Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval

Yi Li   Alistair Moffat   Nicola Stokes   Lawrence Cavedon

NICTA Victoria Laboratory
Department of Computer Science and Software Engineering,
The University of Melbourne, Victoria 3010, Australia

{yli8,alistair,nstokes,lcavedon}@csse.unimelb.edu.au

ABSTRACT

A key problem that arises when unstructured text is being queried is that of properly recognizing and exploiting geographical terms and entities. Here we describe a mechanism for probabilistic toponym resolution, and our experiments with the new method in the setting of the 2005 GeoCLEF queries and judgments. The new method gives improved retrieval effectiveness on a subset of the topics.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Content Analysis and Indexing – *Indexing Methods*

Keywords

Geo-spatial information retrieval, toponym resolution

1. INTRODUCTION

Geographical Information Retrieval (GIR) systems exploit geographical references in queries and document collections in order to improve retrieval effectiveness. They support enabling technologies such as a map-based display of retrieval results, and localized search. In this paper, we describe a probabilistic approach to GIR that overcomes some of the problems typically associated with disambiguation in an IR-type setting. A fundamental pre-processing step for many GIR systems is toponym recognition and resolution. But precise disambiguation can be a vexing task, even for humans. Instead, we argue for a probabilistic disambiguation, where the balance of probabilities is shifted between competing alternative groundings, but none is completely eliminated. The assigned probabilities are then used as weighting factors in the similarity computation used to rank the documents in the collection.

The new mechanism has been implemented and tested using the topics and judgments developed for the 2005 pilot track at GeoCLEF, see `http://ir.shef.ac.uk/geoclef/2005/`. The new method gives improved retrieval effectiveness on the subset of the topics for which it is applicable.

2. SYSTEM DESCRIPTION

Figure 1 shows the architecture of our approach to probabilistic 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 use 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. We are currently making use of the LingPipe open-source NERC system: a Hidden Markov modeling approach trained on a collection of news articles (`http://www.alias-i.com/lingpipe/`). Sections 3 and 4 provide a detailed explanation of the remaining system components.

3. TOPONYM RESOLUTION

Toponym resolution is the task of assigning a location to an ambiguous place name. As a task it is similar to word sense disambiguation in that the context surrounding the place name in the text is used to determine its exact geographical coordinates. We use a novel knowledge-based approach to TR that 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

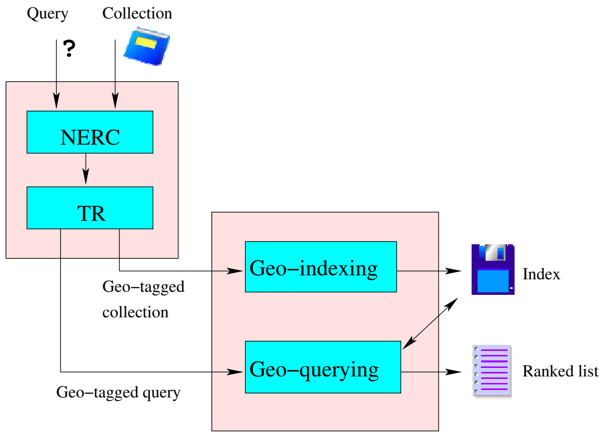

Figure 1: Components of the GIR system described in this paper.

to a particular geographical reference can be found in a gazetteer resource.

The Getty Thesaurus, available from `http://www.getty.edu/vow/TGNServlet`, was used as an aid in our experiments. For example, Getty indicates that the geo-term "Dublin" corresponds to twenty-four distinct locations in four countries: Ireland, the United States, Canada and Australia. Our algorithm assigns initial probabilities to each location candidate based on the significance level of its corresponding location type as defined by the gazetteer. More specifically, since the location type for "Dublin, Ireland" is *capital city*, it receives a higher probability than the other location candidates, all of which are classified lower in the tree-structured hierarchy as being simply *inhabited places*.

Once initial probabilities have been assigned to all location candidates in the text, our TR approach uses the following external and contextual knowledge to boost the probabilities of the most likely candidates in a given document.

- *Local contextual information*: identifying geo-term pairs that occur in close proximity to each other in the text (for example, in the same sentence) provides useful evidence in the disambiguation process, especially in the news domain [Amitay et al., 2004; Garbin and Mani, 2005; Smith and Mann, 2003]. For example, the occurrence of "Washington, D.C." unequivocally means that the writer was not referring to Washington State.

- *Population statistics*: locations that have, according to the World Gazetteer (`http://www.world-gazetteer.com/`), high populations, are considered to be the more likely candidates.

- *Geographical trigger words*: geo-terms that are preceded or succeeded by a geographical trigger word such as "county" or "lake", will have the probability of the appropriate location candidate boosted.

- *Global contextual information*: global geo-spatial information such as the occurrence of countries or states can be used to boost location candidates if the document makes reference to one of its ancestors in the hierarchy. For example, a reference elsewhere in the document to "Australia", which is an ancestor of "Melbourne", will boost the estimated probability that an appearance of "Melbourne" in this document refers to a location in Victoria rather than in Florida.

Once all location candidates in a document have been processed in this manner, the final probability assignments are normalized across the complete set of possible candidates for each particular geo-term. None of these candidates are eliminated entirely. Instead, the improbable ones end up with non-zero, but small, weights.

At this stage there are no public evaluation resources for toponym resolution; and an important element in our future research will be the development of an appropriate evaluation methodology. Having such a resource will allow us to measure the effect of errors in toponym resolution on end-to-end GIR performance.

The next section describes how the probabilistic geo-spatial assignment is integrated into the GIR retrieval process.

4. PROBABILISTIC GEOGRAPHICAL IR

The next step in the pipeline, after the probabilistic annotation of the collection has been performed, is to build the spatial index. We define spatial indexing as the process of introducing spatial concepts into the usual term index as queryable concepts, thus allowing spatial relationships to be queried or otherwise exploited. In our

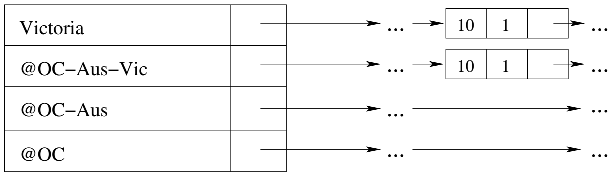

Figure 2: Inverted lists without document expansion.

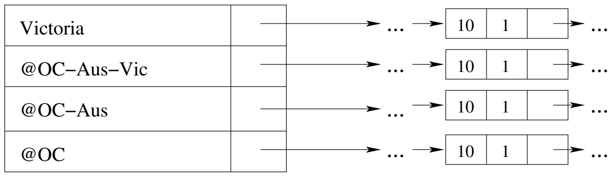

Figure 3: Inverted lists with document expansion.

experiments, we add *geo-terms* into a baseline text retrieval system Zettair [2006]. For a given reference identifier in the gazetteer, the corresponding geo-term in the spatial index has the following hierarchical format:

```
@CONTINENT/LOC1
 -NATION/LOC2
 -STATE/PROVINCE/LOC3
 -CITY/TOWN/COUNTY/LOC4
 -OTHER/LOC5
```

For example, the geo-term `@OC-Aus-Vic-Melbourne` represents the location Melbourne, Victoria, Australia (in Oceania). That is, a geo-term is a string concatenation of its Getty gazetteer identifier with those of its ancestors.

Once the spatial index has been built (off-line), geo-tagged queries can be submitted to the system. A similar retrieval and ranking method is used for text terms and geo-terms with the exception that geo-terms in the query are only searched against geo-terms in the spatial part of the index. The hierarchical geo-term structure makes it easy to expand each location in the query to all its children and nearby locations. Currently, our system supports two distinct geo-term expansion methods: *document expansion* and *query expansion*.

Document Expansion. Document expansion (or *redundant indexing*) is the addition of a posting to all the geo-terms which are ancestors of each expanded geo-term in a given document. For example, the geo-term `@Victoria` in (say) document 10 is indexed as if document 10 contained explicit mention of all of `@Victoria`, `@Australia`, and `@Oceania` (Figures 2 and 3). The document expansion occurs during the indexing process.

The main advantage of document expansion is that it reduces query processing time by eliminating the need for query expansion. However, document expansion also has a number of disadvantages. Firstly, it increases the size of the term index. For example, assuming that place names account for 5% of the terms in a collection, the size of the index will increase by 25% if the expansion of each place name results in the addition of five geo-terms. Secondly, it is harder to assign different weights to the expanded terms, because these redundant geo-terms were added into the postings during indexing. Thirdly, document expansion does not support directional geo-queries such as "cities 100km from Frankfurt" or "in the north-

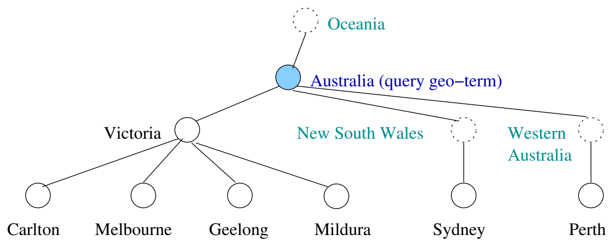

Figure 4: Downward expansion. The collection contains references to all of the locations indicated by solid circles.

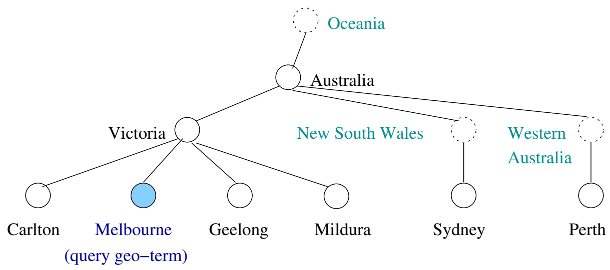

Figure 5: Upward expansion. The collection contains references to all of the locations indicated by solid circles.

east of Iraq". These queries should be processed at query time, for example, using a constrained expansion technique based on the distances between locations calculated using geographical coordinates found in Getty.

Query Expansion. The second method, query expansion, has a number of advantages over document expansion: query expansion doesn't affect the size of the index, and it can be adapted to support directional geo-queries. It also allows more flexible weighting schemes, in which different weights can be assigned to documents which are relevant at different hierarchical levels or spatial distances.

Using the hierarchical structure, a geo-term can be expanded in two directions: *downward* and *upward*. Downward expansion extends the influence of a geo-term to some or all of its descendents in the hierarchical gazetteer structure, to encompass locations that are part of, or subregions of, the location specified in the query. For example, in Figure 4 the geo-term `@Australia` might be expanded to seven geo-terms.

Upward expansion extends the influence of a geo-term to some or all of its ancestors, and then possibly downward again into other siblings of those nodes. This facilitates the expansion of geo-terms in the query to their nearby locations. For example, in Figure 5 the geo-term `@Melbourne` might be expanded to reach all seven other locations that appear in the index.

In the experiments described in Section 5, downward expansion is used for geo-terms proceeded by an "in" spatial relation, while upward expansion is used for the "close/near" relations. For relations such as "in_east", "close_west", and so on, the coordinates of the expanded geo-terms in the gazetteer are compared with those of the query geo-term in order to remove all outlying, irrelevant locations.

After the expansion, we need to rank or assign weights between zero and one to all expanded geo-terms, to reflect their estimated similarities to the source query geo-term. This similarity is calculated from the *hierarchical distance*, the *overlapping area ratio*, or the *spatial distance*. In our experiments we use the hierarchical distance for downward expansion, and the spatial distance for upward expansion. For example, for the indicated query geo-term `@Australia` in Figure 4, the hierarchical distances of the geo-terms resulting from a downward expansion are "1" to `@Victoria`, and "2" to `@Melbourne`. Since geo-terms with a shorter hierarchical distance are considered more similar, `@Victoria` is assigned a higher weight than `@Melbourne`.

Looking now at an example of upward expansion, in the arrangement shown in Figure 5 the query geo-term `@Melbourne` is expanded to `@Victoria`, and from there to `@Carlton`, `@Geelong`, and `@Mildura`; and also potentially as far afield as `@Sydney` and `@Perth`. According to the coordinates in the gazetteer, `@Carlton` is closer to `@Melbourne` than any of the others, and is accordingly assigned the highest weight. By employing the hierarchical and spatial similarity relationships we ensure that documents in which geo-terms with a higher similarity value appear, out-rank other documents containing lower value geo-terms.

In our experiments, we use the *a priori* version (without relevance information) of Okapi BM-25 [Walker et al., 1997], as implemented in Zettair. The BM-25 approach calculates the sum of the scores for each term in the query. In order to calculate the score of the location concept in the query, we first calculate the similarity of each geo-term by multiplying its original Okapi score with its hierarchical or spatial similarity, and the document and query probabilities from the toponym resolution step. We then use a normalization algorithm to get a single score for the location concept by combining the similarity scores of its geo-term (`@Australia`), text term (Australia) and expanded geo-terms (such as `@Melbourne` and `@Sydney`). Without this step, an irrelevant document which contains many of the expanded geo-terms in the query will be incorrectly favored.

Rauch et al. [2003] also looked at a probabilistic framework for integrating disambiguation confidence scores into the retrieval process. However, this work did not consider expanded geo-terms when calculating the similarity between the query and a document.

5. EVALUATION

In this section we outline the evaluation methodology provided by the GeoCLEF retrieval task, and then describe our experimental results on the GeoCLEF 2005 topics. GeoCLEF is part of the larger Cross Language Evaluation Forum (CLEF): an annual European TREC-style event concerned with issues related to cross-language IR. Despite GeoCLEF's emphasis on multi-lingual retrieval, our primary interest here is on the results from the mono-lingual English track.

In the mono-lingual GeoCLEF track participants were provided with a list of 25 geo-spatial topics, and a newswire collection consisting of around 16,000 articles from the *Glasgow Times* and the *LA Times*. Each of the 25 topics consists of a title, a longer description, and a narrative part providing additional clues about the set of documents being sought as "answers". In addition, the topic concept (for example, "murders and violence"), a spatial relation (perhaps "in the south-west of") and one or more geo-term(s) (to complete the same example, "Scotland") are also explicitly listed.

Retrieval systems are evaluated using the standard metrics of mean average precision (MAP) and precision at depth 20 (P@20). Four runs were carried out, building queries from the title and description fields of the GeoCLEF topics:

- `TdBaseline` is a Zettair baseline run; that is, without geo-

Run	MAP		P@20	
`TdBaseline`	.3539		.3780	
`TdGeoQexp`	.3540	+0.03%	.3820	+1.06%
`TdGeoDexp`	.3540	+0.02%	.3860	+2.12%

Table 1: Retrieval effectiveness (MAP and P@20) over all 25 GeoCLEF 2005 topics.

spatial document or query expansion.

- `TdGeoDexp` is a run using the same queries, but now with document expansion (redundant indexing).
- `TdGeoQexp` is a run using query expansion instead.

Table 1 shows the results of these initial runs. Both the Geo-IR experiments (`TdGeoQexp` and `TdGeoDexp`) achieve an imperceptible gain in performance with respect to the baseline run. Although these results are somewhat disappointing, it was noted by many of last year's GeoCLEF participants that the topics focused on the expansion of large land masses such as Europe, and countries paired with topics that are unlikely to require geo-expansion such as "Genocide in Rwanda". These types of geo-spatial queries may not be the most receptive benefactors of spatially-aware technologies. The recently released GeoCLEF 2006 topics address some of these concerns. We expect that additional experiments using these topics will allow a more accurate analysis of the performance gains (or losses) accruing through the use of expansion, and will allow us to refine the techniques we have been using.

Although the average MAP scores in Table 1 imply that our expansion techniques are having little or no effect on the retrieval process, in the following Section, we will show that the Geo-IR runs perform better than the baseline on certain topics.

6. DISCUSSION

Table 2 shows the MAP score and corresponding absolute performance difference between the Geo-IR and the baseline runs for each of the 25 topics. Figure 6 shows that the performance of the Geo-IR systems is highly correlated across the topics. In addition, the cluster of points around the origin indicates that any performance gains or loses achieved by the Geo-IR runs are small. The two outlying points represent the first and last entries in Table 2. We anticipate more diversity between these systems on queries that contain the "close to" relation since the document expansion mechanism only performs downward expansion and this spatial relation requires upward expansion. However, with the exception of queries GC025, all other GeoCLEF topics use the "in" spatial operator. This explains the similarity of the Geo-IR runs.

We conducted a detailed analysis of the ranked lists returned by each system in order to determine why certain topics benefited from our expansion techniques while others didn't. In general, many of the gains and loses observed can be attributed to document re-ranking rather than the introduction of novel "judged" documents to the ranked retrieval list. That is, by adding geo-terms to a query, the final document score is adjusted by the contribution of the query geo-terms and their expanded geo-terms.

During this analysis we identify the following beneficial side-effects of geo-spatial term expansion on the retrieval process:

1. *Synonym effect*: Referencing geo-terms using their Getty id, enables our Geo-IR runs to pick up synonymous relationships between query and document geo-terms. For example, for topic GC010, "Flooding in Holland and Germany", there are relevant documents in the collection (such as GH950201-000048) that mention the "Netherlands" but not "Holland".

2. *Phrasal effect*: Topic GC014 is about "Environmentally hazardous incidents in the North Sea". Both "north" and "sea" are very common terms in the document index, but the phrase `@North Sea` is not. By adding this phrasal geo-term into our query, documents talking about the North Sea but not "north" or "sea" are boosted.

3. *Query expansion effect*: In general, the performance of topics containing the geo-term and spatial relation "in Europe" improved. Similarly, directional queries such as GC022, "Restored buildings in Southern Scotland" that contain a compass point constraint, respond well to expansion. For example, many of the relevant documents to this query are from the Glasgow Times. Hence, references to cities such as Glasgow and Edinburgh are not grounded by "Scotland" because it is expected that the reader's geographical knowledge will allow them to readily disambiguate these place names. This shows that geo-expansion has a significant role to play in a retrieval scenario, where local news sources are being searched by a user who is unfamiliar with the region.

Our analysis also indicated that many GeoCLEF 2005 queries do not benefit from query expansion simply because relevant documents in the collection made reference to these geo-terms. For example, documents relevant to query GC005 on "Japanese rice exports", always mention Japan. In the second half of Table 2, we have also identified 4 topics that experience major gains in performance when their geo-terms are dropped from their query. The average MAP scores for a baseline run on these topics using the concept part of the query only is shown in Table 3. Obviously, we cannot expect our Geo-IR runs to make any gains on these topics.

Other performance drops can be attributed to poor Getty location coverage. For example, some topics such as GC020 contain geo-terms such as "Scottish Islands" that are not listed in the gazetteer. In addition, some locations in Getty cannot be expanded because their geographical coordinates, and related children, are not listed. These locations are members of the "general region" location type

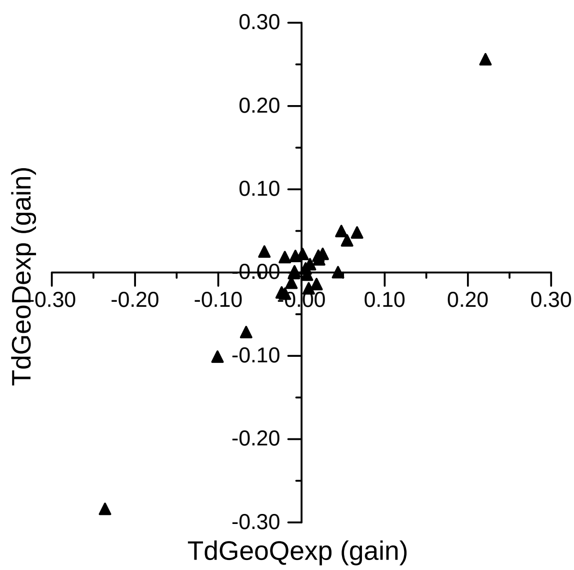

Figure 6: Correlation of MAP scores for Geo-IR runs.

Topic	TdBaseline	TdGeoQexp		TdGeoDexp		spatial query terms
GC010	.5612	.7823	+0.2211	.8174	+0.2562	in Holland and Germany
GC023	.0306	.0974	+0.0668	.0788	+0.0482	in Southwest Scotland
GC013	.5328	.5876	+0.0548	.5715	+0.0387	in Germany
GC022	.3600	.4079	+0.0479	.4099	+0.0499	in Southern Scotland
GC019	.1554	.1993	+0.0439	.1559	+0.0005	in Europe
GC015	.6810	.7065	+0.0255	.7035	+0.0225	in Rwanda
GC018	.3037	.3249	+0.0212	.3196	+0.0159	in Scotland
GC014	.2815	.3018	+0.0203	.3018	+0.0203	in and close to North Sea
GC006	.3656	.3837	+0.0181	.3518	−0.0138	in Europe
GC016	.8608	.8710	+0.0102	.8709	+0.0101	in Siberia and Caspian Sea
GC002	.0666	.0753	+0.0087	.0477	−0.0189	in Europe
GC008	.0399	.0465	+0.0066	.0374	−0.0025	in Europe
GC024	.5887	.5937	+0.0050	.5937	+0.0050	in Scottish Highlands
GC009	.4240	.4253	+0.0013	.4464	+0.0224	in Asia
GC007	.0991	.0918	−0.0073	.1191	+0.0200	in Europe
GC011	.0467	.0383	−0.0084	.0482	+0.0015	in UK and Germany
GC004	.1853	.1762	−0.0091	.1847	−0.0006	in Europe and USA
GC001	.6832	.6711	−0.0121	.6710	−0.0122	off Australia and California
GC005	.5842	.5643	−0.0199	.5589	−0.0253	in Japan
GC020	.3593	.3393	−0.0200	.3779	+0.0186	in islands of Scotland
GC003	.0391	.0153	−0.0238	.0153	−0.0238	in Latin America
GC012	.1603	.1157	−0.0446	.1856	+0.0253	in Europe, United Kingdom and Russia
GC017	.4156	.3491	−0.0665	.3442	−0.0714	in Sarajevo, Bosnia-Herzegovina
GC021	.5639	.4630	−0.1009	.4630	−0.1009	in North Sea
GC025	.4583	.2222	−0.2361	.1746	−0.2837	in and close to Trossachs, Scotland

Table 2: Retrieval effectiveness (MAP) for 25 topics.

Topic	TdBaseline	ConceptOnly		concept terms
GC003	.0391	.3963	+0.3572	Amnesty International
GC004	.1853	.4670	+0.2817	Fur Industry
GC007	.0991	.4961	+0.3970	Trade Unions
GC012	.1603	.2108	+0.0505	Cathedrals

Table 3: Retrieval effectiveness (MAP) for concept dominated topics.

and include toponyms such as "Siberia", "Latin America" and the "North Sea". In the next Section, we discuss the significance of these results within the context of the official GeoCLEF 2005 results.

7. RELATED WORK

For GeoCLEF 2005, participants submitted two mandatory runs: one using the title and description fields of the topic (the TD run); and the other using the title, description and contents of the geographical tags (TDG run). Our baseline `TdBaseline` system would have ranked second overall for the TD run; while the Geo-IR runs `TdGeoQexp` and `TdGeoDexp` would have ranked second for the TDG run.

Looking at the top performing systems at GeoCLEF, no system gained any significant advantage over their baseline run by exploiting the geo-spatial information provided in the topics [Gey et al., 2005b]. For example, the top performing TGN run submitted by Berkeley2 [Petras and Gey, 2005] achieved a MAP of 0.3937 and a baseline TD score of 0.3613. Interestingly, the Berkeley2 TNG run didn't employ any Natural Language Processing techniques. Instead the geo-terms were simply added to the original TD query, which increased the weight of importance of these toponyms during the ranking process. Their system then performed a second retrieval on the collection; this time augmenting their initial query with the top thirty most frequent terms in the documents from the initial ranking. This blind feedback method proved to be very effective at boosting performance, which illustrates the benefit of simple topic expansion to geo-spatial queries.

There were, however, a number of participants that made use of toponym resolution and expansion methods [Cardoso et al., 2005; Ferrndez et al., 2005; Ferrs et al., 2005; Lana-Serrano et al., 2005; Leidner, 2005; Leveling et al., 2005]. Our GIR approach differs from theirs in two important ways: the probabilistic toponym resolution and expansion methods, and an expansion normalization algorithm that treats each query geo-term and its expanded geo-terms as one term. This query normalization step, as already stated, was intended to minimize the influence of documents which contain many of the expanded geo-terms but are not relevant to the overall topic.

In conclusion then, it is our hope that future GeoCLEF tracks will include more directional queries and queries containing the spatial relation "close" as we believe that they stand to gain most from the Geo-IR techniques discussed in this paper.

Acknowledgments. 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 Center of Excellence programs.

References

E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging web content. In *Proc. SIGIR '04*, pages 273–280, 2004.

N. Cardoso, B. Martins, M. S. Chaves, L. Andrade, and M. Silva. The XLDB group at GeoCLEF 2005. In Gey et al. [2005a]. URL http://www.clef-campaign.org/2005/working_notes/.

O. Ferrndez, Z. Kozareva, A. Toral, E. Noguera, A. Montoyo, R. Muoz, and F. Llopis. The University of Alicante at GeoCLEF 2005. In Gey et al. [2005a]. URL http://www.clef-campaign.org/2005/working_notes/.

D. Ferrs, A. Ageno, and H. Rodrguez. The GeoTALP-IR system at GeoCLEF-2005: Experiments using a QA-based ir system, linguistic analysis, and a geographical thesaurus. In Gey et al. [2005a]. URL http://www.clef-campaign.org/2005/working_notes/.

E. Garbin and I. Mani. Disambiguating toponyms in news. In *Proc. Human Language Technology Conference (HLT-EMNLP'05)*, pages 363–370, Vancouver, BC, October 2005. ACL.

F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough, editors. *GeoCLEF @ CLEF 2005: Cross-Language Geographical Information Retrieval*, September 2005a. URL http://www.clef-campaign.org/2005/working_notes/.

F. Gey, R. Larson, M. Sanderson, H. Joho, and Clough P. GeoCLEF: The CLEF 2005 cross-language geographic information retrieval track. In Gey et al. [2005a]. URL http://www.clef-campaign.org/2005/working_notes/.

S. Lana-Serrano, J. Goi-Menoyo, and J. Gonzlez-Cristbal. MIRACLEs 2005 approach to geographical information retrieval. In Gey et al. [2005a]. URL http://www.clef-campaign.org/2005/working_notes/.

J. L. Leidner. Preliminary experiments with geo-filtering predicates for geographic IR. In Gey et al. [2005a]. URL http://www.clef-campaign.org/2005/working_notes/.

J. Leveling, S. Hartrumpf, and D. Veiel. University of Hagen at GeoCLEF 2005: Using semantic networks for interpreting geographical queries. In Gey et al. [2005a]. URL http://www.clef-campaign.org/2005/working_notes/.

V. Petras and F. Gey. Berkeley2 at GeoCLEF: Cross-language geographic information retrieval of German and English documents. In Gey et al. [2005a]. URL http://www.clef-campaign.org/2005/working_notes/.

E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proc. HLT-NAACL'03 Workshop on Analysis of Geographic References*, pages 50–54, 2003.

D. Smith and G. Mann. Bootstrapping toponym classifiers. In A. Kornai and S. Sundheim, editors, *Proc. HLT-NAACL'03 Workshop on Analysis of Geographic References*, pages 45–49, 2003.

S. Walker, S. Robertson, M. Boughanem, G. Jones, and K. S. Jones. Okapi at TREC-6 automatic ad hoc, VLC, routing, filtering and QSDR. In *Proc. Third Text Retrieval Conference (TREC 3)*, 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.