MetaCarta at GeoCLEF 2005

András Kornai
MetaCarta Inc.
kornai@metacarta.com

In Memoriam Erik Rauch

Abstract

In this paper we divide the processing steps required for the GeoCLEF task into two parts: those that are likely common to all participants and those that are specific to the MetaCarta system. After analyzing the 2005 task we conclude that it has surprisingly little geographic content.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2 [**Database management**]: H.2.4 Textual databases; H.2.8 Spatial databases and GIS; I.5 [**Pattern Recognition**]: I.5.4 Text processing

General Terms

Measurement, Performance, Experimentation

Keywords

Information Extraction, Information Retrieval, Question Answering

0 Introduction

MetaCarta participated in the GeoCLEF 2005 conference on a limited basis: only the English data (Glasgow Herald and LA Times) was used; the German material was only considered in passing. We came away from the evaluation with the impression that this was a keyword search task with little geography-specific ability required from the participating systems.

In Section 1 we describe the processing steps we used, with special emphasis on whether we consider any given step manual or automatic. In Section 2 we describe the MetaCarta results, and in Section 3 we consider the larger issue of whether the query texts require true geographical capabilities or are answerable by generic systems as well.

1 System description

The MetaCarta system was not configured or prepared for this evaluation in any way that differs from the standard setup: no changes were made to the underlying codebase, and no parameters were tuned. The GeoCLEF topics as they stand are not suitable for input to the MetaCarta system; we performed several mechanical conversion steps from one to the other.

First, we ran the 25 topics through the MetaCarta tagger. On the 124 geographic entities we had a precision of 100% (we had no false positives) and a recall of 96.8%: we missed *Scottish*

Islands (twice), *Douglas*, and *Campeltown.* This suggests two evaluation paths: on the **discard** path missed entities are treated as plain (nongeographic) text, and on the **pretend** path we pretend the system actually found these.

Second, we removed meta-guidance such as *find information about* or *relevant documents will describe* since the relevant documents will not have the word "relevant" or "document" in them. This step is performed by the `defluff.sed` script (included with our submission) which, arguably, is closer to manual fluff removal than automated conversion. However, MetaCarta does not encounter the fluffy question style in any context outside GeoCLEF, and it makes no sense for MetaCarta to develop a fully automated module for the defluffing task.

Third, we removed stopwords (defined as everything that has more than 1% of the frequency of the word "the" in a terabyte corpus we used for frequency analysis). While the script `defunc.sed` may look ad hoc, it was itself generated by a script based on the "1% of *the*" criterion and as such we consider this step fully automated.

Fourth, we removed geographic metawords in a manner similar to defluffing: when the task description asks for countries involved in the fur trade the word "country" will not be in the docs. The `degeo.sed` script is also included with our submission.

We believe that these steps (though not necessarily in this order) are *generic* in the sense that every geographic IR system must perform them one way or another. After the generic steps, the topics (only `title` and `desc` fields kept) look as follows (autodetected geographic entities in **boldface**):

GC001 Shark Attacks **Australia California** shark attacks humans

GC002 Vegetable Exporters **Europe** exporters fresh dried frozen vegetables

GC003 AI **Latin America** Amnesty International human rights **Latin America**

GC004 Actions against fur industry **Europe USA** protests violent acts against fur industry

GC005 Japanese Rice Imports reasons consequences first imported rice **Japan**

GC006 Oil Accidents Birds **Europe** damage injury birds caused accidental oil spills pollution

GC007 Trade Unions **Europe** differences role importance trade unions European

GC008 Milk Consumption **Europe** milk consumption European

GC009 Child Labor **Asia** child labor **Asia** proposals eliminate improve working conditions children

GC010 Flooding **Holland Germany** flood disasters **Holland Germany** 1995

GC011 Roman **UK Germany** Roman **UK Germany**

GC012 Cathedrals **Europe** particular cathedrals **Europe United Kingdom Russia**

GC013 Visits American president **Germany** visits President Clinton **Germany**

GC014 Environmentally hazardous Incidents **North Sea** environmental accidents hazards **North Sea**

GC015 Consequences genocide **Rwanda** genocide **Rwanda** impacts

GC016 Oil prospecting ecological problems **Siberia** and **Caspian Sea** Oil petroleum development related ecological problems **Siberia Caspian Sea**

GC017 American Troops **Sarajevo Bosnia Herzegovina** American troop deployment **Bosnia Herzegovina Sarajevo**

GC018 Walking holidays **Scotland** walking holidays **Scotland**

GC019 Golf tournaments **Europe** golf tournaments held European

GC020 Wind power Scottish Islands electrical power generation using wind power islands **Scotland**

GC021 Sea rescue **North Sea** rescues **North Sea**

GC022 Restored buildings Southern **Scotland** restoration historic buildings southern **Scotland**

GC023 Murders violence South-West **Scotland** violent acts murders South West part **Scotland**

GC024 Factors influencing tourist industry **Scottish Highlands** tourism industry Highlands **Scotland** factors affecting

GC025 Environmental concerns around Scottish **Trossachs** environmental issues concerns **Trossachs Scotland**

Table 1: Preprocessed Queries

Note how well the results of stopword removal from the `desc` section approximate the `title` section: aside from the last three topics, (where the `desc` section is really narrative) the two are practically identical. Therefore, our first run is based on titles only (see below).

1.1 MetaCarta-specific steps

The natural mode of operation for the MetaCarta system is to use the map as a filter: select a region of interest, such as the Trossachs, and type in some keywords such as *environmental* or *pollution* and see what documents are displayed as a result.

To emulate this operation, we created bounding boxes for each of the regions in the topics. While we didn't get around to fully automating the process of querying the database for polygons and creating the bounding boxes automatically, there is nothing in this step that requires human intervention and for the purposes of submission we consider this automated. The following table was used:

```
Asia 25.0 179.9 6.0
Australia 112.9 159.1 -9.1 -54.7
Bosnia Herzegovina 15.7 19.6 45.2 42.5
California -124.4 -114.1 42.0 32.5
Caspian Sea 47.0 54.0 47.0 36.0
Europe -11.0 60.0 72.00 32.00
Germany 5.8 15.0 55.0 47.2
Holland 3.3 7.2 53.5 50.7
Japan 122.9 153.9 45.5 24.0
Latin America -118.0 -35.0 32.0 -55.0
North Sea -4.0 8.0 65.0 51.0
Russia 26.0 60.0 72.0 41.1
Rwanda 28.8 30.8 -1.0 -2.8
Scotland -8.0 0.0 61.0 55.0
Scottish Highlands -8.0 -2.0 59.3 56.0
* Scottish Islands -8.0 0.0 61.0 56.0
Siberia 60.0 179.9  82.0 48.0
* Trossachs -4.5 -4.25 56.5 56.0
United Kingdom -8.6 2.0 60.8 49.0
United States -125.0 -66.0 49.0 26.0
```

Table 2: Bounding Boxes

Items marked by * did not have a bounding box in the database and reflect manual assignment, a fact that is reflected in our notion of **discard** versus **pretend** evaluation.

Given a fixed collection of documents, such as the English dataset provided for GeoCLEF, a MetaCarta query has three parameters: `maxdocs` is the maximum number of document IDs we wish to see, typically 10 for "first page" results, `bbleft bbright bbtop bbbottom` are longitudes and latitudes for the bounding box, and an arbitrary number of keywords, implicitly ANDed together.

2 Results

In run 0 we only took the title words, the automatically detected regions, created a query as described in 1.1 with `maxdocs` set at 200 (since the system returns results in rank order, to create a first page one can just apply `head` to the result set). When the query implied logical OR rather than AND, we run the queries separately and sorted the results together by relevance.

Run 0 mimicked a **true geographic** search where the geographic portion of the query is input through the map interface. Run 1 is a **true keyword** search where everything (including geographic words) is treated just as a keyword (so the discard and the pretend paths coincide). Running `trec_eval` produces the summaries:

def	**Run 0**	**Run 1**
num_q	22	15
num_ret	1494	1002
num_rel	895	765
num_rel_ret	289	132
map	0.1700	0.1105
R-prec	0.2155	0.1501
bpref	0.1708	0.1148
recip_rank	0.6748	0.6522
ircl_prn.0.00	0.6837	0.6633
ircl_prn.0.10	0.4178	0.2904
ircl_prn.0.20	0.3443	0.2188
ircl_prn.0.30	0.2977	0.1700
ircl_prn.0.40	0.1928	0.1103
ircl_prn.0.50	0.0971	0.0676
ircl_prn.0.60	0.0435	0.0365
ircl_prn.0.70	0.0261	0.0109
ircl_prn.0.80	0.0130	0.0109
ircl_prn.0.90	0.0000	0.0109
ircl_prn.1.00	0.0000	0.0089
P5	0.4455	0.3467
P10	0.3182	0.2333
P15	0.2667	0.1867
P20	0.2500	0.1867
P30	0.2182	0.1644
P100	0.1141	0.0740
P200	0.0636	0.0410
P500	0.0263	0.0176
P1000	0.0131	0.0088

Table 3: Results on Runs 0 and 1

3 Conclusions

As can be seen from comparing the two runs, the non-geographic and the geographic results are remarkably close, which supports the conclusion we arrived at from an informal, manual assessment that very little, if any, geographic specialization is required on these tasks.

First, the selection of geographic entities is limited, and most of them fit in what MetaCarta calls "Tier 1", a small set (2350 entries) of core place names whose approximate locations are known to everyone with a high school education. With the possible exception of the Scottish Islands (a class better defined by listing than by coherent geography) and the Trossachs (whose boundaries are clearly explained in the narrative task) a system with a small post-hoc gazetteer table could handle most of the questions: the only entries missing from the Tier 1 gazetteer are *Argyll, Ayr, Callander, Loch Achray, Loch Katrine, Loch Lomond, Perthshire, Scottish Islands* and *Trossachs*, and these do not even appear in the non-narrative sections.

Given that the problem of avoiding false positives is increasingly hard as we add more and more entities to the gazetteer, a task that encourages the use of trivial gazetteers will not serve the overall evaluation goals well. As it is, MetaCarta has an F-measure of 98.36% which would be quite impressive, were it produced on a more realistic test set.

Second, even within this limited set, one has the feeling (perhaps unsubstantiated, the guidelines didn't address the issue) that many of the toponyms are used metonymically. In particular, *Europe* seems to refer to the EU as a political entity rather than to the continent (see in particular items 4 and 8).

Because the bar is set so low, it is expected that almost all systems will pass the geographic hurdle in GeoCLEF, and their comparison will amount to a comparison of their non-geographic capabilities. To be sure, the issues that come to the fore are classical, and fascinating issues in Information Retrieval:

1. stopword filtering
2. stemming (morphological analysis)
3. selecting keywords for a concept (vocabulary enrichment)
4. good handling of disjuncts and negation (Booleans)
5. fluff removal

There is little doubt that all of these issues are worthy of formal evaluation, though not necessarily as part of an evaluation focusing on geographic IR.

1. We believe that our stopword filtering may be overly generous (there are 75 words that meet the "more than 1% of *the*" criterion) inasmuch as it includes content words like *people*, *part*, *two*, or *time*, but the task is clearly not geared to test this further.

2. We have purposely refrained from stemming, which decreased our overall scores significantly (e.g. on GC001 we missed LA041994-0146 "WOMEN BITTEN BY SHARK HAD WON LEUKEMIA BATTLE" since it has the word *attacked* rather than *attack*) but did not affect our main point, which is to compare geographically oriented search to true keyword search.

3. Where conjunction provides too few results, a reasonable approach would be to decrease the number of keywords: e.g. topic GC006 finds nothing with keywords `oil accident birds` so reducing `oil accident` to `oil` would be helpful. The queries display a clear preference for semantic IR, asking e.g. for "consequences", "concerns", and other highly abstract concepts generally considered beyond the ken of mainstream IR techniques. Ideally, one's system should understand the underlying concepts, but in reality even the human judges are on slippery ground (for example, we do not believe that documents such as LA083194-0133 are actually responsive to GC015) and of course MetaCarta has no software to make a semantic decision. While this is clearly a weakness in the state of the art, again we feel the issue has little to do with geography.

4. To automate booleans a reasonable blind approach would be to take a random word from each conjunct and try the combinations e.g. given *oil prospecting and ecological problems* form

"oil ecological", "oil problems", "prospecting ecological", and "prospecting problems" – we have not implemented this and handle the cases of conjunction in toponyms *North and South America* by a mechanism that was not exercised by the test. The descriptive queries in particular offer a fascinating glimpse into other problems that are viewed as important research topics such as negation: "Reports regarding canned vegetables, vegetable juices or otherwise processed vegetables are not relevant". We do not deny the importance of these problems, but we doubt the wisdom of burdening this task with these.

5. As we have stated earlier, our fluff removal (including geographic fluff) was post hoc, and we do not believe that our scripts would generalize particularly well for future tests of a similar nature. However, again we would suggest reengineering the task rather than building infrastructure for the way it stands now.

However fascinating these problems might be, tacking "in Rwanda" on a question does not make it truly geographic, in fact there is reason to believe that the easy part of geography (continents and countries) is not any different from any other topic hierarchies. GeoCLEF 2005 did not go beyond the easy parts; let's hope next time will be better.

Acknowledgements

Special thanks to Bradley Thompson and the MetaCarta team, very much including Erik Rauch, whose sudden death fills us all with sorrow.