BBK-UFRGS@CLEF2009: Query Expansion of Geographic Place Names

Richard Flemmings¹, Joana Barros¹, André P. Geraldo², Viviane P. Moreira² ¹ Department of Geography, Environment and Development Studies Birkbeck, University of London, Malet Street, London WC1E 7HX, UK j.barros@bbk.ac.uk, richflemmings@gmail.com ² Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS) Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil [apgeraldo, viviane]@inf.ufrgs.br

Abstract

For our first participation on CLEF, our aim was to compare plain information retrieval strategies and query expansion and emphasis of geographic terms. ANNIE was used to recognise geographic entities which were expanded using Google's Hierarchical List of Geographical Place Names. The idea was that the expansion would produce more accurate answers. The results have shown the opposite. Our best performing run was the baseline. Future work will include further experiments and a deeper analysis of our results in order to enable the design of a better performing strategy.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Linguistic processing. H.3.4 [Systems and Software]: Performance evaluation

Free Keywords

Experimentation, performance measurement, placenames

1 Introduction

This paper reports on monolingual question answering experiments performed for GikiCLEF. Our aim was to compare the performance of straightforward Information Retrieval techniques with query expansion using geographic terms. 50 Topics were issued in multiple languages. The GikiCLEF task required the Topics to be used to query Wikipedia articles. The Topics issued for the GikiCLEF task each contain a geographically significant reference. By identifying and emphasising this geographic reference within a topic, we anticipated that the queries could provide more focused and accurate answers.

2 **Experiments**

2.1 Description of Runs and Resources

The text collection used was the Portuguese version of Wikipedia. Some details of this collection are given in Table 1.

Number of documents	1,630,303
Total number of terms	193,623,264
Number of unique terms	1,120,786
Average document length	1,173

Table 1 - Details of the Portuguese data collection

We worked on the html files of the Wikipedia articles. The files were pre-processed to extract the textual contents and to remove redirects. We also removed stop-words according to the lists available from Snowball¹. The IR system we used was Zettair [4], which is a compact and fast search engine developed by RMIT University (Australia) distributed under a BSD-style license. Zettair implements a series of IR metrics for comparing queries and documents. We used Okapi BM25 as some preliminary tests we performed on other data collections showed it achieved the best results.

The top 15 retrieved documents for each query were considered as correct answers. We did not perform any filtering of classification of the answers. Answer selection was based solely on the similarity score between the article and the query.

Three different runs were performed. Below we describe each one:

- Run 1 The unchanged topic sentences were used for searching. This was our baseline.
- Run 2 The original topic sentences were submitted to ANNIE [2], an information extraction system of GATE (General Architecture for Textual Engineering) version 4.0. The GATE software allows a body of text to be input and searched for significant words. The ANNIE (A Nearly New Information Extraction System) extension to GATE has a built in gazetteer of place names. This allowed place names within the GikiCLEF topic list to be highlighted ("annotated"). All terms identified as geographic entities by ANNIE had their weights increased when performing Run 2. In order to emphasise geographic terms, we used a modified version of BM25 [3] which was originally designed to promote rare terms.
- Run 3 In order to place a geographical word or phrase into context, and ensure a more diverse range of geographical words could be searched for, the geographic terms identified by Annie, were submitted to Google's Hierarchical List of Geographical Place Names [1]. By replacing individual geographical words within a GikiCLEF Topic with multiple words, it was anticipated that a greater number of documents would be returned. A tool was built to automatically search for multiple words associated with the geographic word(s) from each Topic. Each geographic word was submitted to the tool and all the geographic words from the hierarchical level below were returned. The tool sent the geographic terms in English as queries and received as the result all geographic place names one level below in Portuguese.

For example, Topic 47 contains the word Germany. This topic was submitted using the word Germany, but also using the geographical level below Germany; the German Bundesland (Figure 1). Using an OR operator each original topic has been expanded using the original geographic term, along with terms from the geographical level below it.



Figure 1 – Example Hierarchical search result

2.2 Results

Our results are presented in table 2. The scores are very low. The best result was achieved by Run 1 (the baseline), which was able to return eight out of 85 correct answers. Run 1 was able to retrieve correct answers for three topics (12, 23, and 34). Run 2 retrieved three correct answers for the same topic. Run three was able to

¹ http://snowball.tartarus.org/

retrieve only one correct answer, but for a different topic (47) "Which cities in Germany have more than one university?". The expansion of the original query with the term "Hamburgo" was able to retrieve one correct article.

Run ID	Score	#Correct Answers
Run 1	0.088	8
Run 2	0.012	3
Run 3	0.000	1

Compared to other participants, our best run was ranked in 11th place out of 17 runs.

3 Conclusions

This paper reported on monolingual question answering experiments performed for GikiCLEF. The aim was to compare traditional Information Retrieval, term weighting, and query expansion.

The results have shown that our strategies for improving the results did not produce the expected outcome. It was anticipated that Run 2 and Run 3 would return a greater number of correct answers as the geographical words within each topic were expanded upon. However, our experiments did not take into account the fact that many words within the Topics that can be classed as geographic may not be contained within the Google Hierarchical List of Geographical Place Names. For example, Topic 08 refers to the "Bohemian Forest". The simplistic approach of word substitution used in Run 3 requires further development to take into account the vague geographies that exist within some of the Topics. Further work will include a deeper analysis of the experimental results and the study and development of new techniques for Geographical Question Answering.

Acknowledgements

This work was partially supported by CNPq (Brazil).

References

- 1. *Google's Hierarchical List of Geographical Place Names.* 10-May-2009]; Available from: http://code.google.com/intl/en/apis/maps/documentation/geocoding/index.html.
- 2. Cunningham, H., et al., *Developing Language Processing Components with GATE Version 4 (a User Guide)*. 2007: The University of Sheffield <u>http://gate.ac.uk/sale/tao/</u>.
- 3. Geraldo, A.P. and V.M. Orengo, *UFRGS@CLEF2008: Using Association rules for Cross-Language Information Retrieval*, in *Working Notes of CLEF2008*, F. Borri, A. Nardi, and C. Peters, Editors. 2008: Aarhus, Denmark.
- 4. Zettair. Zettair. [cited 2007 11/06/07]; Available from: www.seg.rmit.edu.au/zettair/.