



Introduction

- Traditional information retrieval (IR): stemming is applied to all words in a text
- Geographical information retrieval (GIR): use named entity recognition and classification; avoid stemming location names
- Typically, locations are indicated by proper nouns only
- GIRSA (Geographic Information Retrieval by Semantic Annotation): aims at a broader GIR approach not solely based on location names, but on location indicators

Location Indicators

Definition: Location indicators are text segments from which the geographic scope of a document can be inferred.

- Adjectives corresponding to a location.
Example: *tunesisch/Tunisian* for *Tunesien/Tunisia*
- Demonyms, e.g. the name for inhabitants originating from a location.
Example: *Franzose, Französin/Frenchman, Frenchwoman* for *Frankreich/France*
- Codes for a location name.
Example: *HU21* for *Tolna County, Hungary* (FIPS region code)
- Abbreviations and acronyms for a location name, including adjectives.
Example: *franz.* for *französisch/French* (mapped to *Frankreich/France*);
- Orthographic variants, exonyms, historic names.
Example: *Lower Saxony* for *Niedersachsen*.
- Unique entities associated with a geographic location, e.g. headquarters of an organization, persons, buildings.
Example: *Molière* for *France*;
- The location names itself (full names and short forms).
Example: *Republik Korea/Republic of Korea* for *Südkorea/South Korea*.

Location Indicator Normalization

- Character level: diacritical marks are replaced with non-accented characters to create synonym sets of orthographic variants (synsets); elements of a synset are normalized by selecting a representative for the synonym set.
Example: *Québec* → *Quebec*.
- Morphologic level: Inflectional endings are identified and separated using a set of manually created rules and large lists of exceptions. Typical German inflectional endings of a word form are removed. Multi-word expressions may contain inflectional morphology. Morphologic variations of location names are reduced to their base form.
Examples: *Berlins* → *Berlin*; *das Rote Meer* → *Rote Meer*; *des Roten Meer(e)s* → *Rote Meer*. Derivational morphology is part of associating adjectives with location names.
Example: *bayrisch* → *Bayern*; *dänisch* → *Dänemark*.
- Semantic level: Prefixes indicating compass directions are separated from the name. Location indicators are mapped to location names.
Examples: *Norddeutschland* → *Nord-Deutschland*; exception: *Südafrika* → *Südafrika*.
- Lexical level: Name variations are normalized using synset representatives.
Example: *Burma, Birma* → *Myanmar*.

Semantic Analysis for GIR

- Extension of semantic representation matching approach, GIR-InSicht (Leveling et al. (2006)), derived from the deep question answering (QA) system InSicht (Hartrumpf and Leveling (2007))
- GIR-InSicht matches reduced semantic representations of the topic description or topic title with the semantic representations of sentences from the document collection
- 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; Helbig (2006))
- Query decomposition* was implemented, i.e. a query can be decomposed into two dependent queries, the subquery and the main query
- The subquery is answered by the QA system InSicht; the answers were integrated into the main query on the semantic network level

Example:

- Whiskyherstellung auf den schottischen Inseln* ('Whiskey production on the Scottish Islands'; topic 10.2452/57-GC)
 - subquery *Nenne schottische Inseln* ('Name Scottish islands') (by query decomposition)
 - subquery *Nenne Inseln in Schottland* ('Name islands in Scotland') (by inferential query expansion)
 - partial answers *Iona* and *Islay* (by answering the subqueries on the GeoCLEF corpus and the German Wikipedia)
 - new queries *Whiskyherstellung auf Iona* ('Whiskey production on Iona') and *Whiskyherstellung auf Islay* ('Whiskey production on Islay') (by query paraphrasing)
 - In total, 80 different subqueries were produced for the 25 topics

Experimental Setup

- GeoCLEF 2007 documents: 275,000 German newspaper articles from *Frankfurter Rundschau*, *Schweizerische Depeschagentur*, and *Der Spiegel* from the years 1994 and 1995 (see Gey et al. (2006))
- performance of GIRSA is evaluated on 25 topics from from GeoCLEF 2007 with a title, a short description, and a narrative part
- Setup similar to previous GIR experiments on GeoCLEF data Leveling et al. (2006); Leveling (2007)
- Different indexes:
 - S: All words in the document text are stemmed
 - SL: Location indicators are identified and normalized to a base form of a location name
 - SLD: In addition, decomposing is applied to the words in the text
 - O: Documents and queries are represented as semantic networks and GIR is seen as (a form of) QA
- Queries and documents are processed in the same way
- Simple processing: No named entity recognizer; no large gazetteer for query expansion; no blind feedback
- For bilingual (English-German) experiments: translate queries using the Prompt web service for machine translation (<http://www.e-prompt.com/>)

The following parameter settings were used in different retrieval experiments:

- query language: German (DE), English (EN) – cf. run ID
- 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);
- query fields: combinations of title (T), description (D), and locations from narrative (N).

Results and Discussion

Table 1: Results for German GeoCLEF 2007 data.

Run	Parameters		Results		
	index	fields	rel_ret	MAP	P@5
FUHtd1de	S	TD	597	0.119	0.280
FUHtd2de	SL	TD	707	0.191	0.288
FUHtd3de	SLD	TD	677	0.190	0.272
FUHtdn4de	SL	TDN	722	0.236	0.328
FUHtdn5de	SLD	TDN	717	0.258	0.336
FUHtd6de	SLD/O	TD	680	0.196	0.280
GIR-InSicht	O	TD	52	0.067	0.104
FUHtd1en	S	TD	490	0.114	0.216
FUHtd2en	SL	TD	588	0.146	0.272
FUHtd3en	SLD	TD	580	0.145	0.224
FUHtdn4en	SL	TDN	622	0.209	0.352
FUHtdn5en	SLD	TDN	619	0.188	0.272

Precision

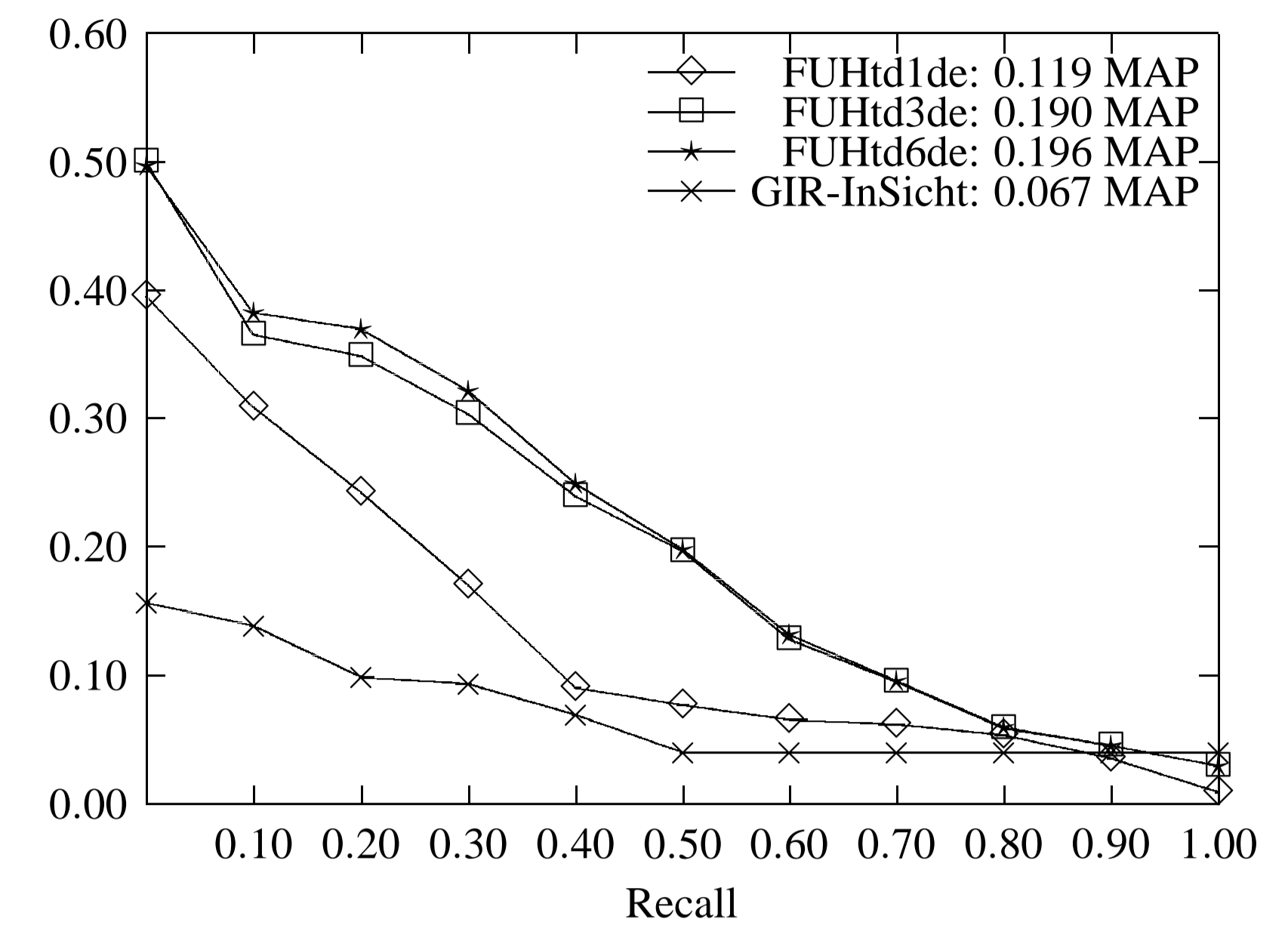


Figure 1: Recall-precision graph for German monolingual GeoCLEF experiments.

Conclusion and Outlook

- Baseline run (FUHtd1de) is clearly outperformed
- Adding location names from the narrative notably improves performance
- Hybrid approach for GIR proved interesting, and even a few additional relevant documents were found
- Planned improvements for GIRSA:
 - estimate the importance (weight) of different location indicators, possibly depending on the context (e.g. *Danish coast* → *Denmark*, but *German shepherd* → *Germany*)
 - apply part-of-speech tagger and named entity recognizer to identify location names
 - investigate the combination of means to increase precision (e.g. recognizing metonymic uses of location names) with means to increase recall (e.g. recognizing and normalizing location indicators)

References

- Gey, Fredric; Ray Larson; Mark Sanderson; Hideo Joho; Paul Clough; and Vivien Petras (2006). GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005* (edited by Peters, Carol; Fredric C. Gey; Julio Gonzalo; Henning Müller; Gareth J. F. Jones; Michael Kluck; Bernardo Magnini; and Maarten de Rijke), volume 4022 of *LNCS*, pp. 908–919. Berlin: Springer.
- Hartrumpf, Sven and Johannes Leveling (2007). Interpretation and normalization of temporal expressions for question answering. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum* (edited by Peters, Carol; Paul Clough; Fredric C. Gey; Jussi Karlgren; Bernardo Magnini; Douglas W. Oard; Maarten de Rijke; and Maximilian Stempfhuber), volume 4730 of *LNCS*, pp. 432–439. Berlin: Springer.
- Helbig, Hermann (2006). *Knowledge Representation and the Semantics of Natural Language*. Berlin: Springer.
- Leveling, Johannes (2007). Experiments on the exclusion of metonymic location names from GIR. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum* (edited by Peters, Carol; Paul Clough; Fredric C. Gey; Jussi Karlgren; Bernardo Magnini; Douglas W. Oard; Maarten de Rijke; and Maximilian Stempfhuber), volume 4730 of *LNCS*, pp. 901–904. Berlin: Springer.
- Leveling, Johannes; Sven Hartrumpf; and Dirk Veiel (2006). Using semantic networks for geographic information retrieval. In *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum* (edited by Peters, Carol; Fredric C. Gey; Julio Gonzalo; Gareth J. F. Jones; Michael Kluck; Bernardo Magnini; Henning Müller; and Maarten de Rijke), volume 4022 of *LNCS*, pp. 977–986. Berlin: Springer.

Acknowledgement

The 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).