# University of Groningen at GeoCLEF 2007

Geoffrey Andogah and Gosse Bouma

Computational Linguistics Group,
Centre for Language and Cognition Groningen (CLCG),
University of Groningen, Groningen, The Netherlands
{g.andogah,g.bouma}@rug.nl

**Abstract.** This paper describes the approach of the University of Groningen to GeoCLEF task for CLEF 2007. We used geographic scope based approach to rank documents.

## 1 Introduction

This paper describes non-geographic similarity, geographic similarity and combined similarity measures employed to approach GeoCLEF task for CLEF 2007. The motivation for our participation was to test geographic scope (geo-scope) based relevance ranking for geographic information retrieval (GIR). We participated in monolingual English task and our evaluation result shows no significant improvement for geo-scope based approach.

## 2 Approach

### 2.1 Resources

**Geographic Knowledge Base.** We used the World Gazetteer[1], GEOnet Names Server[2] (GNS), Wikipedia[3] and WordNet[4] as the bases for our Geographic Knowledge Base (GKB) for several reasons: free availability, multilingual, coverage of most popular and major places, etc.

**Geographic Tagger.** Alias-I LingPipe[5] was used to detect named entities (location, person and organisation), geographic concepts (continent, region, country, city, town, village, etc.), spatial relations (near, in, south of, north west, etc.) and locative adjectives (e.g. Ugandan).

[1] http://www.world-gazetteer.com
[2] http://earthinfo.nga.mil/gns/html
[3] http://www.wikipedia.org
[4] http://wordnet.princeton.edu
[5] http://alias-i.com/lingpipe

**Lucene Search Engine.** Apache Lucene[6] is a high-performance, full-featured text search engine library written entirely in Java. Lucene's default similarity measure is derived from the vector space model (VSM). Lucene was used to index and search both indexes.

### 2.2 Document Indexing

The reference document collection provided for experimentation was indexed using Lucene. Document HEADLINE and TEXT contents were combined to create document content for indexing (see Table 1 for details). Before indexing, the documents were processed with the Porter stemmer and the default Lucene English stopword list.

**Table 1.** Reference document index structure

| Field | Lucene Type | Description |
|---|---|---|
| docid | Field.Keyword | Document unique identification |
| content | Field.Unstored | Combination of HEADLINE and TEXT tag content |

### 2.3 Similarity Measure

**Non-Geographic Similarity Measure.** We used the Apache Lucene IR library to perform non-geographic search. Lucene's default similarity measure is derived from the vector space model (VSM). The Lucene similarity score formula combines several factors to determine the document score for a query [4]:

$$NonSim(q,d) = \sum_{t\,in\,q} tf(t\,in\,d)\,.\,idf(t)\,.\,bst\,.\,lN(t.field\,in\,d) \tag{1}$$

where, $tf(t\,in\,d)$ is the term frequency factor for term $t$ in document $d$, $idf(t)$ is the inverse document frequency of term $t$, $bst$ is the field boost set during indexing and $lN(t.field\,in\,d)$ is the normalization value of a field given the number of terms in the field.

**Geographic Similarity Measure.** As in [2, 5], we chose to use geographic scopes assigned to queries and documents to perform geographic relevance ranking of documents. Our geographic scope resolver [1] assign multiple scopes to documents ranking the scopes from the most relevant to the least relevant, thereby associating a document with multiple scopes or associating a scope with several documents. We limit geographic scopes (geo-scopes) of population centers

[6] http://jakarta.apache.org/lucene

to continent, continent-directional (e.g. western Europe, eastern Africa, eastern Europe, etc.), country, country-directional (e.g. north-of Italy), province[7], and province-directional (e.g. northern California) level.

Equation 2 depicts our geographic similarity measure formula between query $q$ and document $d$:

$$GeoSim(q,d) = \begin{cases} SF \times WTS & \text{if } SF > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where;

$$SF = \begin{cases} \sqrt{\dfrac{N_{(d,q)}}{N_d + N_q + |N_d - N_q|}} & \text{if } N_{(d,q)} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$WTS = \sum \sqrt{wt_{(q,s)}} \times log(1 + wt_{(d,s)}) \tag{4}$$

and, where; $N_q$ is the number of scopes in the query scope set, $N_d$ is the number of scopes in the document scope set, $N_{(d,q)}$ is the number of document scopes present in query scope set, $wt_{(q,s)}$ is the weight assigned to scope $s$ in query $q$ by the scope resolver and $wt_{(d,s)}$ is the weight assigned to scope $s$ in document $d$ by the scope resolver. For a given query, $N_q$ is invariable whilst $N_d$ and $N_{(d,q)}$ vary per document retrieved. The motivation for designing Equation 3 is to mitigate effects of arbitrarily large variations of $N_d$ and $N_q$ to a reasonable level. The values of Equation 3 are within $0.0 < SF \leq 0.5$. Equation 4 as arranged provides the best overall performance for Equation 2 (that is, applying square root weighting to $wt_{(q,s)}$ and logarithmic weighting to $wt_{(d,s)}$).

**Geo-IR Similarity Measure.** The final similarity score formula is directly derived from Equation 1 and Equation 2 similarity score formulae.

$$Sim(q,d) = \lambda_T \ NonSim(q,d) + \lambda_G \ GeoSim(q,d) \tag{5}$$

$$\lambda_T \ + \ \lambda_G \ = \ 1 \tag{6}$$

where; $\lambda_T$ is the non-geographic interpolation factor and $\lambda_G$ is the geographic interpolation factor. Before the ranked list for non-geographic and geographic relevance ranking are linearly combined, their respective scores are normalized to $[0, 1]$.

### 2.4 Query Formulation for Official GeoCLEF Runs

This section describes the University of Groningen official runs for GeoCLEF 2007. In particular we describe which topic components are used for query formulation and which similarity measures were used to perform relevance ranking.

[7] Here a province represents first order administrative division of a country.

**Topic Categorization.** In [3], geographic topics are categorized into eight according to the way they depend on a place (e.g. UK, St. Andrew, etc.), geographic subject (e.g. city, river, etc.) or geographic relation (e.g. north of, western, etc.). GeoCLEF 2007 topics generation followed similar classification, and in our experiment we grouped the topics into two: (1) topics whose geographic scopes can easily be resolved to a place (GROUP1), and (2) topics whose geographic scopes cannot be resolved to a place (GROUP2).

We performed geographic expansion on the following members of GROUP1 – 51, 59, 60, 61, 63, 65, 66, 70. The motivation for geographic expansion on these topics is that they lack sufficient geographic information or the geographic information provided are too ambiguous. For example, topic 59 is expanded by adding the names of major cities in Bolivia, Columbia, Ecuador and Peru. The Lucene boost factor of 0.45F is assigned to placenames for geographic query expansion while the boost factor for placenames in the original query is left at the default value of 1.0F.

Members of GROUP2 are topics – 56, 67, 68, 72. These topics fall under *geographic subject with non-geographic restriction* with exception of topic 72 which is more complex. Resolving geographic scope of these topics to a specific place is a non trivial undertaking. The most reasonable scope for these topics is geographic subject scope such as lake, river, beach, city, etc. For example, topic 56 concern documents with scope lake.

**CLCGGeoEET00, CLCGGeoEETD00 and CLCGGeoEETDN00.** Queries for these runs are formulated by the content of topic TITLE (T), TITLE-DESC (TD) and TITLE-DESC-NARR (TDN) tags respectively. GROUP1 topics are ranked according to Equation 5 with $\lambda_T = 1.0$ and $\lambda_G = 0.0$. GROUP2 topics are ranked according to Equation 1. However, the query for CLCGGeoEETDN00 was mistakenly formulated by the content of topic TITLE instead of TDN.

CLCGGeoEETDN00P is CLCGGeoEETDN00 with query formulated by topic TDN tag content.

**CLCGGeoEETDN01.** The query for this run should have been formulated by the content of topic TDN tags, however, the official result submitted erroneously used TITLE tag content. GROUP1 topics are ranked according to Equation 5 with $\lambda_T = 0.85$ and $\lambda_G = 0.15$. GROUP2 topics are ranked according to Equation 1.

CLCGGeoEETDN01P is CLCGGeoEETDN01 with query formulated by topic TDN tag content.

**CLCGGeoEETDN01B.** The query for this run should have been formulated by the content of topic TDN tags, however, the official result submitted erroneously used TITLE tag content. GROUP1 topics are ranked according to Equation 5 with $\lambda_T = 0.85$ and $\lambda_G = 0.15$.

For GROUP2 topics, we scan each document retrieved and ranked according to Equation 1 for geographic types (geo-types) as well as determine the geo-types of geographic names found in the documents. Each geo-type found in the document is assigned a weight. For documents containing query geo-type, we add the geo-type weight to Lucene score and then re-rank documents based on the new score.

CLCGGeoEETDN01BP is CLCGGeoEETDN01B with query formulated by topic TDN tag content.

## 3 Evaluation and Future Work

Table 2 shows the result of our runs. The best performing official run is CLCGGeoEETD00. However, as mentioned in the previous section the queries for runs CLCGGeoEETDN00, CLCGGeoEETDN01 and CLCGGeoEETDN01B were erroneously formulated by using the TITLE tag content instead of TITLE-DESC-NARR tags. As such these runs perform poorly. We made correction and they provided the best performance as shown by rows CLCGGeoEETDN00P, CLCGGeoEETDN01P and CLCGGeoEETDN01BP respectively.

The results of runs CLCGGeoEETDN00P and CLCGGeoEETDN01P are statistically equivalent, but 5.2 % better than result for CLCGGeoEETD00 which used TD content. From this we conclude that geographic information may be useful in improving the performance of an IR system in answering geography constrained user information need. However, scope based relevance ranking as implemented shows no significant improvement.

**Table 2.** Individual Run Performance as measured by Mean Average Precision and R-Precision

| Run | MAP | R-Precision |
|---|---|---|
| CLCGGeoEET00 | 0.2023 | 0.2186 |
| CLCGGeoEETD00 | **0.2515** | **0.2595** |
| CLCGGeoEETDN00 | 0.2023 | 0.2186 |
| CLCGGeoEETDN01 | 0.2053 | 0.2234 |
| CLCGGeoEETDN01B | 0.1847 | 0.2019 |
| CLCGGeoEETDN00P | 0.2647 | 0.2743 |
| CLCGGeoEETDN01P | *0.2681* | *0.2878* |
| CLCGGeoEETDN01BP | 0.2442 | 0.2579 |

Our future work will focus on investigating: (1) better geographic similarly measure formulae, (2) the use of geographic scopes selected by the searcher from the returned documents for relevance feedback and (3) term-relevance feedback based on geographic terms (e.g.placenamess) extracted by the searcher afterexaminingg retrieved documents. We are already testing some of these ideas and results are promising.

## 4 Concluding Remarks

We described non-geographic similarity, geographic similarity and combined similarity measures employed to approach GeoCLEF task for CLEF 2007. We tested geographic scope (geo-scope) based relevance ranking for geographic information retrieval (GIR). Our evaluation result shows no significant improvement for geo-scope based approach in monolingual English task.

## 5 Acknowledgements

This work is supported by NUFFIC within the framework of Netherlands Programme for the Institutional Strengthening of Post-secondary Training Education and Capacity (NPT) under project titled ”Building a sustainable ICT training capacity in the public universities in Uganda”.

## References

1. G. Andogah, G. Bouma, J. Nerbonne, and E. Koster. Resolving geographical scope of documents with Lucene. To appear in the near future, 2007.
2. L. Andrade and M. J. Silva. Relevance ranking for geographic IR. In *Workshop on Geographical Information Retrieval, SIGIR'06*, August 2006.
3. F. Gey, R. Larson, M. Sanderson, K. Bischoff, T. Mandl, and C. Womser-Hacker. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In *Working Notes for CLEF 2006 Workshop (CLEF 2006)*, Alcante, Spain, September 2006.
4. O. Gospodnetic and E. Hatcher. *Lucene in Action*. Manning Publications Co., 206 Bruce Park Avenue, Greenwich, CT 06830, 2005.
5. B. Martins, M. J. Silva, and L. Andrade. Indexing and ranking in Geo-IR systems. In *Workshop on Geographical Information Retrieval, SIGIR'05*, November 2005.