Blind Relevance Feedback and Named Entity based Query Expansion for Geographic Retrieval at GeoCLEF 2006

Kerstin Bischoff, Thomas Mandl
and Christa Womser-Hacker

Information Science, University of Hildesheim, Germany

7th Workshop of the Cross-Language Evaluation Forum (CLEF), September 20-22, 2006, Alicante, Spain

Outline

- Challenges for Geographic Information Retrieval (GIR) at GeoCLEF 2006
- Our Approach: Geographic expansion via Blind Relevance Feedback (BRF)
- Post submission runs with (Geo)BRF and Boolean Retrieval
- Current Work: Heuristics for expansion

Challenges

GC 048-title: Forest fires in Northern Portugal

- Geo-Parsing: Named Entity Recognition and Classification (NER/NEC)
- Geo-Coding for spatial retrieval and ranking

  or expansion for text retrieval

Some problems:

- NER/NEC, disambiguation and translation
- Resources: availability and coverage of gazetteers
- Heuristics for geographical expansion

Our approach

- Resource problem for topics like *the Middle East*, *the Tropics* or *Northern Portugal*

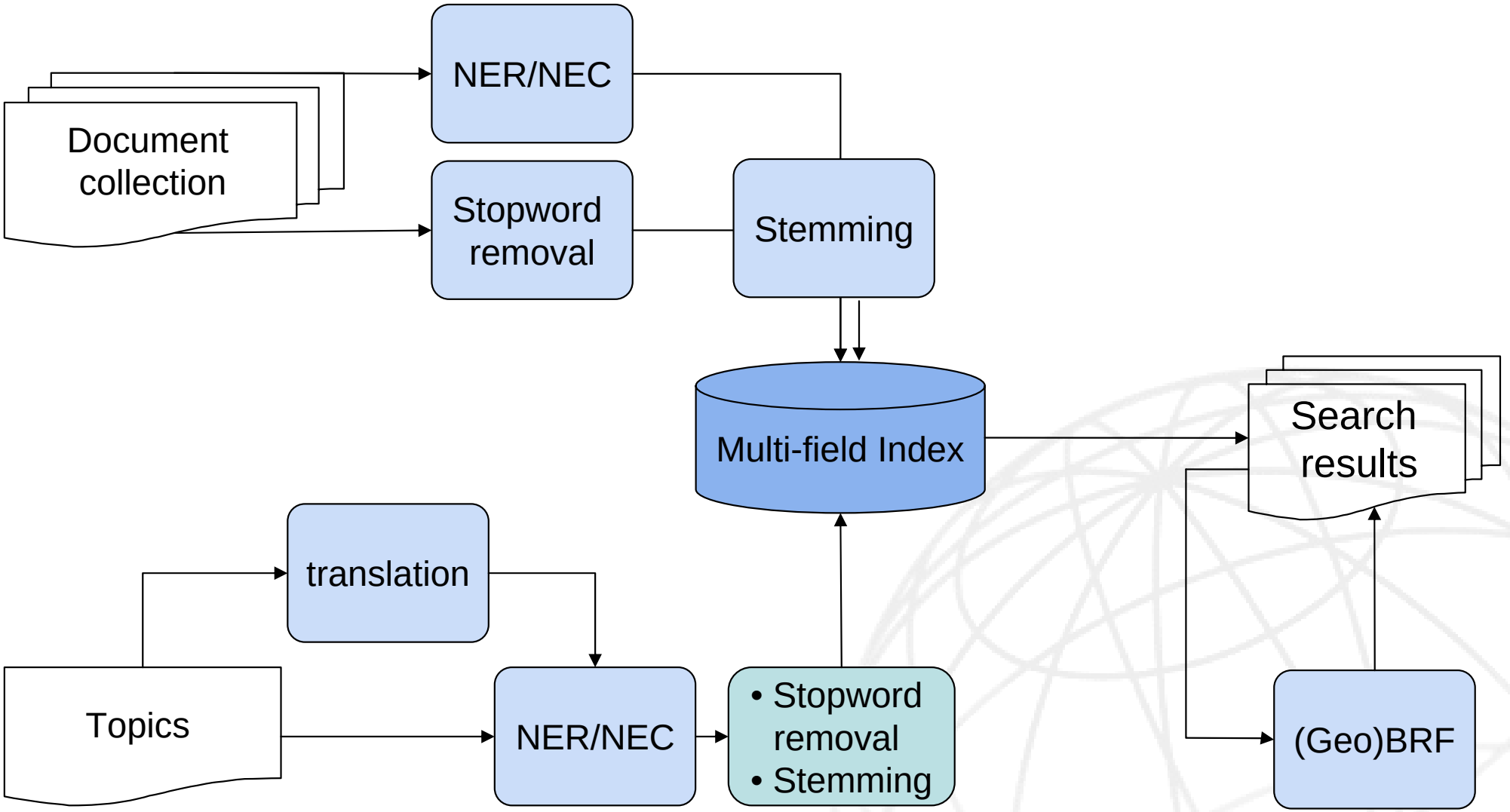
➔ Geographic query expansion via BRF

- Index collection with (geographic) Named Entities (NE)
- Force the addition of geographic NEs from the top-ranked documents within the BRF
- Expand with/without Boolean retrieval

Our approach

- Processing steps and tools used
  - (Translation of topics via Babelfish, Linguatec and FreeTranslation)
  - (NER/NEC via the machine learning tool Alias-I-LingPipe)
  - Stopword removal: lists by the University of Neuchatel‘s
  - Stemming: Lucene for German, Snowball for English
  - Ranked Boolean retrieval via Lucene
  - (BRF or GeoBRF with high weights for geographic NEs)

Our approach



Our approach

- Translation and NER/NEC
  - Translation of NEs often insufficient „*New narrow country*“, „*Vila actually*“ or „*grandmas*“
  - NER should precede a gazetteer lookup for name variants
  - (multiple) translation improved NER/NEC for German
  - Listings and compound names problematic for NER/NEC
- GeoBRF – an example
  - GC 37 *the Middle East*: *Syria*, *Israel*, *Cairo*, *Egypt*, *Lebanon*, *Gaza*, *Jerusalem*, *Beirut*, *Gulf*

Post submission runs

Table 1: Effects of NE-weighting and (Geo)BRF for monolingual German

Fields	BRF (docs,terms)	**MAP**	NEs weighted **MAP**	With GeoBRF **MAP**	NEs weighted and GeoBRF **MAP**
TD	–	25.73	27.67		
TDN	–	23.43	25.65		
TD	5, 25	19.72	23.59	26.43	29.54
TDN	5, 25	22.28	23.33	25.71	25.65
TD	10, 20	21.82	24.72		29.07
TDN	10, 20	21.05	24.70		27.90
TD	10, 30	19.25	24.98	30.34	31.20
TDN	10, 30	21.60	24.03	28.65	28.72

23.73% for manually corrected NEs

33.38% for manually corrected NEs

Post submission runs

Table 2: Effects of NE-weighting and (Geo)BRF for monolingual English

Fields	BRF (docs,terms)	**MAP**	NEs weighted **MAP**	With GeoBRF **MAP**	NEs weighted and GeoBRF **MAP**
TD	–	18.11	20.38	–	–
TDN	–	16.27	18.42	–	–
TD	5, 25	15.71	19.03	19.00	18.65
TDN	5, 25	15.47	18.30	18.12	18.20
TD	10, 20	15.74	18.65	15.69	14.70
TDN	10, 20	16.84	17.70	14.22	14.83
TD	10, 30	14.66	17.88	17.44	17.84
TDN	10, 30	16.41	18.04	15.53	17.50

Post submission runs

Table 3: Effects of NE-weighting and (Geo)BRF with Boolean Retrieval for monolingual English

Fields	BRF (docs,terms)	**MAP**	non-geographic NEs weighted **MAP**	With GeoBRF **MAP**	NEs weighted and GeoBRF **MAP**
TD	–	21.23	21.23	–	–
TDN	–	17.60	19.41	–	–
TD	5, 25	18.09	18.09	14.08	14.22
TDN	5, 25	19.49	19.49	14.91	14.96
TD	10, 20	18.99	18.99		18.07
TDN	10, 20		20.05		20.92
TD	10, 30		18.95	13.49	13.53
TDN	10, 30	20.18	20.18	18.56	18.56

17.41% for manually corrected NEs

21.13% for manually corrected NEs

Current Work

- Intellectual analysis and manual query optimization for GeoCLEF topics 2005 and 2006
  - in order to find good strategies and develop heuristics

- Example heuristic for German for country names:

```
if   country is U.S.A or is in Europe
then expand with the biggest cities
if   country equals collection home
then expand with divisions and biggest cities
else no expansion/only capital
```

- Factors like proximity, elite nation, relevance may influence publicity of location names

References

[1] Amitay, Einat; Har'El, Nadav; Sivan, Ron; Soffer, Aya (2004): *Web-a-Where: Geotagging Web content*. In: SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. July 2004, Sheffield, UK. ACM, pp. 273-280.

[2] Chaves, Marcirio Silveira; Martins, Bruno; Silva, Mário J. (2005): *Challenges and Resources for Evaluating Geographical IR*. In: Proceedings of the 2nd International Workshop on Geographic Information Retrieval, CKIM 2005. Nov. 2005, Bremen, Germany. pp. 65-69.

[3] Clough, Paul (2005): *Extracting Metadata for Spatially-Aware Information Retrieval on the Internet*. In: Proceedings of the 2nd International Workshop on Geographic Information Retrieval, CKIM 2005. Nov. 2005, Bremen, Germany. pp 25-30.

[4] Clough, Paul; Joho, Hideo; Purves, Ross (2005): *Identifying imprecise regions for geographic information retrieval using the web*. In: Proceedings of the GIS RESEARCH UK 13th Annual Conference. Glasgow, UK, 2005. pp. 313-318.

[5] Ageno, Alicia; Ferrés, Daniel; Rodríguez, Horacio (2005): *The GeoTALP-IR System at GeoCLEF-2005: Experiments Using a QA-based IR System, Linguistic Analysis, and a Geographical Thesaurus*. In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF. Sep. 2005, Vienna, Austria. http://www.clef-campaign.org/2005/working_notes/.

References

[6] Gey, Fredric; Larson, Ray; Sanderson, Mark; Joho, Hideo; Clough, Paul (2005): *GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track*. In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF. Sep. 2005, Vienna, Austria. http://www.clef-campaign.org/ working_notes/.

[7] Gey, Frederic; Petras, Vivien (2005): *Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents*. In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF. Sep. 2005, Vienna, Austria. http://www.clef-campaign.org/ working_notes/.

[8] Hackl, René; Mandl, Thomas; Womser-Hacker, Christa (2005): *Mono- and Cross-lingual Retrieval Experiments at the University of Hildesheim*. In: Peters, Carol; Clough, Paul; Gonzalo, Julio; Kluck, Michael; Jones, Gareth; Magnini, Bernard (eds): Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Berlin et al.: Springer [LNCS 3491] pp. 165-169.

[9] Larson, Ray R. (2005): *Chesire II at GeoCLEF: Fusion and Query Expansion for GIR*. In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF. Sep. 2005, Vienna, Austria. http://www.clef-campaign.org/2005/working_notes/

[10]Mandl, Thomas; Schneider, René; Strötgen, Robert (2006): *A Fast Forward Approach to Cross-lingual Question Answering for English and German*. In: Gey, Fredric C.; Gonzalo, Julio; Jones, Gareth J.F; Kluck, Michael; Magnini, Bernardo; Müller, Henning; Peters, Carol; de Rijke, Maarten (eds.). Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF. Sep. 2005, Vienna, Austria. Revised Selected Papers. LNCS 4022; Springer, pp. 332-336.