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

CLEF 2007 Workshop, Budapest, Hungary

Outline

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, proper nouns only); employ geographic knowledge
- 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.

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 $\rightarrow$ *Tunesien*
(*Tunisian* $\rightarrow$ *Tunisia*)

Location Indicators

Definition

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

- Demonyms, e.g. the name for inhabitants originating from a location.

Example:
Franzose, Französin $\rightarrow$ *Frankreich*
(*Frenchman, Frenchwoman* $\rightarrow$ *France*)

Location Indicators

Definition

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

- Codes for a location name.

Example:
HU21 $\rightarrow$ *Tolna County, Hungary* (FIPS region code)

Location Indicators

Definition

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

- Abbreviations and acronyms for a location name, including adjectives.

Example:
franz. $\rightarrow$ *französisch* $\rightarrow$ *Frankreich*
(*French* $\rightarrow$ *France*)
TX $\rightarrow$ *Texas*

Location Indicators

Definition

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

- Orthographic variants, exonyms, historic names.

Example:
Lower Saxony $\rightarrow$ *Niedersachsen*

Location Indicators

Definition

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

- Unique entities associated with a geographic location, e.g. headquarters of an organization, persons, buildings.

Example:
Eiffel Tower $\rightarrow$ *Paris*
Moliére $\rightarrow$ *France* (?)
VW $\rightarrow$ *Wolfsburg* (?)

Location Indicators

Definition

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

- The location names itself (full names and short forms).

Example:
Republik Korea $\rightarrow$ *Südkorea*
(*Republic of Korea* $\rightarrow$ *South Korea*)

Location Indicator Normalization

Normalization on surface (character), morphologic, syntactic, semantic, and lexical level.

Character level

- Diacritical marks replaced with non-accented characters
- Orthographic variants normalized by selecting a representative

Example:
Québec $\rightarrow$ *Quebec*

Location Indicator Normalization

Normalization on surface (character), morphologic, syntactic, semantic, and lexical level.

Morphologic level

- Inflectional endings are identified and removed
- Morphologic variations of location names are reduced to their base form
- Derivational morphology: adjective →location name

Examples:
des Roten Meer(e)s →Rote Meer
bayrisch →Bayern
dänisch →Dänemark

Location Indicator Normalization

Normalization on surface (character), morphologic, syntactic, semantic, and lexical level.

Semantic level

- Prefixes are separated from the name
- Location indicators are mapped to location names

Examples:
Norddeutschland $\rightarrow$*Nord-Deutschland*
exception:
Südafrika $\rightarrow$*Südafrika*

Location Indicator Normalization

Normalization on surface (character), morphologic, syntactic, semantic, and lexical level.

Lexical level

- Name variations are normalized using synset representatives

Example:
Burma $\rightarrow$ *Myanmar*
Birma $\rightarrow$ *Myanmar*

Semantic Analysis for GIR

- Extension of semantic network matching approach, GIR-InSicht (Leveling et al. (2006)), derived from the deep question answering (QA) system InSicht (Hartrumpf and Leveling (2007))
- Query semantic network was allowed to be split in parts at specific semantic relations, e.g. at a LOC(ATION) relation
- *Query decomposition*: a query can be decomposed into two dependent queries, the subquery and the main query
- The subquery is answered by the QA system InSicht; answers are integrated into the main query

Semantic Analysis Example

Topic 10.2452/57-GC

Whiskyherstellung auf den schottischen Inseln/
"Whiskey production on the Scottish Islands"

Inferential query expansion followed by query decomposition

→Subquery *Nenne schottische Inseln/*
"Name Scottish islands"
Subquery *Nenne Inseln in Schottland/*
"Name islands in Scotland" (inferences)

Semantic Analysis Example

Topic 10.2452/57-GC

Whiskyherstellung auf den schottischen Inseln/
"Whiskey production on the Scottish Islands"

Answering the subqueries on the GeoCLEF corpus and the German Wikipedia

→Partial answers *Iona* and *Islay*
→Better gazetteer entry points

Semantic Analysis Example

Topic 10.2452/57-GC

Whiskyherstellung auf den schottischen Inseln/
"Whiskey production on the Scottish Islands"

New queries (paraphrased)

→New queries *Whiskyherstellung auf Iona/*
"Whiskey production on Iona"
and *Whiskyherstellung auf Islay/*
"Whiskey production on Islay"

Semantic Analysis Example

Topic 10.2452/57-GC

Whiskyherstellung auf den schottischen Inseln/
"Whiskey production on the Scottish Islands"

→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 Depeschenagentur*, and *Der Spiegel* from the years 1994 and 1995
- GIRSA evaluated on 25 GeoCLEF topics with a title (T), a short description (D), and a narrative part (N)
- Setup similar to previous GIR experiments on GeoCLEF data Leveling et al. (2006); Leveling (2007)

Methods for GIR (1/3)

PoS-Tagger/NERC (TnT, Lingpipe etc.):

- Andogah, Bouma et al. (U. Groningen)
- Buscaldi, Rosso (U. Valencia)
- Ferrés, Rodríguez (U. Catalunya)
- Kölle, Heuwing et al. (U. Hildesheim)
- Lana-Serrano, Villena-Román et al. (U. Madrid)
- Overell, Magalhães et al. (IC London)
- Perea-Ortega, García-Cumbreras et al. (U. Jaén)

List lookup:

- Leveling, Hartrumpf (U. Hagen)
- Larson (U. C. Berkeley)

→Only part of the solution, but GIRSA needs this, too!

Methods for GIR (1/3)

PoS-Tagger/NERC (TnT, Lingpipe etc.):

- Andogah, Bouma et al. (U. Groningen)
- Buscaldi, Rosso (U. Valencia)
- Ferrés, Rodríguez (U. Catalunya)
- Kölle, Heuwing et al. (U. Hildesheim)
- Lana-Serrano, Villena-Román et al. (U. Madrid)
- Overell, Magalhães et al. (IC London)
- Perea-Ortega, García-Cumbreras et al. (U. Jaén)

List lookup:

- Leveling, Hartrumpf (U. Hagen)
- Larson (U. C. Berkeley)

→Only part of the solution, but GIRSA needs this, too!

Methods for GIR (2/3)

Gazetteers/GKB (GNS, WordNet etc.):

- Andogah, Bouma et al. (U. Groningen)
- Buscaldi, Rosso (U. Valencia)
- Cardoso, Cruz et al. (U. Lisbon)
- Ferrés, Rodríguez (U. Catalunya)
- Guillén (CSU)
- Lana-Serrano, Villena-Román et al. (U. Madrid)
- Larson (U. C. Berkeley)
- Li, Wang et al. (Microsoft Asia)
- Nasikhin, Adriani (U. Indonesia)
- Overell, Magalhães et al. (IC London)

Small name lists (about 250,000 entries):

- Leveling, Hartrumpf (U. Hagen)

→GIRSA does not use geographic knowledge, yet.

Methods for GIR (2/3)

Gazetteers/GKB (GNS, WordNet etc.):

- Andogah, Bouma et al. (U. Groningen)
- Buscaldi, Rosso (U. Valencia)
- Cardoso, Cruz et al. (U. Lisbon)
- Ferrés, Rodríguez (U. Catalunya)
- Guillén (CSU)
- Lana-Serrano, Villena-Román et al. (U. Madrid)
- Larson (U. C. Berkeley)
- Li, Wang et al. (Microsoft Asia)
- Nasikhin, Adriani (U. Indonesia)
- Overell, Magalhães et al. (IC London)

Small name lists (about 250,000 entries):

- Leveling, Hartrumpf (U. Hagen)

→GIRSA does not use geographic knowledge, yet.

Methods for GIR (3/3)

Blind Feedback:

- Cardoso, Cruz et al. (U. Lisbon)
- Ferrés, Rodríguez (TALP) – Relevance Feedback
- Guillén (CSU)
- Kölle, Heuwing et al. (Hildesheim)
- Larson (U. C. Berkeley)
- Nasikhin, Adriani (U. Indonesia)
- Overell, Magalhães et al. (IC London)

No Blind Feedback:

- Leveling, Hartrumpf (U. Hagen)

→GIRSA will not utilize ad-hoc blind feedback!

Methods for GIR (3/3)

Blind Feedback:

- Cardoso, Cruz et al. (U. Lisbon)
- Ferrés, Rodríguez (TALP) – Relevance Feedback
- Guillén (CSU)
- Kölle, Heuwing et al. (Hildesheim)
- Larson (U. C. Berkeley)
- Nasikhin, Adriani (U. Indonesia)
- Overell, Magalhães et al. (IC London)

No Blind Feedback:

- Leveling, Hartrumpf (U. Hagen)

→GIRSA will not utilize ad-hoc blind feedback!

Experimental Setup

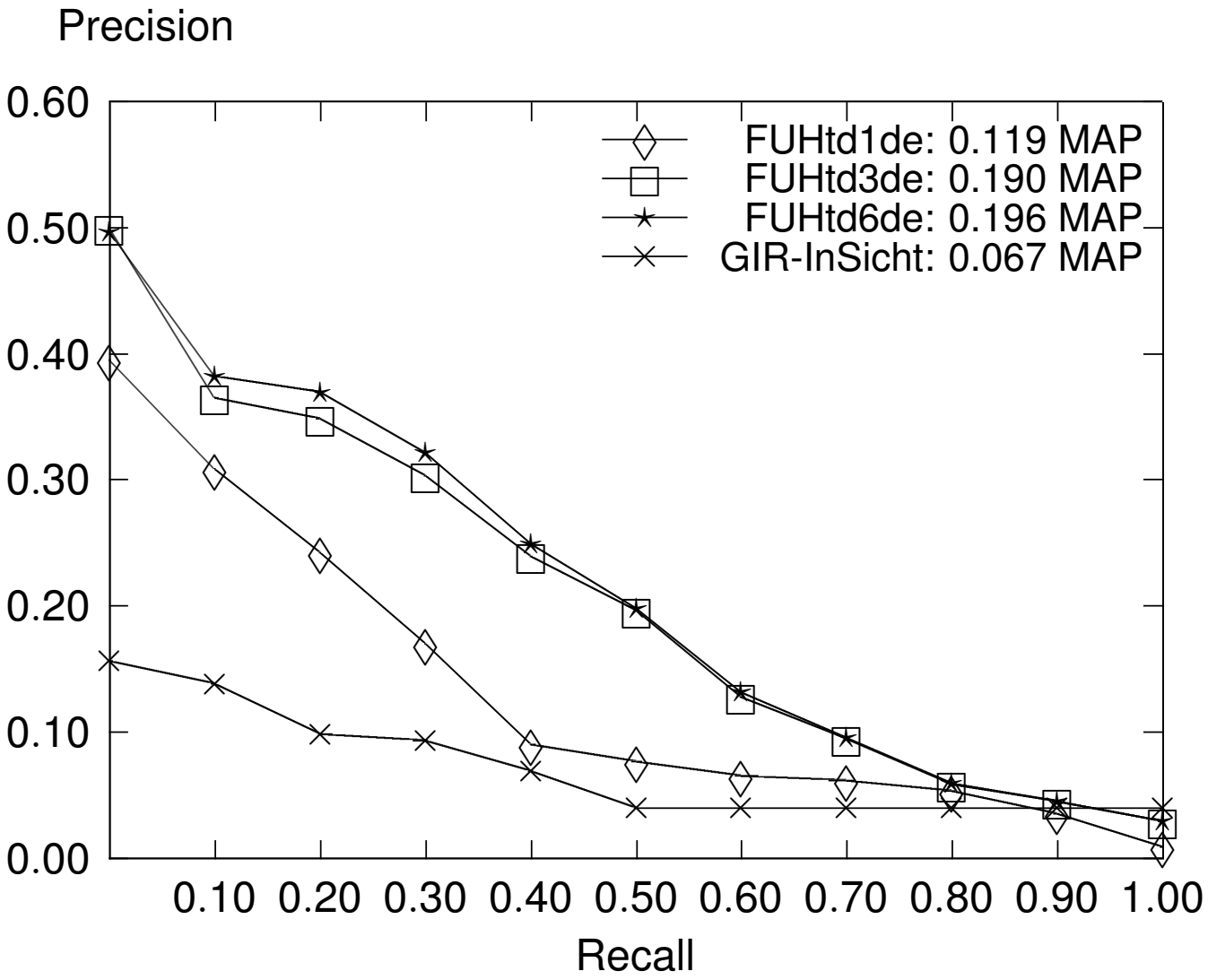
Different indexes:

- 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, decompounding 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

Results and Discussion

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

Results for Monolingual German



Conclusion

- Baseline run (FUHtd1de) is clearly outperformed
- Adding selected location names (from the narrative) notably improves performance
- Hybrid approach (with GIR-InSicht) for GIR proved interesting:
  even a few additional relevant documents were found

Outlook

Planned improvements for GIRSA:

- Estimate the importance (weight) of different location indicators, possibly depending on the context:
  Danish coast $\rightarrow$ *Denmark*, but
  German shepherd $\not\rightarrow$ *Germany*
- Apply part-of-speech tagger and named entity recognizer to identify location names
- Investigate the combination of means to increase precision (metonymic uses of location names) with means to increase recall (normalizing location indicators)

Selected References

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, CLEF 2006* (edited by et al., Carol Peters), volume 4730 of *LNCS*, pp. 432–439. 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, CLEF 2006* (edited by et al., Carol Peters), 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, CLEF 2005* (edited by et al., Carol Peters), volume 4022 of *LNCS*, pp. 977–986. Berlin: Springer.