NICTA I2D2 Group at GeoCLEF 2006

Yi Li Nicola Stokes Lawrence Cavedon Alistair Moffat

National ICT Australia, Victoria Research 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, as well as post-GeoCLEF evaluations and improvements to the submitted system. In particular, we used techniques to assign probabilistic likelihoods to geographic candidates for each identified geo-term, and a probabilistic IR engine. A *normalisation* process that adjusts term weights, so as to prevent expanded geo-terms from overwhelming non-geo terms, is shown to be crucial.

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 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 is built on top of the Zettair [1] IR engine, extending the base engine with a probabilistic retrieval technique to handle ambiguous geographic references. The documents are pre-processed with Language Technology components to identify and resolve geo references: we use the LingPipe *named entity recognition and classification* system; and a *toponym resolution* component. The toponym resolution component assigns probabilistic likelihoods to potential candidate locations, found in a gazetteer, for geographic terms identified by LingPipe.

This paper describes the system architecture and a number of experiments performed for the purpose of GeoCLEF, as well as evaluating different approaches to geospatial-retrieval. In particular, 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, [2]), 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

* NICTA is funded by the Australian Government's "Backing Australia's Ability" initiative, in part through the Australian Research Council.

refined version saw a slight increase in overall MAP over the non-GIR baseline run. During these post–submission experiments we also noticed that our linguistic annotation was having an unpredicatable effect on document rankings, due to the increase in average document length (in bytes) which is used by the Okapi BM–25 ranking method [3]. This problem is also address in this paper, and is described in Section 3.

Overall, our baseline for GeoCLEF 2006 topics seemed low (MAP score of 0.2312); adding the geographic retrieval techniques increased overall MAP by 1.86% (document expansion) and 3.11% (query expansion) over the baseline. Variation in performance was seen across topics; this is discussed in Section 4.

2 System Description

There are four steps involved in our probabilistic geospatial information retrieval (GIR) process: *named entity recognition and classification* (NERC); *probabilistic toponym resolution* (TR); *geo-spatial indexing*; and *retrieval*. For maximum flexibility and easy re-configuration, we used the UIMA[1] document processing architecture to augment documents with extra annotations and to perform the indexing.

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 uses the LingPipe open-source NERC system, which employs a Hidden Markov model trained on a collection of news articles (`http://www.alias-i.com/lingpipe/`). For further GIR system details, see [4].

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

[1] `http://www.research.ibm.com/UIMA/`

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: *local contextual information*, for example, geo-terms occurring in close proximity mutually disambiguate each other, in particular, city–state pairs; *population information*, when available; *specified trigger words* such as "County" or "River"; and *global contextual information*, such as occurrences in the document of "country" or "state" type geo-terms that are gazetteer ancestors to the candidate. Final probability assignments are then 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 302 tagged GeoCLEF news articles; a total of 2207 tagged locations. The overall precision of LingPipe on this dataset was 46.88% precision and 69.77% recall. With respect to disambiguation accuracy our system achieved an accuracy of 71.74%. Using additional geographical evidence from Wikipedia we were able to increase the accuracy to 82.14%. A detailed description of this experiment, as well as further detail on the impact of NLP errors, is presented in [5].

Probabilistic Geographical IR

Our Geographical Information Retrieval system involves an extension of Zettair [1], to which we add 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 may be expanded either *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. For example, downward expansion was used for geo-terms proceeded by an "in" spatial relation, while upward expansion was used for "close/near" relations.

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 [3] (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. Our ranking algorithm works as follows.

In a query, all query terms are divided into two types: *concept terms* t_c and *location terms* (annotated with reference to gazetteers) t_l. In a query example "wine Australia", the concept term is "wine" and the location term is "Australia". The final score is contributed to by both concept and locations terms:

$$sim(Q, D_d) = sim_c(Q, D_d) + sim_l(Q, D_d) \tag{1}$$

where $sim_c(Q, D_d)$ is the concept similarity score and $sim_l(Q, D_d)$ is the location similarity score. The concept similarity score is calculated using the same calculation as Okapi. The location similarity score can then be denoted as:

$$\begin{aligned} sim_l(Q, D_d) &= \sum_{t \in Q_l} sim_t(Q, D_d) \\ &= \sum_{t \in Q_l} Norm_t\Big(sim_{text}(Q, D_d), \\ &\qquad sim_{geo1}(Q, D_d), \ldots, sim_{geoT}(Q, D_d)\Big) \end{aligned} \tag{2}$$

where Q_l is the aggregation of all location terms in the query, T is the number of all corresponding geo-terms (including expanded geo-terms) of a location $t \in Q_l$, and $Norm_t()$ is a normalization function which normalizes the similarity scores of a location t's textual terms, geo-term and its expanded geo-terms.

To define the normalization function, assume that we have T similarity scores from terms belonging to the same location (text, geo-term and expanded geo-terms), and after sorting them into descending order, they are: sim_1,..., sim_T. We use a geometric progression to compute the final normalization score:

$$Norm(sim_1, \ldots, sim_T) = sim_1 + \frac{sim_2}{a} + \cdots + \frac{sim_T}{a^{T-1}} \quad (a > 1) \tag{3}$$

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. 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.

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, geo-terms are disambiguated, but geo-terms are *not* further expanded with related geo-terms.
4. `MuTdQexpPrb`: Probabilistic version of Zettair. Documents are automatically annotated and disambiguated 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 overall mean average precision (MAP) scores for runs submitted to GeoCLEF. These scores are significantly lower than the MAP score obtained by the baseline system (`MuTdTxt`) run over GeoCLEF 2005 topics: 0.3539.

Run	`MuTdTxt`	`MuTdnTxt`	`MuTdRedn`	`MuTdQexpPrb`	`MuTdManQexpGeo`
MAP	0.2312	0.2444	0.2341	0.2218	0.2400
$\%\Delta$		+5.71%	+1.25%	−4.07%	+3.81%

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

Subsequent to submitting to GeoCLEF 2006, we made some improvements to the normalization step described in Section 2. Further, we discovered, after annotation, that documents have different changes in length ratio, depending on the density of the spatial references they contain. This change has an impact on the final ranking of these documents. To evaluate the impact of different length normalization strategies, we performed some further experiments. First, the unannotated collection was indexed. The average document length was 3532.7 bytes. We then indexed an annotated version of the collection, in which the average document length was 4037.8 bytes. Finally, we executed 25 topics against these two indexes using only text terms, and found large variations in the average precision value differences (standard deviation of 15.8%). To counteract the effect of byte length variations, all the annotated documents should use the same document byte length as their baseline unannotated documents when indexed. We used this technique to index the same annotated collection. From Figure 1 we can see that for most of the topics their MAP is not changed at all, the variations in the average precision value differences obtained is very small with a stadand deviation of only 0.71%.

We re-ran our two geographic runs again using both improved normalization and baselined document length. The results are provided in Table 2.

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 set of queries. To discover the performance of our query normalization, we also performed a further run without normalization. Its overall MAP is 0.1994, which is a significant decline.

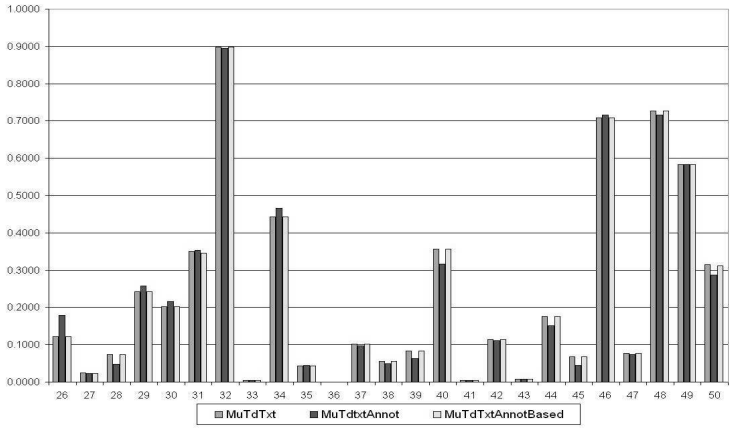

Fig. 1. Average Precision per topic, for each run, of three different text-only runs. `TTxt`: unannotated documents; `TTxtAnnot`: annotated documents with different document length; `TTxtAnnotBased`: annotated documents with baselined document length.

Run	`MuTdTxt`	`MuTdnTxt`	`MuTdRedn`	`MuTdQexpPrb`	`MuTdManQexpGeo`
MAP	0.2312	0.2444	0.2355	0.2384	0.2400
$\%\Delta$		+5.71%	+1.86%	+3.11%	+3.81%

Table 2. MAP scores over all 25 topics, using improved normalization and baselined document length.

We also performed a non-probabilistic run whose overall MAP is 0.2388, indicating little difference between probabilistic and non-probabilistic techniques.

4 Analysis and Conclusions

One of the underlying assumptions of our current GIR architecture is that every query that makes reference to a geographical entity requires geospatial expansion. From a detailed analysis of our results we observed that this is not always the case. For example, topic 28 (*Snowstorms in North America*) benefits more from concept expansion ("snow", "storm") than geo-term expansion. The frequent occurrence of queries of this nature in the GeoCLEF topics may explain the high performance of GeoCLEF systems that disregard the geographical nature of the queries and expand instead through relevance feedback methods [2]. It may also explain why our manual expansion run did not significantly outperform the baseline run, as the human annotator added geo-terms only.

However even when queries are suitable candidates for geospatial expansion, the query expansion methods describe in this paper often do not perform as expected. One reason for this is that our gazetteer lacks some vital information

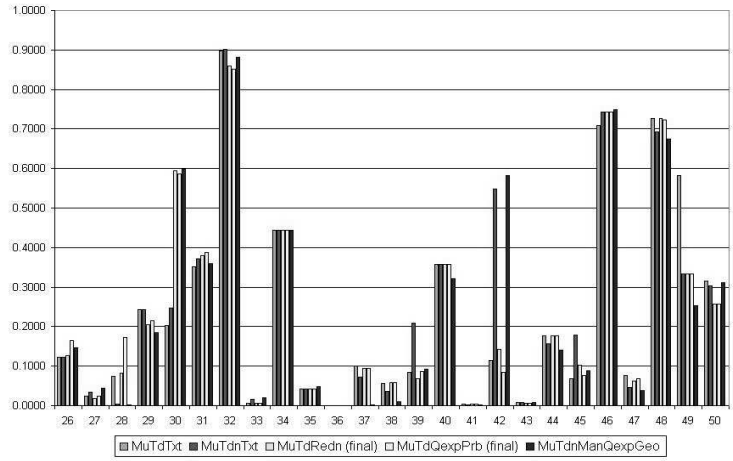

Fig. 2. Average Precision per topic, for each run, using improved normalization and baselined document length.

needed for geospatial expansion. For example, topics that mention politically defunct locations, such as "former Yugoslavia" and "the eastern Bloc", are not listed by the Getty thesaurus. In addition, locations that fall under the "general region" category in Getty, such as "Southeast Asia", "the Ruhr area" and "the Middle East", cannot be expanded as neither their related children nor their associated long/lat coordinates are listed in the gazetteer. This explains the poor to average performance of our geo–runs on topics 33, 34, 35, 37, 38, and 44.

Another assumption of our GIR architecture is that GeoCLEF queries that will benefit from geospatial expansion only require the additional concepts provided by a standard gazetteer; that is, neighboring locations and synonyms. Some GeoCLEF queries require adjective–form expansion of their geo-terms; this is particularly important when a document mentions the concept of the query (for example, "diamonds" in topic 29) but alludes to the location of the event through the use of a geographical adjective (for example, "a South African-based organization"). Examples such as these indicate that metonymic references to locations do not necessarily reduce the performance of GIR systems, as was implied in [6]. The adjective form of place names can be captured from a thesaural resource, such as WordNet.

In addition, some geospatial queries require entity type expansion that is constrained by geographic boundaries. For example, in topic 43 (*Scientific research in New England universities*), ignoring the syntactic relationship between the location "New England" and the entity "universities", will ensure that documents that mention "Harvard" will appear less relevant, since Harvard isn't (by coincidence) a state, county, town or suburb in New England. Information

such as this may only be found using an additional knowledge source such as Wikipedia.

Despite these drawbacks, expansion of the query has had a positive effect on some topics, with the normalization step seeming to have alleviated the query overloading problem. 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*).

Missed recognitions and misclassifications by LingPipe, and incorrectly disambiguated locations in the queries and collection, will also have compromised the performance of our GIR runs. In additional experiments, not discussed in this paper, we have found that NERC systems such as LingPipe are significantly underperforming on GeoCLEF relative to the performance scores reported at evaluation forums such as MUC;[2] re-training of off-the-shelf systems for GeoCLEF will significantly reduce such errors. However, some NERC errors are more critical than others; maximizing NERC recall rather than precision will improve the quality of the annotated data presented to the GIR system, as the toponym resolution step filters out much of the NERC misclassified named entities. Further discussion of this work can be found in [5].

References

1. The Zettair search engine (2006) http://www.seg.rmit.edu.au/zettair/index.php.
2. Gey, F., Petras, V.: Berkeley2 at GeoCLEF: Cross-language geographic information retrieval of German and English documents. In: GeoCLEF 2005 Working Notes, Vienna (2005) http://www.clef-campaign.org/2005/working_notes/.
3. Walker, S., Robertson, S., Boughanem, M., Jones, G., Sparck Jones, K.: Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR. In: Proc. Sixth Text Retrieval Conference (TREC 6), Gaithersburg, Maryland (November 1997)
4. Li, Y., Moffat, A., Stokes, N., Cavedon, L.: Exploring probabilistic toponym resolution for geographical information retrieval. In: SIGIR Workshop on Geographical Information Retrieval, Seattle (2006)
5. Stokes, N., Li, Y., Moffat, A., Rong, J.: An empirical study of the effects of NLP components on Geographic IR performance. Technical report, National ICT Australia Technical Reports Series (2006)
6. Leveling, J., Hartrumpf, S.: On metonymy recognition for geographic ir. In: SIGIR Workshop on Geographical Information Retrieval, Seattle (2006)

[2] `http://www-nlpir.nist.gov/related_projects/muc/`