UB at GeoCLEF 2006

Miguel E. Ruiz [(1)], Stuart Shapiro [(2)], June Abbas [(1)], Silvia B. Southwick [(1)] and David Mark [(3)]
State University of New York at Buffalo
[(1)]Department of Library and Information Studies
[(2)] Department of Computer Science and Engineering
[(3)] Department of Geography
E-mail: meruiz@buffalo.edu

Abstract

This paper summarizes the work done at the State University of New York at Buffalo (UB) in the GeoCLEF 2006 track. The approach presented uses pure IR techniques (indexing of single word terms as well as word bigrams, and automatic retrieval feedback) to try to improve performance of queries with geographical references. The main purpose of this work is to identify the strengths and shortcomings of this approach so that it serves as basis for future development of a geographical reference extraction system. We submitted four runs to the monolingual English task, 2 automatic runs and two manual runs, using the title and description fields of the topics. Our official results are above the median system (auto=0.2344 MAP, manual=0.2445 MAP). We also present an unofficial run that uses title description and narrative which shows a 10% improvement in results with respect to our baseline runs. Our manual runs were prepared by creating a Boolean query based on the topic description and manually adding terms that are consulted from geographical resources available on the web. Although the average performance of the manual run is comparable to the automatic runs, a query by query analysis shows significant differences among individual queries. In general, we got significant improvements (more that 10% average precision) in 8 of the 25 queries. However, we also noticed that 5 queries in the manual runs perform significantly below the automatic runs.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Experimentation,

Keywords

Geographical Information Retrieval, Query Expansion

1 Introduction

For Our participation in GeoCLEF 2006 we used pure information retrieval techniques to expand geographical terms present in the topics. We used a version of the SMART[3] system that has been updated to handle modern weighting schemes (BM25, pivoted length normalization, etc.) as well as multilingual support (ISO-Latin1encoding and stemming for 13 European languages using Porter's stemmer).

We decided to work only with English documents and resources since they were readily available. Section 2 and 3 present the details on collection preparation and query processing. Section 4 presents the retrieval model implemented with the SMART system. Section 5 shows results on the GeoCLEF 2005 data that was used for tuning parameters. Section 6 presents results using the official GeoCLEF 2006 topics as well as a brief analysis and discussion of the results. Section 7 presents our conclusions and future work.

2 Collection Preparation

Details about the GeoCLEF document collection are note discussed in this paper but the reader is referred to the GeoCLEF overview paper. Our document collection consists of 169,477 documents from LA Timas and The Glasgow Herald. Processing of English documents followed a standard IR approach discarding stop words and using Porter's stemmer. Additionally we added word bigrams that identify pairs of contiguous non-stop words to form a two word phrase. These bigrams allowed a stop word to be part of the bigram if they included the word "of" since it was identified as common component of geographical names (i.e. "United_Kingdom" and "City_of_Liverpool" would be a valid bigrams). Documents were indexed using the vector space model (as implemented in the SMART system) with two ctypes. The first ctype was used to index words in the title and body of the article while the second ctype represented the indexing of the word bigrams previously described.

3 Query processing

To process the topics we followed the same approach described above (using stop words, stemming, and adding word bigrams). Each query was represented using two ctypes. The first ctype for single word terms extracted from the parts that will be used in the query (i.e. title and description). For our official runs we only use the title and description.

We designed a way to identify geographical features and expand them using geographical resources but due to the short time available for developing we could not include it in our official runs. For this reason we submitted results using a pure IR approach for this year and work on the development of the geographical feature extraction for next year. Our results should be considered as baseline results. One of the authors created a manual version of the queries using geographical resources available on the internet and writing a Boolean query. This manual run was included in the official results. We also explore automatic retrieval feedback of both automatic and manual queries.

4 Retrieval Model

We use a generalized vector space model that combines the representation of the two ctypes and weights the contribution of each part in the final similarity score between document $\vec{d}_i$ and query $\vec{q}$. The final score is computed as the linear combination of ctype1 (words) and ctype2 (bigrams) as follows:

$$sim(\vec{d}_i,\vec{q}) = \lambda * sim_{words}(\vec{d}_i,\vec{q}) + \mu * sim_{bigrams}(\vec{d}_i,\vec{q})$$

Where λ and μ are coefficients that control the contribution of each of the two ctypes. The values of these coefficients are computed empirically using the optimal results in the GeoCLEF 2005 topics. The similarity values are computed using pivoted length normalization weighting scheme[4] (pivot=347.259, slope= 0.2).

We also performed automatic retrieval feedback by retrieving 1000 documents using the original query and assuming that the top n documents are relevant and the bottom 100 documents are not relevant. This allows us to select the top m terms ranked according to Rocchio's relevance feedback formula[2]:

$$w_{new}(t) = \alpha * w_{orig}(t) + \beta * \frac{\sum_{i \in \mathrm{Re}l} w(t,d_i)}{|\mathrm{Re}l|} - \gamma * \frac{\sum_{i \in \neg \mathrm{Re}l} w(t,d_i)}{|\neg \mathrm{Re}l|}$$

Where α, β, and γ are coefficients that control the contribution of the original query, the relevant documents (*Rel*) and the non-relevant documents (¬*Rel*) respectively. The optimal values for these parameters are also determined using the CLEF 2005 topics. Note that the automatic query expansion adds m terms to each of the two ctypes.

5 Preliminary Experiments Using CLEF2005 Topics

We first tested our baseline system using the GeoCLEF2005 topics. We used the title, description and geographic tags. Table 1 shows the performance values for the baseline run and for the best run submitted to GeoCLEF 2005[1] (BKGeoE1). The mean average precision for this baseline run is 0.3592 which is pretty good and would have been among the top 3 systems in GeoCLEF 2005. This certainly indicates that a pure IR system was enough to answer most of the topics proposed last year.

Table 1 performance of our baseline system against best run in GeoCLEF 2005

	UB Baseline	**UB retrieval feedback**	**Best Run (BKGeoE1)**

Parameters	λ= 10, μ= 1	n= 5, m= 50
		α= 16, β= 96, γ= 8
		λ= 10, μ= 1	

MAP	36.42%	37.93%	39.36%
P@5	59.20%	58.40%	57.60%
P@10	46.80%	50.40%	48.00%
P@20	35.60%	37.20%	41.00%
P@100	18.16%	18.80%	21.48%

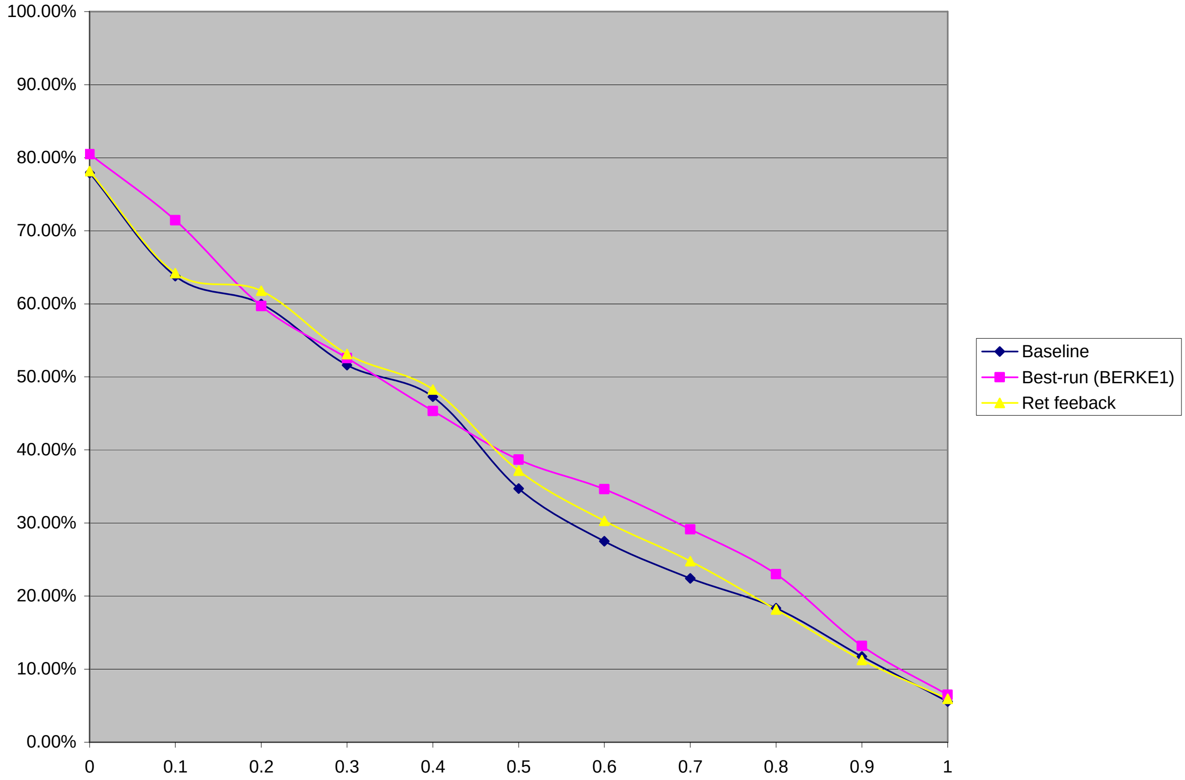

Figure 1 Recall -Precision graph of our baseline and Ret feedback systems against the best run in CLEF 2005

A query by query analysis reveals that the IR approach performs well in many topics but there are a few that could be improved (See Table 2). The system did not perform well in topics 2, 7, 8, 11 and 23. After analyzing these topics we conclude that most of them could have performed better if we had use some sort of expansion of continents using the countries located in them (i.e. European countries).

Table 2 Query by query evaluation of baseline run

Qid	#Relev	#relret	Avg-P	exact-P	P@5	P@10	P@20
1	14	14	60%	57%	80%	60%	40%
2	11	8	10%	18%	20%	20%	15%
3	10	8	44%	40%	80%	40%	20%
4	43	39	35%	37%	80%	70%	55%
5	27	25	52%	56%	100%	80%	55%
6	13	11	28%	38%	40%	30%	30%
7	85	57	6%	9%	0%	10%	5%
8	10	10	4%	0%	0%	0%	0%
9	19	17	46%	42%	100%	80%	40%
10	12	12	82%	75%	100%	90%	50%
11	21	13	7%	5%	20%	10%	5%
12	76	57	14%	17%	60%	50%	35%
13	7	7	52%	43%	60%	40%	25%
14	43	41	39%	49%	40%	70%	50%
15	110	110	74%	72%	60%	70%	75%
16	15	15	88%	80%	100%	100%	65%
17	129	129	50%	49%	80%	90%	60%
18	48	41	29%	38%	80%	50%	50%
19	100	79	16%	24%	0%	30%	30%
20	9	8	11%	22%	40%	20%	10%
21	29	27	44%	38%	100%	80%	55%
22	46	42	48%	52%	80%	80%	75%
23	43	19	3%	5%	20%	10%	5%
24	105	104	50%	56%	60%	50%	65%
25	3	3	59%	67%	60%	30%	15%
All	1028	896	38%	40%	58%	50%	37%

6 Results Using GeoCLEF 2006 Topics

We submitted four official runs: two using automatic query processing and two using manual methods. As expected our results (both automatic and manual) performed above the median system. Results are presented in Table 3.

The automatic runs perform slightly above the median system which indicates that the set of topics for this year where harder to solve using only IR techniques. After taking a look to the official topics we realize that we could have used a better expansion method using the geographical resources (i.e identifying queries that have specific latitude and longitude references to restrict the set of retrieved results).

On the other hand, the manual queries perform in average similarly to the automatic runs but a query by query analysis reveals that there are quite a few queries that outperform significantly the automatic runs. However, at the same time there are two queries that perform significantly below the automatic systems. Note that the first manual run (UBGManual1) does not use automatic feedback while the second manual run (UBGManual2) uses automatic retrieval feedback. This merits further analysis to identify those strategies that are successful in improving performance.

Table 3 Performance of GeoCLEF 2006 Topics

Official Runs		
Run Label	**Mean Avg. P**	**Parameters**

UBGTDrf1
(automatic feedback)
	0.2344	n= 10, m= 20
		α= 16, β= 96, γ= 8,
		λ= 10, μ= 1

UBGTDrf2
(automatic feedback)
	0.2330	n= 5, m= 50
		α= 16, β= 96, γ= 8,
		λ= 10, μ= 1

UBGManual1 (Manual run only)	0.2307	λ= 10, μ= 1

UBGManual2
(automatic feedback)
	0.2446	n= 10, m= 20
		α= 16, β= 96, γ= 8,
		λ= 10, μ= 1

Unofficial Runs		

UBGTDNrf1	0.2758	n= 5, m= 5
		α= 16, β= 96, γ= 8,
		λ= 10, μ= 1

We also noted that our best run (not submitted) performs quite well with respect to our baseline official runs. This run uses title, description and narrative, and conservative retrieval feedback parameters (n=5 documents and m=5 terms). It is also encouraging that this run, when compared to the manual run, captures several of the good terms that were added manually.

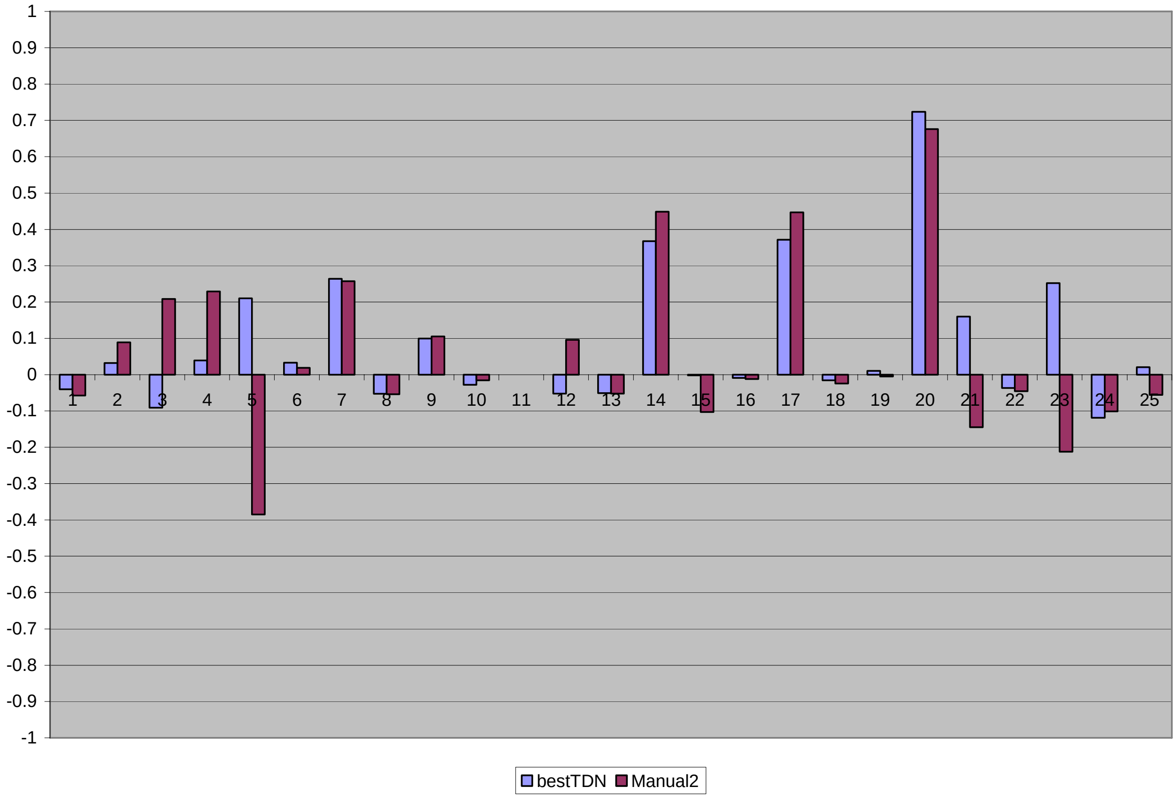

Figure 2 Comparison of best manual run and best automatic run using our system

7 Conclusion

This paper presents an IR based approach to Geographical Information retrieval. Although this is our baseline system we can see that the results are competitive, especially if we use the long topics (title description and narrative). We still need to do more in depth analysis of the reasons why some manual queries improved significantly with respect to the median system and the problem presented in 5 queries that did perform significantly below the median. We plan to explore way to generate automatic geographic references and ontology based expansion for next year.

References

1. Gey, F., Larson, R., Sanderson, M., Joho, H. and Clough, P. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track *Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria*, 2005.
2. Rocchio, J.J. Relevance feedback in information retrieval. in Salton, G. ed. *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice Hall, Englewood Cliff, NJ, 1971, 313-323.
3. Salton, G. (ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice Hall, Englewood Cliff, NJ, 1971.
4. Singhal, A., Buckley, C. and Mitra, M., Pivoted Document Length Normalization. in *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1996), ACM Press, pages, 21-29.