NICTA I2D2 Group at GeoCLEF 2006

Yi Li Nicola Stokes* Lawrence Cavedon Alistair Moffat
National ICT Australia, Victoria Laboratory
Department of Computer Science and Software Engineering
The University of Melbourne, Victoria 3010, Australia
{yli8,nstokes,lcavedon,alistair}@csse.unimelb.edu.au

Abstract

We report on the experiments undertaken by the NICTA I2D2 Group as part of GeoCLEF 2006. We experimented with geographic-based query expansion, using a gazetteer to extend geospatial terms to "nearby" locations, and included sublocations. The processing pipeline of the geographic information retrieval system included: a *named entity recognition* system for identifying location names; a *toponym resolution* component for assigning probabilistic likelihoods to geographic candidates obtained from a gazetteer (the Getty Thesaurus); and a probabilistic approach to Geographic Information Retrieval. We experimented with approaches involving expanding location names in both documents and queries. We used a normalization process to adjust term weights to ensure that geographic terms added to a query do not overwhelm the contribution of non-geographic query terms. We submitted five runs to the English-only GeoCLEF monolingual task, ranging from a baseline task of text-only retrieval based on topic title and description, to queries expanded using gazetteer-based toponym resolution. Our submitted runs showed little improvement for GIR runs over the baseline run. A refinement to the normalization process (post-submission) resulted in GIR runs showing 6.57% and 5.84% improvement over the baseline in overall MAP.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content analysis and indexing – *indexing methods*; H.3.2 [**Information Storage and Retrieval**]: Information storage – *file organization*; H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval – *search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and software – *performance evaluation*.

General Terms

Measurement, Performance, Experimentation

Keywords

Geographic Information Retrieval, Query Expansion

1 Introduction

I2D2 (Interactive Information Discovery and Delivery) is a project being undertaken at National ICT Australia (NICTA), with the goal of enhancing user interaction with information hidden in large document

*Author for correspondence.

collections. A specific focus of I2D2 is in detecting geographically salient relationships, with Geographic Information Retrieval (GIR) being an important challenge.

The system used in the I2D2 2006 submission to GeoCLEF differs significantly from the I2D2 system used in the 2005 submission. We used an updated version of the Zettair [2006] IR engine as our baseline: this resulted in an increase of baseline MAP scores from 0.2514 to 0.3539 over the GeoCLEF 2005 topics. Other differences included introducing Language Technology components: use of the LingPipe *named entity recognition and classification* system; and a *toponym resolution* component. The toponym resolution component assigns probabilistic likelihoods to potential candidate locations for geographic terms identified by LingPipe. Candidates are obtained from a gazetteer; for the GeoCLEF experiments, we used the Getty Thesaurus of Geographic Names. The core retrieval approach involved extending Zettair with a probabilistic technique, described below.

For the problem of retrieving documents containing locations related to those explicitly mentioned in the query, we experimented with both *document expansion* and *query expansion*, that is, replacing geographic terms in documents or queries with a list of related terms, as described below. To combat the drop in precision resulting from geographic expansion, as noted by GeoCLEF 2005 participants (for example, Gey and Petras [2005]), we implemented a *normalization step* to ensure that the added location names do not overwhelm other terms in the topic. The submitted runs used an early version of this normalization; a subsequent refined version saw a significant increase in overall MAP over the non-GIR baseline run.

Overall, our baseline for the GeoCLEF 2006 topics seemed rather low (MAP=0.2312); adding the geographic retrieval techniques increased overall MAP by 6.57% (document expansion) and 5.84% (query expansion) over the baseline, which is a slight improvement over adding the topic's *narrative* field to the query. Variation was seen across topics, often depending on the type of geographic reference, and is discussed in Section 4.

2 System Description

Figure 1 shows the architecture of our approach to probabilistic geospatial information retrieval (GIR). There are four steps involved in the process: *named entity recognition and classification* (NERC); *probabilistic toponym resolution* (TR); *geo-spatial indexing*; and *retrieval*.

We used a named entity recognition and classification system to differentiate between references to the names of places (which we are interested in), and the names of people and organizations (which we are not). A surprising number of everyday nouns and proper nouns are also geographic entities, for example, the town "Money" in Mississippi. Errors in this part of the pipeline can have a significant effect on the accuracy of the disambiguation process. Our system made use of the LingPipe open-source NERC system, which makes use of a Hidden Markov model trained on a collection of news articles (`http://www.alias-i.com/lingpipe/`).

For further system details, see Li et al. [2006].

Toponym Resolution

Toponym resolution (TR) is the task of assigning a location to each place name identified by the named entity recognizer. Many place names are ambiguous; context surrounding a place name in the text can be used to determine the correct candidate location. Our approach to TR assigns probability scores to each location candidate of a toponym based on the occurrence of hierarchical associations between place names in the text. Hierarchical associations and location candidates pertaining to a particular geographical reference can be found in a gazetteer resource. For this purpose, we used the Getty Thesaurus, available from `http://www.getty.edu/vow/TGNServlet`.

Probabilities are allocated to candidate locations based on a five-level normalization of the gazetteer. For example, a candidate that is classified as a *continent* or *nation* receives a significant probability, while a candidate that is classified as an *inhabited place* (which includes cities) initially receives a much smaller probability, and so on. Initial probability assignments are then adjusted based on a variety of evidence, such as:

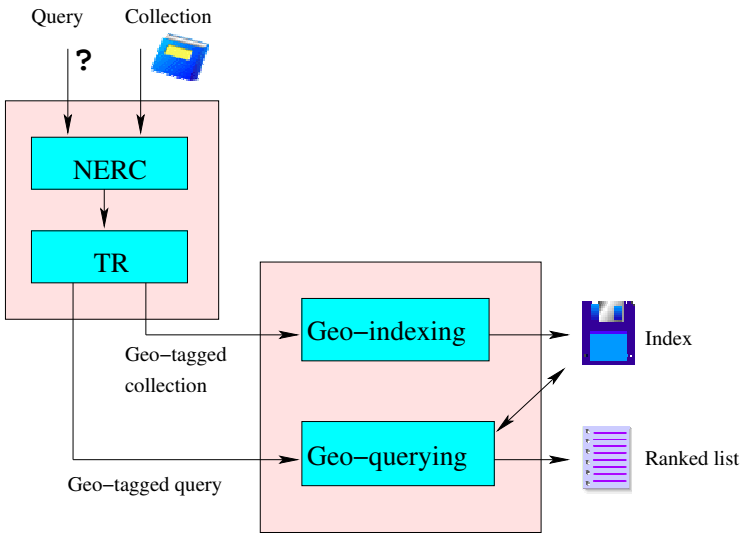

Figure 1: Components of the GIR system described in this paper.

- *Local contextual information*: for example, geo-pairs occurring in close proximity to each other, in particular, *City*, *State* pairs;
- *Population information*, when available;
- *Specified trigger words* such as “County” or “River”;
- *Global contextual information*, such as occurrences in the document of country or state geo-terms that are gazetteer ancestors to the candidate; and
- *Mutual disambiguation* across geo-terms: candidates for different geo-terms that are closely related in the gazetteer hierarchy boost each others’ probability assignment for their respective terms.

Final probability assignments are normalized across the complete set of possible candidates for each geo-term.

We used a hand annotated subset of the GeoCLEF corpus to determine the performance of the Named Entity Classification system, and our toponym disambiguation algorithm. This annotated corpus consisted of 106 Glasgow Herald and 196 LA Times news articles, which contained 2311 tagged locations in total. The overall precision of LingPipe on this dataset was 50% precision and 65% recall, that is, 1502 tags were correctly identified. With respect to disambiguation accuracy our system achieved an accuracy of 90.3% on the 1502 correctly identified LingPipe place names; this gives us an idea of how well our disambiguator would work on 100% accurately NERC tagged data. However, when disambiguation accuracy is calculated with respect to the total number of tagged locations in the dataset, we achieve an accuracy of 60.8%. This indicates that we could make significant gains by improving the performance of the NERC component of our pipeline architecture.

Probabilistic Geographical IR

Our Geographical Information Retrieval system involves an extension of Zettair [2006], to which we spatial-term indexing. Hierarchically expanded geo-terms (in each case a concatenated string consisting of a candidate and its ancestors in the gazetteer) are added to an index. Geo-tagged queries can then be processed by matching geo-terms in the query to geo-terms in the spatial index.

The system supports both *document expansion* and *query expansion* techniques for matching the location in a query to all its gazetteer children and nearby locations. Document expansion (or *redundant*

indexing) involves adding spatial terms to the index for each of a geo-term's ancestors in the gazetteer hierarchy. Query expansion involves expanding terms in the query. This technique allows more flexible weighting schemes, whereby different weights can be assigned to documents which are more relevant at different hierarchical levels or spatial distances.

A geo-term in a query expansion may be expanded *upwards* or *downwards*. Downward expansion extends the influence of a geo-term to some or all of its descendants in the gazetteer hierarchy to encompass locations that are part of, or subregions of, the specified location. Upward expansion expands the influence of a geo-term to some or all of its ancestors, and then possibly downward to siblings of these nodes. This upward-downward expansion facilitates expansion of geo-terms in a query to their nearby locations. For example, downward expansion was used for geo-terms proceeded by an "in" spatial relation, while upward expansion was used for "close/near" relations. Spatial relations such as "in_east" or "close_west" are not currently handled in our system.

After expansion, weights are assigned to all expanded geo-terms, reflecting their estimated similarities to the source query geo-term. We used *hierarchical distance* for downward expansion and *spatial distance* for upward expansion. Finally, the *a priori* Okapi BM-25 approach [Walker et al., 1997] (as implemented in Zettair) is used to calculate the sum of scores for the query. We apply a *normalization step* to obtain a single score for each location concept by combining the similarity scores of its geo-term, text term, and expanded geo-terms. Without this step, irrelevant documents that contain many of the expanded geo-terms in the query will be incorrectly favored. The contribution of the (potentially numerous) geo-terms added to an expanded query might then overwhelm the contribution of the non-geo terms in the topic.

For further details of the probabilistic geographical IR approach, see Li et al. [2006].

3 Experimental Results

All of our GeoCLEF 2006 submitted runs were based on the Zettair system, some using baseline Zettair and others using the probabilistic IR techniques described in the previous section. The runs submitted were:

1. `MuTdTxt`: Baseline Zettair system run on unexpanded queries formed from topic *title* and *description* only. All retrieval is purely text-based. We take this to be our baseline.

2. `MuTdnTxt`: Baseline Zettair system run on unexpanded queries formed from topic *title* and *description* and location words from the *narrative* field. All retrieval is purely text-based.

3. `MuTdRedn`: Baseline Zettair system run on queries formed from topic *title* and *description*. Documents are automatically toponym-resolved and expanded with related geo-terms (as described in the previous section). The query (*title* and *description*) is automatically annotated and geo-terms are disambiguated, but these terms are *not* further expanded with (hierarchically) related geo-terms.

4. `MuTdQexpPrb`: Probabilistic version of Zettair (as described above). Documents are automatically annotated and disambiguated (that is, toponym resolution is performed) but are *not* expanded with related geo-terms. Query (*title* and *description*) is automatically resolved for geo-terms, and geo-terms are expanded with related geo-terms. This is our most complete Geographic IR configuration.

5. `MuTdManQexpGeo`: Baseline Zettair using purely text-based retrieval, but for which the query is *manually* expanded (with related text place-names).

Table 1 shows the overall mean average precision (MAP) scores for the runs submitted to GeoCLEF. These scores are significantly lower than the MAP score obtained by the baseline system (`MuTdTxt`) run over the GeoCLEF 2005 topics: 0.3539.

Subsequent to submitting to GeoCLEF 2006, we made some improvements to the normalization step described in Section 2. The results of the same runs as above but using this improved system are presented in Table 2. This improvement of the normalization step results in an increase in MAP for the probabilistic runs (`MuTdRedn` and `MuTdQexpPrb`).

Figure 2 displays the Average Precision scores for each topic and for each run, after including the newer, improved normalization step. There is a high degree of variance in the performance obtained across the

	Run descriptor				
	MuTdTxt	MuTdnTxt	MuTdRedn	MuTdQexpPrb	MuTdManQexpGeo
MAP	0.2312	0.2444	0.2341	0.2218	0.2400
%Δ		+5.71%	+1.25%	−4.07%	+3.81%

Table 1: MAP scores for each submitted run over all 25 topics.

Run	MuTdTxt	MuTdnTxt	MuTdRedn	MuTdQexpPrb	MuTdManQexpGeo
MAP	0.2312	0.2444	0.2464	0.2447	0.2400
%Δ		+5.71%	+6.57%	+5.84%	+3.81%

Table 2: MAP scores over all 25 topics, using improved normalization step.

set of queries. We plan to undertake a failure-mode analysis to try and ascertain reasons why the toponym resolution did not always yield improved retrieval effectiveness.

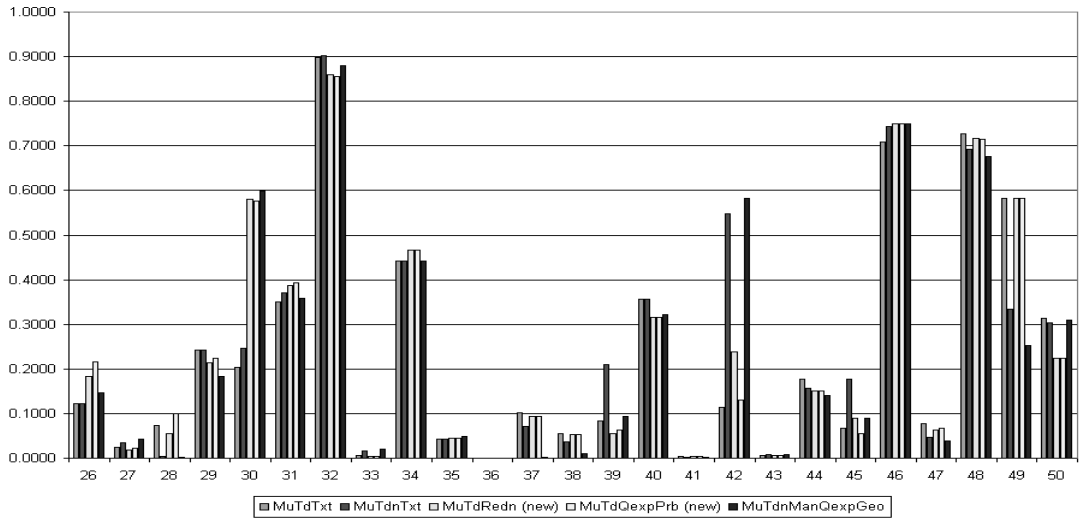

Figure 2: Average Precision per topic, for each run, using improved normalization.

4 Discussion

On several topics, Average Precision is very low for all of our runs, including some for which a significant number of documents were deemed relevant: for example, topic 33 (*International sports competitions in the Ruhr area*) had 20 documents judged as relevant, of which our baseline retrieved only 8. We are currently performing a failure analysis to determine why this is the case.

Regarding the geographical IR techniques, downward expansion of the query seems to have had a positive effect on some topics, with the normalization step seeming to have alleviated the problem with decreased precision. Topic 26 (*Wine regions around rivers in Europe*) sees an increase over both baseline and the manually-annotated-query runs. Improvement over baseline was also seen with queries involving spatial relations such as "near" (for example, topic 30, *Car bombing near Madrid*); and with geographic specialization, such as *Northern* (for example, topic 42, *Regional elections in Northern Germany*). Note, however, that our handling of spatial relations did not extend to specific distance specifications, which may have contributed to a slight drop in precision from baseline for topic 27 (*Cities within 100km of Frankfurt*).

Also, the automatic GIR approaches resulted in significantly lower AP on topic 42 as compared to using a manually expanded query.

Other topics for which we saw decreased performance for GIR compared to baseline include those involving general regions that are not expanded in the Getty Thesaurus. Regions such as the *Caucasus* (topic 39, *Russian troops in the southern Caucasus*) and *Yugoslavia* (currently synonymous with *Serbia and Montenegro* in the Getty Thesaurus, with topic 44 asking about *Arms sales in former Yugoslavia*) were not handled well. Such non-expanded (in the Getty Thesaurus) general regions were also involved in topics that saw no improvement over very low baseline AP, such as topics 41 (*Shipwrecks in the Atlantic Ocean*) and 43 (*Scientific research in New England Universities*). Topic 50 (*Cities along the Danube and the Rhine*) saw significantly decreased AP as compared to both baseline and the manually-expanded query runs; rivers without a clear central location present a special challenge. We expect topic 33 (*International sports competitions in the Ruhr area*), while involving a general region not expanded in the Getty Thesaurus, could be handled by transforming *in the X area* to the form *near X*.

Missed recognitions and misclassifications by LingPipe in the queries did not seem to be a significant problem; however, LingPipe's performance on the document collection is more likely to have compromised the GIR runs – this was noted as an area for improvement in Section 2. Interestingly, LingPipe's correct recognition of *Southeast Asia* in topic 38 (*Solar or lunar eclipse in Southeast Asia*) probably resulted in worse performance than the alternative of recognizing *Asia* as a location with the spatial specialization of *southeast* handled separately, since *Southeast Asia* does not have an expansion in the Getty Thesaurus.

Acknowledgments. Brianna Laugher, Jiawen Rong, and Daniel Walmsley made important contributions to system implementation, document preparation, and performance of experiments. National ICT Australia (NICTA) is funded by the Australian Government's "Backing Australia's Ability" initiative, in part through the Australian Research Council.

References

F. Gey and V. Petras. Berkeley2 at GeoCLEF: Cross-language geographic information retrieval of German and English documents. In F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough, editors, *GeoCLEF @ CLEF 2005: Cross-Language Geographical Information Retrieval*, Vienna, September 2005. URL `http://www.clef-campaign.org/2005/working_notes/`.

Y. Li, A. Moffat, N. Stokes, and L. Cavedon. Exploring probabilistic toponym resolution for geographical information retrieval. In C. Jones and R. Purves, editors, *SIGIR Workshop on Geographical Information Retrieval*, pages 17–22, Seattle, August 2006.

S. Walker, S. Robertson, M. Boughanem, G. Jones, and K. Sparck Jones. Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR. In *Proc. Sixth Text Retrieval Conference (TREC 6)*, Gaithersburg, Maryland, November 1997. URL `http://trec.nist.gov/pubs/trec6/papers/city_proc_auto.ps.gz`.

Zettair. The Zettair search engine, 2006. URL `http://www.seg.rmit.edu.au/zettair/index.php`.