summarised as follows. Notably we detected no overall performance increase by the use of topics expanded with geospatial entities over the baseline topics.

Interpolated Recall (%)	Precision Average (%)
0	66.80
10	56.28
20	42.09
30	34.56
40	27.47
50	22.17
60	17.15
70	13.38
80	9.08
90	6.24
100	2.72

The average precision (non-interpolated) for all relevant documents (averaged over queries) is 25.14%. It can be observed that the precision average at early recall points is quite promising, while at lower recall points system performance drops markedly.

The performance of the i2d2 systems at a given retrieval depth can be seen in the table below:

Document Cutoff Level (DCL)	**Precision at DCL (%)**
5 docs	46.40
10 docs	37.20
15 docs	33.33
20 docs	30.00
30 docs	26.52
100 docs	16.00
200 docs	11.02
500 docs	5.71
1000 docs	3.11

Overall the R-Precision (precision after R documents retrieved, where R = relevant retrieved) is 27.69%. Again we can see that at a DCL of less than 15-20, the systems performance is quite promising with precision of between 30% and 46%. At larger DCL values, system performance is markedly degraded.

The precision average for individual queries can be seen in the table below:

Topic	**Precision Average (%)**	**Topic**	**Precision Average (%)**
Topic 001	60.43	Topic 014	10.58
Topic 002	0.83	Topic 015	63.91
Topic 003	0.05	Topic 016	29.20
Topic 004	9.38	Topic 017	43.19
Topic 005	54.26	Topic 018	13.96
Topic 006	21.69	Topic 018	11.02
Topic 007	35.90	Topic 020	24.96
Topic 008	6.24	Topic 021	51.56
Topic 009	29.49	Topic 022	17.92
Topic 010	37.67	Topic 023	0.48
Topic 011	5.11	Topic 024	49.07
Topic 012	4.58	Topic 025	29.91
Topic 013	17.12		

The NICTA i2d2 system performed very poorly on a subset of topics (GC002, GC003, GC004, GC008, GC011, GC012, GC023); although notably all systems appeared to perform badly on these topics. Our system performed reasonably well for about half of the topics (GC006, GC007, GC009, GC010, GC013, GC014, GC016, GC017, GC018, GC019, GC020, GC020, GC022, GC024, GC025). In a small number of topics (GC001, GC005, GC015 and GC021), the NICTA i2d2 system performed at an average precision of over 50%.

6 Future Directions

In this final section, we reflect, based on our experience in GeoCLEF and as a result of general research involvement in this area on a number of desirable items which would be contained in a standardised geospatial information retrieval evaluation set, and may be considered in future instantiations of the GeoCLEF task.

A wider variety of geospatial concepts should be included in the queries. In 2005, the queries were mostly geospatial entities at the country regional level, or the continental level. Standard resources such as country gazetteers and the Getty Thesaurus offer entities at much finer grained levels and could be used as a coarse evaluation metric for the diversity of the concepts in a shared evaluation task data set.

Additionally a much larger set of locational relations ('operators') should be included, rather than the dominant "in" operator used in 2005. Independently the i2d2 project has developed a set of 100 locational operators based on mining web query logs, and many of these have specifically geospatial aspects (cardinal directions for example) in addition to locational semantic load.

We would be happy to contribute this taxonomy in a future instantiation, or customise it as appropriate.

The geospatial resolution task is core to any geographical information retrieval system. In this year's tasks, the only resolution type exhibited was a generalised scope expansion task eg expand Europe to a list of countries. There are a range of other resolution types including scope reduction which should be included in subsequent topic development considerations.

Furthermore, it would be useful to have participants all using a baseline geospatial gazetteer, in order to more objectively evaluate the performance of geospatial information retrieval engines. Without knowing the variety of gazetteers used in the 2005 endeavour at the time of writing (one suspects that it is in fact quite large), it is clear that a hierarchically structured gazetteer would offer significant advantages since many of the geospatial entities were general, and required expansion into finer grained units. We would recommend the use of a broad coverage gazetteer like the Getty Thesuarus of Geographic names as a common gazetter baseline; recognising that there is a cost to acquiring this resource, perhaps a more general resource such as the UNLoCode database would offer a similar baseline.

Finally (although perhaps most obviously) a larger number of topics will be required for robust evaluation. The 25 topics used in this year's effort provided a useful basis for evaluating the viability of the GeoCLEF exercise, particularly by ensuring that the barrier to entry was low,

Whilst we realise that all of these desiderata have a human effort impact, we believe that the overal quality of results, and the standardisation of the evaluation effort will benefit from their inclusion.

7 Conclusions

Our approach to the GeoCLEF track in 2005 was largely exploratory within the bounds of our broader NICTA research project in Interactive Information Discovery and Delivery, which encompasses a range of applications for geospatial information retrieval. While our system is not particularly mature, it is clear that it has capacity to perform well in these types of shared evaluation tasks, and the exercise has allowed us to further refine our development directions. In particular, the resources specifically committed to the GeoCLEF exercise in 2005 were quite minimal; despite this, our system performed moderately well. We look forward to participating in future instantiations of GeoCLEF, and welcome discussions on resourcing this effort into the future.

Acknowledgements

National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology, and the Arts and the Australian Research Council through *Backing Australia's Ability* and the ICT Research Centre of Excellence Programs.