WHERE IS THE GEO IN GEOCLEF?

András Kornai
MetaCarta Inc.
kornai@metacarta.com

2005 GeoCLEF Workshop
Vienna Sept 23 2005

In Memoriam Erik Rauch

SPECIAL THANKS TO:

- Brad Thompshon
- Karl Wright
- Doug Cutting

OUTLINE

1. How Geographic IR works for MetaCarta
2. *Generic* vs *geo-specific tasks*
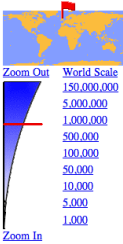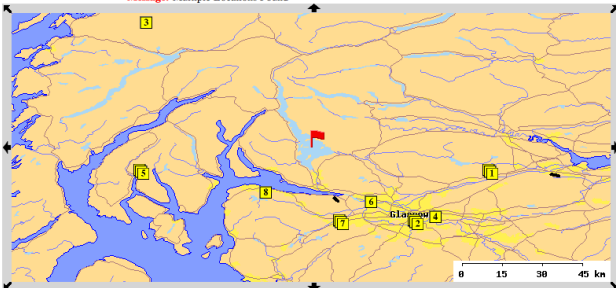3. Results
4. Conclusions

Part I

HOW GEO IR WORKS

SCRIPT-DRIVEN USE

```
./headnode-search.py 200 112.9 159.1 -9.1 -54.7
shark attacks human >>GC001-a
./headnode-search.py 200 -124.4 -114.1 42.0 32.5
shark attacks human >>GC001-c
./headnode-search.py 200 -11.0 60.0 72.00 32.00
Vegetable exporters fresh dried frozen vegetables
>>GC002 ./headnode-search.py 200 -118.0 -35.0
32.0 -55.0 Amnesty International human rights
>>GC003
```

To get from GeoCLEF to MetaCarta we need

AUTO bounding box

To get from GeoCLEF to MetaCarta we need

AUTO bounding box

AUTO geolocation of query

TO GET FROM GEOCLEF TO METACARTA WE NEED

AUTO bounding box
AUTO geolocation of query
SEMI defluffing

To get from GeoCLEF to MetaCarta we need

- AUTO bounding box
- AUTO geolocation of query
- SEMI defluffing
- AUTO stopword removal

TO GET FROM GEOCLEF TO METACARTA WE NEED

AUTO bounding box
AUTO geolocation of query
SEMI defluffing
AUTO stopword removal
SEMI geofluff removal

To get from GeoCLEF to MetaCarta we need

AUTO bounding box

AUTO geolocation of query

SEMI defluffing

AUTO stopword removal

SEMI geofluff removal

NON stemming

To get from GeoCLEF to MetaCarta we need

AUTO bounding box

AUTO geolocation of query

SEMI defluffing

AUTO stopword removal

SEMI geofluff removal

NON stemming

SEMI booleans

To get from GeoCLEF to MetaCarta we need

AUTO bounding box

AUTO geolocation of query

SEMI defluffing

AUTO stopword removal

SEMI geofluff removal

NON stemming

SEMI booleans

NON query enrichment

TO GET FROM GEOCLEF TO METACARTA WE NEED

AUTO bounding box
AUTO geolocation of query
 SEMI defluffing
AUTO stopword removal
 SEMI geofluff removal
 NON   stemming
 SEMI booleans
 NON   query enrichment
 NON   (auto)feedback

Part II

Generic vs geo-specific tasks

generic	geo-specific
defluffing	geofluff removal
stopword removal	doc geolocation
stemming	query geolocation
boolean	bounding boxes
query enrichment	relevance mixing
(auto)feedback	display

Part III

Results

FoM	0	1	(3)	(4)	(0+3)
num_q	22	15	25	25	25
num_ret	1494	1002	820	500	1594
num_rel	895	765	1028	1028	1028
num_rel_ret	289	132	214	144	339
map	0.1700	0.1105	0.1819	0.1653	0.1959
R-prec	0.2155	0.1501	0.2328	0.2040	0.2396
bpref	0.1708	0.1148	0.1796	0.1570	0.1896
recip_rank	0.6748	0.6522	0.5453	0.5970	0.6778
ircl_prn.0.00	0.6837	0.6633	0.6064	0.6344	0.6878
ircl_prn.0.10	0.4178	0.2904	0.5096	0.4757	0.4505
ircl_prn.0.20	0.3443	0.2188	0.3748	0.3338	0.3740
ircl_prn.0.30	0.2977	0.1700	0.1622	0.1765	0.2986
ircl_prn.0.40	0.1928	0.1103	0.1161	0.1453	0.2064

FoM	0	1	(3)	(4)	1+3
ircl_prn.0.50	0.0971	0.0676	0.0976	0.1301	0.1221
ircl_prn.0.60	0.0435	0.0365	0.0687	0.0680	0.0750
ircl_prn.0.70	0.0261	0.0109	0.0687	0.0430	0.0597
ircl_prn.0.80	0.0130	0.0109	0.0663	0.0410	0.0457
ircl_prn.0.90	0.0000	0.0109	0.0513	0.0063	0.0207
ircl_prn.1.00	0.0000	0.0089	0.0394	0.0063	0.0194
P5	0.4455	0.3467	0.4240	0.4160	0.4640
P10	0.3182	0.2333	0.3680	0.3640	0.3560
P15	0.2667	0.1867	0.3627	0.3227	0.3067
P20	0.2500	0.1867	0.3300	0.2880	0.2820
P30	0.2182	0.1644	0.2360	0.1920	0.2427
P100	0.1141	0.0740	0.0856	0.0576	0.1204
P200	0.0636	0.0410	0.0428	0.0288	0.0660
P500	0.0263	0.0176	0.0171	0.0115	0.0271
P1000	0.0131	0.0088	0.0086	0.0058	0.0136

Part IV

Conclusions

PUT THE GEO IN GEOCLEF:

- Make sure geoparsing is exercised
- Make sure gazetteer is exercised
- (Make sure bounding box capability is exercised)
- Make sure stemming/booleans/query enrichment are not limiting factors