

University of Hagen at GeoCLEF 2007: Exploring Location Indicators for Geographic Information Retrieval

Johannes Leveling and Sven Hartrumpf
Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany
firstname.lastname@fernuni-hagen.de

Abstract

Location indicators are text segments from which a geographic scope can be inferred, e.g. adjectives, demonyms (names for inhabitants of a place), geographic codes, orthographic variants, and abbreviations can be mapped to location names in one or more inferential steps. In this paper, the normalization of location indicators and treating morphology of location indicators for geographic information retrieval (GIR) within the system GIRSA (Geographic Information Retrieval by Semantic Annotation) are explored.

Several retrieval experiments are performed on the German GeoCLEF 2007 data, including a baseline IR experiment on stemmed text (0.119 mean average precision, MAP). Results for this experiment are compared to results for experiments with normalized location indicators. Additionally, the latter approach was combined with an approach using semantic networks for retrieval (an extension of an experiment performed for GeoCLEF 2005).

When using the topic title and description, the best performance was achieved by the combination of approaches (0.196 MAP); adding location names from the narrative part increased MAP to 0.258. Results indicate that 1) employing normalized location indicators improves MAP and increases the number of relevant documents found; 2) additional location names from the narrative increase MAP and recall, and 3) the semantic network approach has a high initial precision and even adds some relevant documents which were previously not found.

For bilingual (English-German) experiments, queries were first translated into German before utilizing the translation as input to GIRSA. Performance for these experiments is generally lower, but reflect results for monolingual German. The baseline experiment (0.114 MAP) is clearly outperformed by all other experiments, achieving the best performance for a setup using title, description, and narrative (0.209 MAP).

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*Indexing methods; Linguistic processing*; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Query formulation; Search process*; H.3.4 [Information Storage and Retrieval]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

Location Indicators, Geographic Information Retrieval

1 Introduction

Traditional information retrieval applies stemming to all words in a text. In the context of geographical information retrieval (GIR) on textual information, named entity recognition and classification play an important role to identify location names and to avoid stemming them. GIR is concerned with facilitating geographically-aware retrieval of information. This awareness often results from identifying proper nouns in the text, disambiguating them further into person names, organization names, and location names (geographic entities). Thus, identification of location names is typically restricted to proper nouns only.

The main goal of this paper is to investigate if one should aim at a broader GIR approach which is not solely based on proper nouns corresponding to location names. To test end, the notion of location indicators is introduced and retrieval experiments are performed by the system GIRSA (Geographic Information Retrieval by Semantic Annotation).¹ The experiments are based on documents and topics for GeoCLEF 2007, the geographic information retrieval task at CLEF 2007 (Cross Language Evaluation Forum).

2 Location Indicators

2.1 Definition

In this paper, location indicators are investigated. Location indicators are text segments from which the geographic scope of a document can be inferred. They include, but are not limited to:

- Adjectives corresponding to a location.
Examples: “*tunesisch*”/“*Tunisian*” for “*Tunesien*”/“*Tunisia*”; “*irisch*”/“*Irish*” for “*Irland*”/“*Ireland*”; “*bayrisch, bayerisch*”/“*Bavarian*” for “*Bayern*”/“*Bavaria*”.
- Demonyms, e.g. the name for inhabitants originating from a location.
Examples: “*Franzose, Französin*”/“*Frenchman, Frenchwoman*” for “*Frankreich*”/“*France*”; “*Mongole, Mongolin*”/“*Mongolian*” for “*Mongolei*”/“*Mongolia*”; “*Düsseldorfer, Düsseldorferin*”/“*inhabitant of Düsseldorf*” for “*Düsseldorf*”.
- Codes for a location name, including ISO region codes, postal and zip codes.
Examples: “*HU21*” for “*Tolna County, Hungary*” (FIPS region code); “*GUY*” for “*Guyana*” (ISO 3166-1 alpha-3); “*GY*” for “*Guyana*” (ISO 3166-1 alpha-2); “*EGLL*” for “*Heathrow Airport, London, UK*” or “*LPBJ*” for “*Beja Air Base, Beja, Portugal*” (International Civil Aviation Organization codes).
- Abbreviations and acronyms for a location name, including abbreviations of adjectives.
Examples: “*franz.*” for “*französisch*”/“*French*” (mapped to “*Frankreich*”/“*France*”); “*ital.*” for “*italienisch*”/“*Italian*” (“*Italien*”/“*Italy*”); “*Whv.*” for “*Wilhelmshaven*”; “*NRW*” for “*Nordrhein-Westfalen*”/“*North Rhine-Westphalia*”.
- Orthographic variants, including exonyms and historic names.
Examples: “*Cologne*” for “*Köln*”; “*Lower Saxony*” for “*Niedersachsen*”.
- Language names in the text.
Example: “*Portuguese*” for “*Portuguese speaking countries*” (mapped to “*Portugal, Angola, Cape Verde, East Timor, Mozambique, and Brazil*”).

¹The research described is part of the IRSAW project (Intelligent Information Retrieval on the Basis of a Semantically Annotated Web; LIS 4 – 554975(2) Hagen, BIB 48 HGfu 02-01), which is funded by the DFG (Deutsche Forschungsgemeinschaft).

- Meta-information for a document, i.e. the language a document is written in.
Example: “*Die Katze jagt die Maus*” for “*German language*” (mapped to “*Germany, Austria, and Switzerland*”).
- Unique entities associated with a geographic location, i.e. headquarters of an organization, persons, and buildings.
Examples: “*Boeing*” for “*Seattle, Washington*”; “*Molière*” for “*France*”; “*Galileo Galilei*” for “*Italy*”; “*Eiffel Tower*” for “*Paris*”; “*Pentagon*” for “*Washington, D.C.*”.
- The location names itself, including full names and short forms.
Example: “*Republik Korea*”/“*Republic of Korea*” for “*Südkorea*”/“*South Korea*”.

Typically, location indicators are not included in gazetteers, e.g. the morphology and lexical knowledge for adjectives is missing completely. Distinct location indicators contribute differently to the task of assigning a geographic scope to a document. Their importance depends on their usage and frequency in the corpus (e.g. adjectives are generally frequent) and the correctness of identifying them, because new ambiguities arise (e.g. the ISO 3166-1 code for Tuvalu (TV) is also the abbreviation for television).

2.2 Location Indicator Normalization

The normalization of location indicators to location names takes place on different levels of linguistic analysis in GIRSA.

- Character level: In all entries of the name lexicons, diacritical marks are replaced with non-accented characters to create orthographic variants of names. These resulting orthographic variants are used as elements of a synonym set and normalized by selecting a representative for the synonym set (synset).
Example: “*Québec*” → “*Quebec*”.
- Morphologic level: Inflectional endings for adjective and noun forms are identified and separated using a set of manually created rules and large lists of exceptions. Typical German inflectional endings of a word form (e.g. “-s”, “-es”, “-er”, “-en”, “e”) are removed before the lookup in name lexicons. (Note that location names usually do not have a plural form.)
More complex cases are multi-word expressions which may contain inflectional morphology. Morphologic variations of location names are reduced to its base form.
Examples: “*Berlins*” → “*Berlin*”; “*das Rote Meer*” → “*Rote Meer*”; “*des Roten Meer(es)*” → “*Rote Meer*”.
Derivational morphology is part of connecting adjectives to location names.
Example: “*bayrisch*” → “*Bayern*”; “*dänisch*” → “*Dänemark*”.
- Semantic level: Prefixes indicating compass directions are separated from the name. A database management system may view the hyphenated result as either one or two terms, depending on the search options. Thus, a search for “*Norddeutschland*” will also return documents containing the phrase “*im Norden Deutschlands*”. Also on the semantic level, a mapping between location indicators and location names takes place.
Examples: “*Norddeutschland*” → “*Nord-Deutschland*”; “*Süd-Frankreich*” → “*Süd-Frankreich*”; exception: “*Südafrika*” → “*Südafrika*”.
- Lexical level: Name variations are normalized using synset representatives. The synsets contain elements referencing the same geographic location.
Example: “*Burma*”, “*Birma*” → “*Myanmar*”.

Of course, there is an implicit ordering of normalization steps: morphological variations are identified first, removing inflectional endings before lookup. Then, complex named entities are recognized and represented as a single term. Next, adjectives and acronyms are mapped to the expanded location name. Normalization by mapping to a synset representative is the last operation.

2.3 Semantic Analysis for GIR

This year, the approach of semantic representation matching (GIR-InSicht, derived from the deep QA system InSicht, [5]) was tried again for GeoCLEF. See [8] for details on the first experiment in this direction at GeoCLEF 2005. GIR-InSicht matches reduced semantic representations of the topic description (or topic title) to the semantic representations of sentences from the document collection. This process is quite strict and proceeds sentence by sentence.² Before matching starts, the query semantic network was allowed to be split in parts at specific semantic relations, e.g. at a LOC relation (location of a situation or object) of the MultiNet formalism (multilayered extended semantic networks; [6]), to increase recall while not losing too much precision.

For GeoCLEF 2007, *query decomposition* was implemented, i.e. a query can be decomposed into two dependent queries, the subquery and the main query. The subquery was answered by the QA system InSicht; the answers were integrated into the main query on the semantic network level (thereby avoiding the complicated or problematic integration on the surface level). For example, the title of topic 10.2452/57-GC “*Whiskyherstellung auf den schottischen Inseln*” (‘*Whiskey production on the Scottish Islands*’) and similarly the description of this topic lead to the subquery “*Nenne schottische Inseln*” (‘*Name Scottish islands*’). Decomposition is also applied to the alternative query semantic networks derived by inferential query expansion. In the above example, this leads to the subquery “*Nenne Inseln in Schottland*” (‘*Name islands in Scotland*’). InSicht answers the subqueries on the semantic representations of the GeoCLEF document collection and the German Wikipedia. For the above subqueries, it correctly delivered islands like “*Iona*” and “*Islay*”, which in turn lead to main query semantic networks which could be paraphrased as “*Whiskyherstellung auf Iona*” (‘*Whiskey production on Iona*’) and “*Whiskyherstellung auf Islay*” (‘*Whiskey production on Islay*’). Note that the decomposed queries are processed only as alternatives to the original query.

Another decomposition strategy produces questions aiming at meronymy knowledge based on the geographical type of a location, e.g. for a country *C* in the original query a subquery like “*Name cities in C.*” is generated, whose results are integrated into the main query semantic network. This strategy led to interesting questions like “*Welcher Staat/Welche Region/Welche Stadt liegt im Himalaya?*” (‘*Which country/region/city is located in the Himalaya?*’). In total, both decomposition strategies led to 80 different subqueries for the 25 topics. After the title and description of a topic have been processed independently, GIR-InSicht combines the results. If a document occurs in the title results and the description results, the highest score was taken for the combination.

The semantic matching approach is completely independent of the main approach in GIRSA. Some of the functionality of the main approach is also realized in the matching approach, e.g. some of the location indicators described above are also exploited in GIR-InSicht (adjectives; demonymy for regions and countries). They are not normalized, but the query semantic network is extended by many alternative semantic networks that are in part derived by symbolic inference rules using the semantic knowledge about location indicators. In contrast, the main approach exploits this information on the level of terms.

2.4 Related Work

Nagel [11] describes the manual construction of a place name ontology containing 17,000 geographic entities as a prerequisite for analyzing German sentences. He states that in German, toponyms have a simple inflectional morphology, but a complex (idiosyncratic) derivational morphology.

Buscaldi, Rosso et al. [1] investigate semi-automatic creation of a geographical ontology, using gazetteer data and resources like Wikipedia and WordNet.

Wang et al. [12] introduce the concept of dominant locations (later called implicit locations, [10]). Implicit locations are locations not explicitly mentioned in a text. The only case explored are locations that are closely related to other locations.

Previous work on GIR by members of the IICS includes experiments with documents and queries represented as semantic networks [8], and experiments dealing with linguistic phenomena, such as cases of regular metonymy of location names [7], which was utilized as a means to increase precision in GIR. Due

²But documents can also be found if the information is distributed across several sentences because a coreference resolver processed all document representations.

to time constraints, metonymy recognition was not included in GIRSA. For the experiments for GeoCLEF 2007, we focused on investigating means to increase recall.

3 Experimental Setup

The GeoCLEF 2007 documents constitute a corpus of more than 275,000 German newspaper articles from 'Frankfurter Rundschau', 'Schweizerische Depeschenagentur', and 'Der Spiegel' from the years 1994 and 1995 (see [3]). The performance of GIRSA is evaluated on the test set from GeoCLEF 2007, containing 25 topics with a title, a short description, and a narrative part. As in a setup for previous GIR experiments on GeoCLEF data [8, 9], the documents were indexed with the Zebra database management system [4], which supports a standard relevance ranking (*tf-idf* IR model).

Documents are preprocessed as follows to produce different indexes:

1. S: As in traditional IR, all words in the document text (including location names) are stemmed, using an implementation of the German snowball stemmer.
2. SL: Location indicators are identified and normalized to a base form of a location name.
3. SLD: In addition, decomposing is applied to the words in the text. German decomposing follows the frequency-based approach described in [2].
4. O: Documents and queries are represented as semantic networks and GIR is seen as a form of question answering (see Sect. 2.3).

The following location indicators were normalized in documents and queries for the GIR experiments: adjectives corresponding to locations, demonyms, abbreviations, orthographic variants, language names, and location names. Normalization consists of applying a set of transformation rules (covering regular variations) and looking up locations in specialized exception lists for each type of location indicator.

Basically, queries and documents are processed in the same way. The title and short description were used for creating a query. GeoCLEF topics contain a narrative part describing documents which are to be assessed as relevant. Instead of employing a large gazetteer containing location names as a knowledge base for query expansion, additional location names were extracted from the narrative part of the topic. No meronymy information is utilized for direct query expansion (because these may be just the terms the blind feedback may find and there would be a combined effect).

For the bilingual (English-German) experiments, the queries were translated using the Promt web service for machine translation.³ Query processing then follows the setup for monolingual German experiments.

The following parameter settings were used in different retrieval experiments:

1. query language: German (DE) or English (EN);
2. index type: stemming only (S), identification of locations, not stemmed (SL), decomposition of German compounds (SLD), hybrid (SLD/O), and based on semantic networks (O);
3. query fields: combinations of title (T), description (D), and locations from narrative (N).

Parameters and results for monolingual German and bilingual English-German experiments are shown in Table 1. The table shows relevant and retrieved documents (*rel_ret*), MAP and precision at five, ten, and twenty documents. In total, 904 documents were assessed as relevant for the 25 topics. For the run FUHtd6de, results from GIR-InSicht were merged with results from the experiment FUHtd3de in a straightforward way, using the maximum score.

³<http://www.e-prompt.com/>

Table 1: Results for different retrieval experiments on German GeoCLEF 2007 data.

Run	Parameters			Results				
	query language	index	fields	rel_ret	MAP	P@5	P@10	P@20
FUHtd1de	DE	S	TD	597	0.119	0.280	0.256	0.194
FUHtd2de	DE	SL	TD	707	0.191	0.288	0.264	0.254
FUHtd3de	DE	SLD	TD	677	0.190	0.272	0.276	0.260
FUHtdn4de	DE	SL	TDN	722	0.236	0.328	0.288	0.272
FUHtdn5de	DE	SLD	TDN	717	0.258	0.336	0.328	0.288
FUHtd6de	DE	SLD/O	TD	680	0.196	0.280	0.280	0.260
GIR-InSicht	DE	O	TD	52	0.067	0.104	0.096	0.080
FUHtd1en	EN	S	TD	490	0.114	0.216	0.188	0.162
FUHtd2en	EN	SL	TD	588	0.146	0.272	0.220	0.196
FUHtd3en	EN	SLD	TD	580	0.145	0.224	0.180	0.156
FUHtdn4en	EN	SL	TDN	622	0.209	0.352	0.284	0.246
FUHtdn5en	EN	SLD	TDN	619	0.188	0.272	0.256	0.208

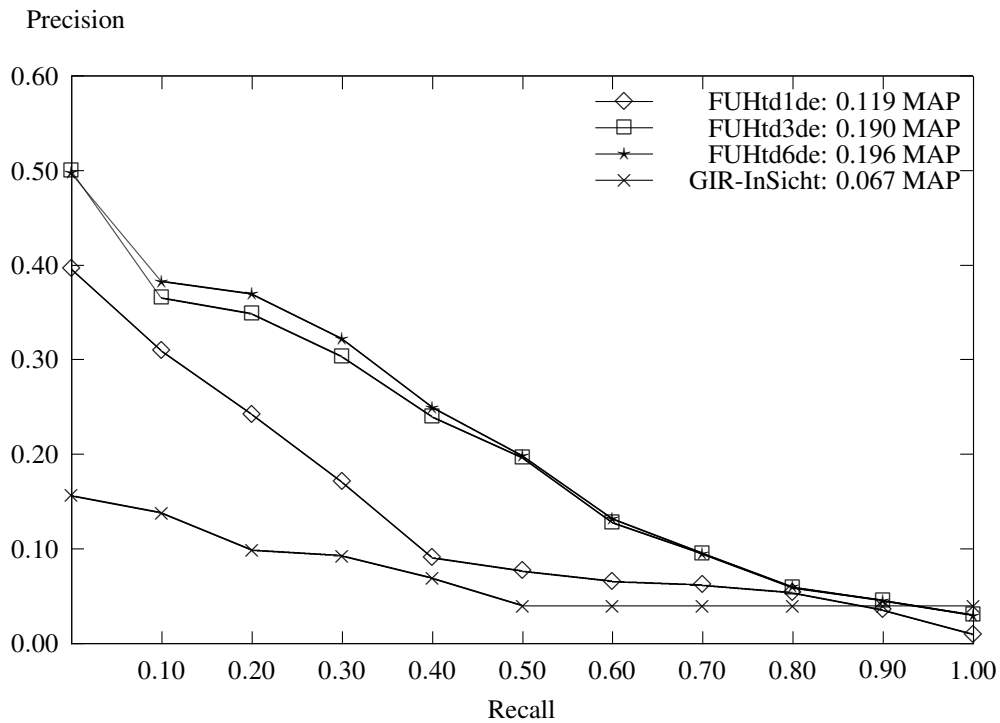


Figure 1: Recall-precision graph for German monolingual GeoCLEF experiments.

Table 2: Comparison of results per topic for different German monolingual runs.

topic	FUHtd3de		GIR-InSicht		FUGtd6de	
	rel_ret/rel	MAP	rel_ret/rel	MAP	rel_ret/rel	MAP
10.2452/56-GC	2/ 5	0.002	1/ 5	0.029	2/ 5	0.030
10.2452/57-GC	1/ 1	0.200	1/ 1	1.000	1/ 1	0.200
10.2452/61-GC	45/ 46	0.530	21/ 46	0.305	45/ 46	0.492
10.2452/65-GC	83/ 95	0.348	3/ 95	0.007	83/ 95	0.291
10.2452/66-GC	1/ 2	0.017	0/ 2	0.000	1/ 2	0.017
10.2452/68-GC	203/269	0.380	0/269	0.000	203/269	0.380
10.2452/69-GC	20/ 45	0.123	2/ 45	0.006	20/ 45	0.111
10.2452/70-GC	17/ 23	0.115	0/ 23	0.000	17/ 23	0.115
10.2452/72-GC	4/ 10	0.022	3/ 10	0.160	7/ 10	0.208
10.2452/75-GC	83/110	0.440	21/114	0.173	83/110	0.460

4 Results and Discussion

Identifying and indexing normalized location indicators, decomposing, and adding location names from the narrative part improves performance considerably, i.e. 120 additional relevant documents are found and MAP is increased from 0.119 (FUHtd1de) to 0.258 (FUHtdn5de) in comparison to the baseline experiment.

Decomposing German nouns seems to have different effects on precision and recall (FUHtd2de vs. FUHtd3de and FUHtdn4de vs. FUHtdn5de): while more relevant documents are retrieved without decomposing, initial precision is higher when utilizing decomposing.

Topic 10.2452/55-GC contains a negation in the topic title and description (“*but not in the Alps*”). However, adding the location names from the narrative part of the topic (“*Scotland, Norway, Iceland*”) did not notably improve precision for this topic (0.005 MAP in FUGtd3de vs. 0.013 MAP in FUHtdn5de).

A small analysis of results found by GIR-InSicht in comparison with the main GIR system reveals that for ten topics, GIR-InSicht retrieved documents, and for seven topics, it returned relevant documents (see Fig. 1 and Table 2). This approach, originating from question answering and based on a strict matching of semantic representations, returns three additional relevant documents for the combination (FUHtd6de). However, the MAP for some topics in the combined run indicates that merging by taking the maximum of two scores might be too simple. For a single topic (10.2452/52-GC), zero relevant documents were retrieved in all experiments.

Results for the bilingual (English-German) experiments are generally lower. As for German, all other experiments outperform the baseline (0.114 MAP). The best performance is achieved by an experiment using topic title, description, and location names from the narrative (0.209 MAP). In comparison with results for the monolingual German experiments, the performance drop lies between 4.2% (first experiment) and 27.1% (fifth experiment).

5 Conclusion and Outlook

In this paper location indicators were introduced as text segments from which location names can be inferred. For the GeoCLEF 2007 experiments, different indexes containing stemmed words and location indicators normalized to location names were created. Results of the GIR experiments show that MAP is higher when using location indicators instead of location names to represent the geographic scope of a document. A broader approach to identify the geographic scope of a document is needed because proper nouns or location names do not alone imply the geographic scope of a document.

In addition, we investigated using location names extracted from the narrative part of a topic (instead of looking up additional location names in large gazetteers). The narrative contains a detailed description about which documents are to be assessed as relevant (and which not), including additional location names. Adding these location names to the query notably improves performance. This result is seemingly in con-

trast to some results from GeoCLEF 2006, were it was found that additional query terms (from gazetteers) degrade performance. A possible explanation is that in this experiment, only a few location names were added (3.16 location names on average for fifteen of the 25 topics with a maximum of thirteen additional location names). When using a gazetteer, one has to decide which terms are the most useful in query expansion. If this decision is based on the importance of a location, a semantic shift in the results may occur, which degrades performance. In contrast, selecting terms from the narrative part increases the chance to expand a query with relevant terms only.

The hybrid approach for GIR proved interesting, and even a few additional relevant documents were found in the combined run. As GIR-InSicht originates from a deep (read: semantic) QA approach, it returns documents with a high initial precision, which may prove useful in combination with a geographic blind feedback strategy. GIR-InSicht performs worse than the IR baseline, because only 102 documents were retrieved for ten of the 25 topics. However, more than half (56 documents) turned out to be relevant.

Several improvements are planned for GIRSA. These include using estimates for the importance (weight) of different location indicators, possibly depending on the context (e.g. “*Danish coast*” → “*Denmark*”, but “*German shepherd*” ↯ “*Germany*”), and using a part-of-speech tagger and named entity recognizer to identify location names. Finally, we plan to investigate the combination of means to increase precision (e.g. recognizing metonymic location names) with means to increase recall (e.g. recognizing and normalizing location indicators).

References

- [1] Davide Buscaldi, Paolo Rosso, and Piedachu Peris Garcia. Inferring geographical ontologies from multiple resources for geographical information retrieval. In *Proceedings of the 3rd Workshop on Geographical Information Retrieval (GIR 2006)*, pages 52–55, Seattle, USA, 2006.
- [2] Aitao Chen. Cross-language retrieval experiments at CLEF 2002. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*, volume 2785 of *LNCS*, pages 28–48, Berlin, 2002. Springer.
- [3] Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *LNCS*, pages 908–919. Springer, Berlin, 2006.
- [4] Sebastian Hammer, Adam Dickmeiss, Heikki Levanto, and Mike Taylor. *Zebra – User’s Guide and Reference*. Copenhagen, Denmark, 2005.
- [5] Sven Hartrumpf and Johannes Leveling. University of Hagen at QA@CLEF 2006: Interpretation and normalization of temporal expressions. In Alessandro Nardi, Carol Peters, and José Luis Vicedo, editors, *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006.
- [6] Hermann Helbig. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, 2006.
- [7] Johannes Leveling and Sven Hartrumpf. On metonymy recognition for GIR. In *Proceedings of the 3rd Workshop on Geographical Information Retrieval (GIR 2006)*, pages 9–13, Seattle, USA, 2006.
- [8] Johannes Leveling, Sven Hartrumpf, and Dirk Veiel. Using semantic networks for geographic information retrieval. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, Henning Müller, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *LNCS*, pages 977–986. Springer, Berlin, 2006.

- [9] Johannes Leveling and Dirk Veiel. University of Hagen at GeoCLEF 2006: Experiments with metonymy recognition in documents. In Alessandro Nardi, Carol Peters, and José Luis Vicedo, editors, *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006.
- [10] Zhisheng Li, Chong Wang, Xing Xie, Xufa Wang, and Wei-Ying Ma. Indexing implicit locations for geographical information retrieval. In *Proceedings of the 3rd Workshop on Geographical Information Retrieval (GIR 2006)*, pages 68–70, Seattle, USA, 2006.
- [11] Sebastian Nagel. An ontology of German place names. *Corela – Cognition, Représentation, Langage, Le traitement lexicographique des noms propres* (Numéros spéciaux), 2005.
- [12] Lee Wang, Chuang Wang, Xing Xie, Josh Forman, Yansheng Lu, Wei-Ying Ma, and Ying Li. Detecting dominant locations from search queries. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '05)*, pages 424–431, New York, USA, 2005. ACM Press.