Monolingual Retrieval Experiments with Spatial Restrictions at GeoCLEF 2007

Ralph Kölle, Ben Heuwing, Thomas Mandl, Christa Womser-Hacker

Information Science, University of Hildesheim,
Marienburger Platz 22
D-31141 Hildesheim, Germany

koelle@uni-hildesheim.de

Abstract

The participation of the University of Hildesheim focused on the monolingual German and English tasks of GeoCLEF 2007. Based on the results of GeoCLEF 2005 and GeoCLEF 2006, the weighting and expansion of geographic named entities (NE) and Blind Relevance Feedback were combined. This year an improved model for German Named Entity Recognition was evaluated.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-Language Information Retrieval, Evaluation, Geographic Information Retrieval Systems

1 Introduction

Retrieval of documents which fulfil a spatial requirement is an important task for retrieval systems. Such geographic information retrieval systems are evaluated within the GeoCLEF track at CLEF. Our experiments expanded an ad-hoc system to allow geographic queries. Based on the participation in GeoCLEF 2006 and some post experiments [Bischoff et al. 2007], we again adopted a (blind) relevance feedback approach which focuses on named geographic entities. To improve the named entity recognition (NER) for German entities we used an optimised model based on the NEGRA[1]-corpus for training.

2 Geographic Retrieval System

The system we augmented for this experimentation with (geographic) NEs in GIR is based on a retrieval system applied to ad-hoc retrieval in previous CLEF campaigns [Gey et al. 2007]. Apache Lucene[2] is the backbone system for stemming, indexing and searching.

Named Entity Recognition was carried out with the open source machine learning tool LingPipe[3], which identifies named entities and classifies them into the categories Person, Organization, Location and Miscellaneous according to a trained statistical model.

[1] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html

[2] http://lucene.apache.org/java/

[3] http://www.alias-i.com/lingpipe

3 Submitted Runs

After experimentation with the GeoCLEF data of 2006 we submitted runs differing in parameters and query processing steps.
Run descriptions and results measured as Mean Average Precision (MAP) are shown in Table 1 for submitted monolingual runs and in Table 2 for the corresponding results with the training topics of 2006.

Table 1. Results monolingual runs (nm = new NER-Model - om = old NER-Model)

Run	***Language***	***Narrative***	***BRF (weight-docs-terms)***	**Geo-NE's** ***(weight-docs-terms)***	***MAP***
HiMoDeBase	German		0.5-5-25	-	0,2019
HiMoDeNe2	German		0.2-5-25	0.2-30-40 (nm)	0,1953
HiMoDeNe2Na	German	x	0.2-5-25	0.15-30-60 (nm)	0,2067
HiMoDeNe3	German		0.2-5-25	1.0-10-4 (nm)	0,1795
HiMoEnBase	English		0.5-5-25	-	0.1405
HiMoEnNe	English		0.2-5-25	0.5-5-20	0.1535
HiMoEnNaNe	English	x	0.2-5-25	0.5-5-20	0.1497
HiMoEnNe2	English		0.2-5-25	2-10-3	0,1268

Table 2. Results for training topics of monolingual runs (nm = new NER-Model - om = old NER-Model)

Run	***Language***	***Narrative***	***BRF (weight-docs-terms)***	**Geo-NE's** ***(weight-docs-terms)***	***MAP***
HiMoDeBase	German		0.5-5-25	-	0.1722
HiMoDeNe1	German		0.2-5-25	0.2-30-40 (om)	0.1811
HiMoDeNe2	German		0.2-5-25	0.2-30-40 (nm)	0.1963
HiMoDeNe2Na	German	x	0.2-5-25	0.15-30-60 (nm)	0.2013
HiMoDeNe3	German		0.2-5-25	1.0-10-4 (nm)	0.1811
HiMoEnBase	English		0.5-5-25	-	0.1893
HiMoEnNe	English		0.2-5-25	0.5-5-20	0.1966
HiMoEnNaNe	English	x	0.2-5-25	0.5-5-20	0.1946
HiMoEnNe2	English		0.2-5-25	2-10-3	0.1795

With the training topics of 2006 best results were made expanding the query with 40 geographic terms from the best 30 documents giving each a relative weight of 0.2 compared to the rest of the query (for German) and using 20 terms from top5 documents with a relative weight of 0.5 for English (Table 2). While in the case of the English topics this hold true for the submitted runs, for German topics the base run without NER performed best (Table 1).
The worse results for the English topics indicate more difficult topics (concerning our retrieval system) for 2007. With the German results remaining on almost the same level, the optimised NER-model for German seems to improve retrieval quality.

Summing up, we could not find a substantial positive impact of additional geographic information, but the effect of investment in optimizing the Geo-NE model seems to be positive.

5 Conclusion and Outlook

Optimised Geo-NE models seem to have positive effect on retrieval quality for monolingual tasks. For future experiments, we intend to integrate geographic ontologies to expand entities with neighbouring places, villages and regions. Furthermore we will integrate Wikipedia as translation tool for Geo-NEs to participate in multilinual tasks of GeoCLEF in the future.

References

Bischoff, Kerstin; Mandl, Thomas; Kölle, Ralph; Womser-Hacker, Christa (2007): Geographische Bedingungen im Information Retrieval: Neue Ansätze in Systementwicklung und Evaluierung. In: Oßwald, Achim; Stempfhuber, Maximilian; Wolff, Christian (Hrsg.): Open Innovation – neue Perspektiven im Kontext von Information und Wissen? Proc 10. Internationales Symposium für Informationswissenschaft (ISI 2007) 30. Köln Mai - 1. Juni 2007. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft 46] pp. 15-26.

Gey, Fredric; Larson, Ray; Sanderson, Mark; Bishoff, Kerstin; Mandl, Thomas; Womser-Hacker, Christa; Santos, Diana; Rocha, Paulo; Di Nunzio, Giorgio; Ferro, Nicola (2007): GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, Carol et al. (Eds.). 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer [Lecture Notes in Computer Science 4730] pp. 852-876.