University of Hagen at GeoCLEF 2005: Using Semantic Networks for Interpreting Geographical Queries

Johannes Leveling, Sven Hartrumpf, Dirk Veiel

Intelligent Information and Communication Systems (IICS)

University of Hagen (FernUniversität in Hagen)

58084 Hagen, Germany

{*Johannes.Leveling,Sven.Hartrumpf,Dirk.Veiel*} *@fernuni-hagen.de*

Abstract

The IICS group at the University of Hagen employs multilayered extended semantic networks for the representation of background knowledge, queries, and documents for geographic information retrieval (GIR). This paper describes our work for the participation at the GeoCLEF task of the CLEF 2005 evaluation campaign (Cross Language Evaluation Forum).

In our approach, geographical concepts from the query network are expanded with concepts which are semantically connected via topological, directional, and proximity relations. We started with an existing geographical knowledge base represented as a large semantic network and expanded it with concepts automatically extracted from the GEOnet Names Server (GNS). Furthermore, we created concept hypotheses by adding a prefix with regular semantics, for example *"Süd"/'South'* and *"Zentral"/'Central'*, and integrated the corresponding semantic relations into our geographical knowledge base.

Several experiments for GIR on German documents have been performed: a baseline corresponding to a traditional information retrieval approach; a variant expanding thematic, temporal, and geographic descriptors from the semantic network representation of the query; and an adaptation of a question answering (QA) algorithm based on semantic networks.

The second experiment is based on a representation of the natural language description of a topic as a semantic network, which is achieved by a deep linguistic analysis. The semantic network is transformed into an intermediate representation of a database query explicitly representing thematic, temporal, and local restrictions. This experiment showed the best performance with respect to mean average precision (MAP): 10.53 percent using the topic title and description or 10.22 percent using title, description, and additional location information.

The third experiment, adapting a QA algorithm, uses a modified version of the QA system InSicht. The system matches deep semantic representations (semantic networks) of queries or their equivalent or similar variants to semantic networks for document sentences. Since this approach was too much oriented towards precision, partitioning a query network was allowed when certain graph topologies exist. For example, local specifications can be split off, so that they can be matched in other sentences of the document under investigation. The geographical knowledge base developed for the other experiments improved the results of this approach, too. To keep answer time low and main memory consumption acceptable, some parameters of the InSicht system had to be adjusted.

In conclusion, we provide a basic architecture for further experiments in geographic information retrieval based on semantic networks. Future research aims at improving the named entity recognition for toponyms, connecting semantic networks and databases, expanding our geographical knowledge base, and investigating the role of semantic relations in geographic queries.

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)*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic networks*

General Terms

Measurement, Performance, Experimentation

Keywords

Geographic information retrieval, Query expansion

1 Introduction

Geographical Information Retrieval (GIR) is concerned with the retrieval of documents involving the interpretation of geographical knowledge by means of topological, directional, and proximity information. Documents typically contain descriptions of events or static situations that are temporally and/or spatially restricted. For example, consider the phrases *"the industrial development after World War II"* and *"the social security system outside of Scandinavia"*. Furthermore, many documents contain ambiguous geographic references. There are, for example, more than 30 cities named *"Zell"* in Germany, and any occurrence of this name in a document can have a different meaning and should be disambiguated from context. In addition, a toponym (a name for a geographic entity) can be referred to with names in different languages or local dialects, with historical names, etc., which will require normalization or translation to enable successful document retrieval. The latter problems are similar to the problems of polysemy and synonymy in traditional information retrieval (IR).

GeoCLEF is a task of the Cross Language Evaluation Forum offering scientific challenges in the interpretation of geographical information retrieval queries. The queries are targeted at existing CLEF document collections of news stories that include a variety of topics and geographical regions; for German, these are the news articles of *"Der Spiegel"*, *"Frankfurter Rundschau"*, and *"Schweizer Depeschenagentur"* from 1994 and 1995. The goal of the GeoCLEF task is to find all and only documents that are relevant to a given topic.

GeoCLEF topics include a short description (title and description), a longer narrative, and location elements consisting of a combination of thematic concepts (*DE-concept*), spatial relations (*DE-spatialrelation*), and place names (*DE-location*). For example, the short description of the first topic includes the natural language title *"Haifischangriffe vor Australien und Kalifornien"*/'Shark attacks off Australia and California'.

A traditional approach to GIR involves at least the following processing steps to identify geographical entities (Jones et al., 2002):

- named entity recognition (NER), including the detection of geographic names (tagging with named entities, including toponyms);

- collecting and integrating information from the contexts of named entities;

- disambiguation of named entities; and

- grounding the entities (i.e. connecting them to the model) and interpreting coordinates.

After identifying toponyms in queries and documents, coordinates can be assigned to them. In GIR, assigning a relevance score to a document for a given query typically involves calculating the distance between geographical entities in the query and the document and mapping it to a score.

Table 1: Overview of synonyms and word senses in GermaNet and GNS data for a selected subset of 169,407 geographical entities in Germany (DE). The data normalization consisted of removing all name variants introduced by the transcription of German umlauts (e.g. the name *"Koln"* is removed if it refers to the same entity as *"Köln"*).

Characteristic	Resource		
	GermaNet	GNS (DE, all)	GNS (DE, normalized)
synsets total	41,777	95,993	95,993
synonyms in synsets	60,646	121,055	103,508
unique literals	52,251	94,187	80,808
synonyms per synset	1.45	1.26	1.08
word senses per literal	1.16	1.29	1.28

One of the major problems for GIR is the disambiguation of toponyms from semantic context and identifying spatial ambiguity (e.g. *'California'* in *'Mexico'* and/or *'California'* in the *'United States of America'*). Table 1 shows a comparison of ambiguity in GermaNet[1] (Kunze and Wagner, 2001) and the GEOnet Names Server data (GNS, described in Section 2.2). As shown in the table, synonymy seems to be a lesser problem for German geographic names (1.08 synonyms per synset vs. 1.45 synonyms per synset in a lexical-semantic net for German), while the role of polysemy (word senses and disambiguation) becomes more important for GIR (1.28 vs. 1.16).

Problems that are less often identified and less investigated in research for GIR are:

- **Toponyms in different languages**. The translation of toponyms plays an important role even for monolingual retrieval when different and external information resources are integrated. In gazetteers, mostly English naming conventions are used.

- **Name variants**. The same geographic object can be referenced by endonymic names, exonymic names, and historical names. An endonym is a local name for a geographic entity, for example, *"Wien"*, *"Köln"*, and *"Milano"*. An exonym is a place name in a certain language for a geographic object that lies outside the region where this language has an official status; for example, *"Vienna"* is the English exonym for *"Wien"*, *"Cologne"* is the English exonym for *"Köln"*, and *"Mailand"* is the German exonym for *"Milano"*. Examples of historical names or traditional names are *"New Amsterdam"* for *"New York"* and *"Colonia Claudia Ara Agrippinensium"* or *"Cöllen"* for *"Köln"*. For GIR, name variants should be conflated.

- **Composite names**. Composite names or complex named entities consist of two or more words. Frequently, appositions are considered to be a part of a name. For example, there is no need for the translation of the word *"mount"* in *"Mount Cook"* (the German translation is *"Mount Cook"*), but *"Insel"* is typically translated in the expression *"Insel Sylt"*/*"island of Sylt"*. For named entity recognition, certain rules have to be established how composite names are normalized. In some composite names, two or more toponyms (geographic names) are employed in reference to a single entity, for example, *"Haren (Ems)"*, *"Frankfurt/Oder"*, or *"Freiburg im Breisgau"*. While additional toponyms in a context allow for a better disambiguation, such composite names require a normalization, too.

- **Semantic relations between toponyms and related concepts**. In named entity recognition and GIR, semantic relations between toponyms and related concepts are often ignored. Concepts related to a toponym such as the language, inhabitants of a place, properties (adjectives), or phrases (*"former Yugoslavia"*) are not considered in geographic tagging. For example, the toponym *"Scotland"* can be inferred for occurrences of *"Scottish"*, *"Scotsman"*, or *"Scottish districts"*.

---

[1] http://www.sfs.nphil.uni-tuebingen.de/lsd/Intro.html

c26$_d$
SUB **region**
$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{QUANT} & \text{one} \\ \text{REFER} & \text{det} \\ \text{CARD} & 1 \end{bmatrix}$

$\xrightarrow[\text{c} \quad \text{ATTR} \quad \text{c}]{}$

c27$_{na}$
SUB **name**
$\begin{bmatrix} \text{QUANT} & \text{one} \\ \text{CARD} & 1 \end{bmatrix}$

$\xrightarrow[\text{c} \quad \text{VAL} \quad \text{s}]{}$ **lateinamerika.0**$_{fe}$

s ↑ ***IN** c

c29$_l$
$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{QUANT} & \text{one} \\ \text{REFER} & \text{det} \\ \text{CARD} & 1 \end{bmatrix}$

c16$_{d\vee io}$
SUB **institution**
$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{QUANT} & \text{one} \\ \text{REFER} & \text{det} \\ \text{CARD} & 1 \end{bmatrix}$

$\xrightarrow[\text{c} \quad \text{ATTR} \quad \text{c}]{}$

c17$_{na}$
SUB **name**
$\begin{bmatrix} \text{QUANT} & \text{one} \\ \text{CARD} & 1 \end{bmatrix}$

s ↑ **LOC** s

s ↓ **ATTCH** s

c ↓ **VAL**

c20$_{as\vee io}$
PRED **menschenrecht**
$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{QUANT} & \text{mult} \\ \text{REFER} & \text{det} \end{bmatrix}$

$\xleftarrow[\text{MCONT}]{}$

c8?$_{ad\vee d\vee io}$
PRED **bericht**
$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{QUANT} & \text{mult} \\ \text{REFER} & \text{indet} \end{bmatrix}$
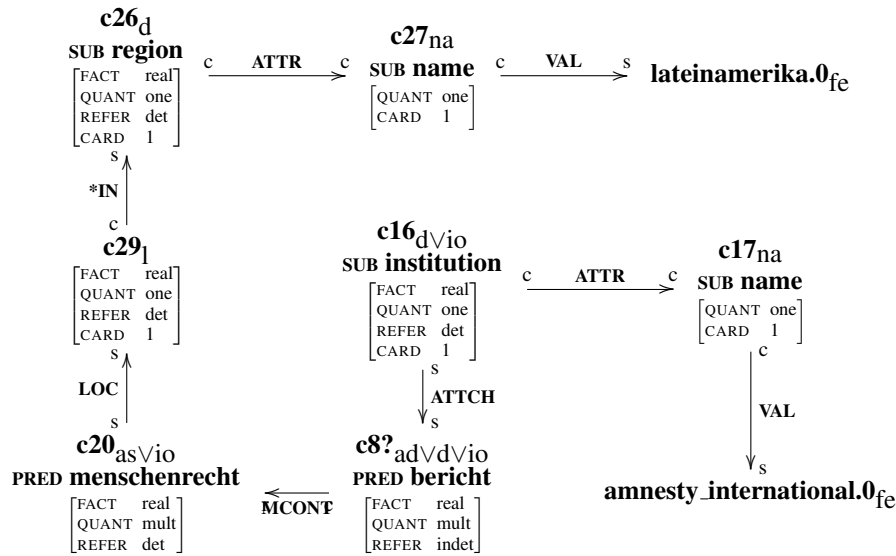
amnesty_international.0$_{fe}$

Figure 1: Automatically generated semantic network for the description of GeoCLEF topic GC003 (*"Finde Berichte von Amnesty International bezüglich der Menschenrechte in Lateinamerika."*, 'Amnesty International reports on human rights in Latin America.'). The relations are explained in Table 2. Note that nodes representing proper names bear a *.0* suffix, and subordinating edges (PRED, SUB) are *folded* below the node name. The imperative verb has already been removed.

- **Temporal changes in toponyms**. Geographic concepts undergo temporal changes. For example, the effect of wars (the geographic area of *"Poland"* during the last centuries) or treaties (e.g. *"the EU"* refers to a different region after the expansion of the European Union) change what a geographic name represents. This is an indication that temporal and spatial restrictions should not be discussed separately.

- **Metonymic usage**. Toponyms are used ambiguously. For example, *"Libya"* occurs in the news corpus as a reference to the *"Libyan government"* (as in *"Libya stated that . . . "*). Similarly, *"British soil"* is used as a reference to *"Great Britain"*.

Currently, there is no practical solution for these problems and their investigation is a long-term issue for GIR. We concentrate on providing a basic architecture for geographic information retrieval with semantic networks, which will be refined later.

2 Interpreting Geographical Queries with Semantic Networks

The IICS group employs a syntactico-semantic parser (WOCADI parser – WOrd ClAss based DIsambiguating parser, Hartrumpf (2003)) to obtain the representation of queries and documents as semantic networks according to the MultiNet paradigm (Helbig, 2005). This approach has been used in experiments for domain-specific IR (Leveling and Hartrumpf, 2005) as well as in question answering (Hartrumpf, 2005). Aside from broadening the application domain of the MultiNet paradigm, its corresponding tools, and its applications, our work for the participation in the GeoCLEF task serves the following purposes:

1. To identify possible improvements for the NER component in the WOCADI parser.

2. To improve the connectivity between semantic networks and large resources of structured information (e.g. databases).

3. To create a larger set of geographical background knowledge by semi-automatic and automatic knowledge extraction from geographic resources.

4. To investigate the role of semantic relations and their interpretation for GIR.

We will discuss these points in the following subsections before presenting our results.

2.1 Improving the NER

Currently, the NER for WOCADI is based on large lists of names including cities, countries, organizations, products, etc. These name lexica contain more than 230,000 proper names. This approach is suitable in a domain where proper names are known in advance or in a limited domain. In general, a method to dynamically identify proper nouns for a semantic analysis is needed. There is a machine learning module in preparation that creates a hypothesis for a word form while parsing a text with WOCADI.

2.2 Connecting Semantic Networks and Databases

Data from the GEOnet Names Server[2] (GNS[3], containing approximately 4.0 million entities with 5.5 million names world-wide and 169,407 entities for Germany) was processed to provide a data set for a gazetteer database. The GNS is a valuable resource for geographical information, but the use of this multilingual gazetteer data has proved problematic in our setup, so far.

- Some data may not be present at all (e.g. *"the Scottish Trossachs"*), so that a geographic interpretation fails for some concepts.

- Geographic names may be present in the native language, in English, or both. For some concepts, there are no native language forms (the GNS data has an American English background). For example, *"The North Sea"* is present as a toponym in the database, but *"Nordsee"* is not. Similarly, there is no entry for *"Schottland"* in the GNS data, but the English spelling *"Scotland"* occurs. This means that the GNS data cannot be obtained (and subsequently used) for a non-English monolingual task without an additional translation phase.

- Geographic objects may cover several areas or may be adjacent to several regions (such as rivers traversing different countries, large forests, seas, or mountains, e.g. *"Bodensee"/'Lake of Constance'*, *"Rhein"/'Rhine'*, *"Donau"/'Danube'*, or *"Alpen"/'Alps'*).

- Relations or modifiers generate name variants which are not covered by a gazetteer (*"Süddeutschland"/'South(ern) Germany'*, *'the southern part of Germany'*), because they are subject to interpretation or do not have corresponding coordinates.

- The data representation may be inconsistent. For example, some rivers (streams) are represented by a set of coordinates (e.g. *"Alter Rhein"*), some are represented by a single coordinate (e.g. *"Main"*) in the GNS data.

- The gazetteer does not provide sufficient information for a successful disambiguation from context (for example, temporal information is missing).

- The ontological basis of the GNS is incomplete. For example, church (CH), religious center (CTRR), monastery (MNSTY), mission (MSSN), temple (TMPL), and mosque (MSQE) are defined (among others) in the GNS data as classes of geographic entities that refer to sacral buildings. A cathedral (GeoCLEF topic GC012) is a sacral building as well but neither is there a corresponding geographic class defined nor is gazetteer data for cathedrals provided.

- The inflection of names is typically not covered in gazetteers. Many names have a special genitive form in German (and English), which the morphology component of the WOCADI parser can analyze. But there are more complicated cases, where parts of a complex name are inflected for grammatical case. For example, the river *"(die) Schwarze Elster"* has the genitive form *"(der) Schwarzen Elster"*.

---

[2] http://earth-info.nga.mil/gns/html/cntry_files.html
[3] The GEOnet Names Server contains data from the National Geospatial-Intelligence Agency and the U.S. Board on Geographic Names database

Because of these problems, we see the GNS data as a general source of information, which should be extended by domain-, language-, and application-specific knowledge.

Gazetteers and derived knowledge bases share the same problems. Both are always incomplete (see Leidner (2004) and Fonseca et al. (2002) for a discussion of problems of gazetteer selection), data in both is not fine-grained or detailed enough for many tasks, and for both, entry points (valid search keys) for access must be known.

2.3 Expanding Geographical Background Knowledge

The IICS created and maintained a large semantic network as a geographical knowledge base for expanding geographical concepts. This knowledge base was automatically extended by generating hypotheses for new geographical concepts and integrating them into the semantic network.

For all concepts from the existing semantic network ontology, hypotheses are generated for meronymy relations. A hypothetical concept is created by concatenating some prefix with regular semantics in geography with the original concept. Typical examples of an implied meronymy are *"Süd-"*/*'South'*, *"Südost-"*/*'Southeast'*, *"Zentral-"*/*'Central'*, and *"Mittel-"*/*'Middle'*. The occurrence frequency of a hypothesis is looked up in the index of base forms for the entire (annotated/tagged) news corpus. Hypothetical concepts with a frequency less than a given threshold (three occurrences) were rejected. The resulting relations were integrated into the semantic network containing the background knowledge. These concepts typically do not occur in gazetteers because they are vague and their interpretation depends on context.

The second approach to expanding the geographic background knowledge at the IICS involved automatically extracting concepts from a database consisting of the GNS data for Germany. The GNS data shares much information with other major resources and services involving geographical information, such as the Getty Thesaurus of Geographical names[4] (TGN, containing about 1.3 million names) or the Alexandria Digital Library project[5] (ADL Gazetteer, containing about 4.4 million entries). Therefore we concentrated on the GNS data.

For each GNS gazetteer entry, a set of geographic codes is provided which can be interpreted to form a geographic path to the geographical object. For example, the database entry for the city of *"Wien"*/*'Vienna'* contains information that the city is located in *"Amerika oder Westeuropa"*/*'America or Western Europe'*, *"Europa"*/*'Europe'*, *"Österreich"*/*'Austria'*, and in the *"Bundesland Wien"* and that *"Vienna"* is a name variant of *"Wien"*. This information is post-processed and transformed into a set of semantic relations. We extracted some 20,000 geographical relations for a subset of 27 geographic classes (out of 648 classes defined in the GNS) from the German data. Relations that can be inferred from transitivity or symmetry properties are not explicitly entered into the geographical knowledge base as they are dynamically generated in our experiments.

2.4 The Role of Semantic Relations in Geographical Queries

The MultiNet paradigm offers a rich repertoire of semantic relations and functions. Table 2 shows and briefly describes the most important relations for representing topological, directional, or proximity information. Note that the length of a path of semantic relations between two related concepts may be used to calculate their (thematic or geographic) proximity or distance as well. Figure 2 shows an excerpt from our geographical knowledge base with MultiNet relations. For the moment, the interpretation of the semantic functions is limited because we do not yet use the assigned coordinates from the GNS data.

3 Monolingual GeoCLEF Experiments (German – German)

Currently, the GeoCLEF experiments follow our established setup for information retrieval tasks. The WOCADI parser is employed to analyze the newspaper and newswire articles, and concepts (or rather:
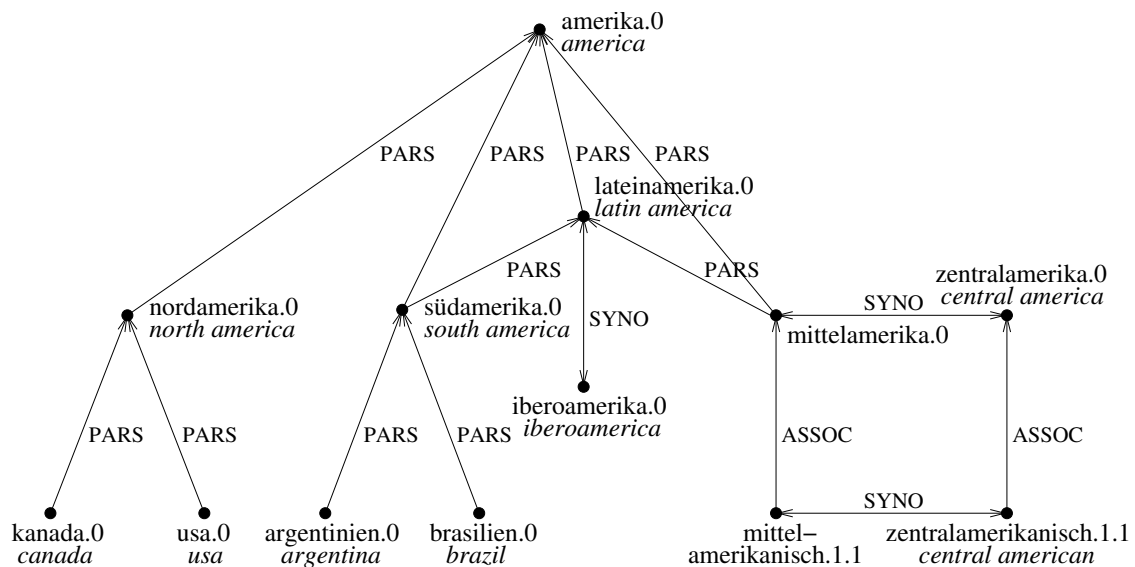
---

[4]http://www.getty.edu/
[5]http://www.alexandria.ucsb.edu/

amerika.0
america

PARS PARS PARS PARS

lateinamerika.0
latin america

nordamerika.0
north america

PARS PARS

südamerika.0
south america

PARS

SYNO

zentralamerika.0
central america

SYNO

mittelamerika.0

PARS PARS

kanada.0
canada

usa.0
usa

PARS PARS

argentinien.0
argentina

brasilien.0
brazil

SYNO

iberoamerika.0
iberoamerica

ASSOC ASSOC

mittel–
amerikanisch.1.1

zentralamerikanisch.1.1
central american

SYNO

Figure 2: Excerpt from the MultiNet representation of a geographical knowledge base.

Table 2: Overview of some important MultiNet relations and functions for the interpretation of geographic queries.

MultiNet Relation	Description
ASSOC(x,y)	concepts associated with toponyms and properties corresponding to toponyms, (e.g. language, inhabitant, or adjective form)
ATTCH(x,y)	attachment between objects; y is attached to x
ATTR(x,y)	attribute of an object; y is an attribute of x
CIRC(x,y)	situational circumstance (semantically non-restrictive)
CTXT(x,y)	contextual restriction
DIRCL(x,y), ORNT(x,y)	direction and orientation of events and situations (e.g. *"the flight to Berlin"*)
EQU(x,y)	name variants and equivalent names, including endonyms, exonyms, and historical names
LOC(x,y)	specifying locations; x takes place at y; x is located at y
PARS(x,y)	meronymy, holonymy (PART-OF); x is part of y
PRED(x,y)	predication; every z from set x is a y
SUB(x,y)	subordination (IS-A); x is a y
SYNO(x,y)	synonyms and near-synonyms
TEMP(x,y)	temporal specification
VAL(x,y)	value specification; y is value for attribute x
*IN(x,y)	semantic function; x is contained in y
*NEAR(x,y)	semantic function; x is close to y
*OUTSIDE(x,y)	semantic function; x is not inside y
*SOUTH_OF(x,y)	semantic function; x is south of y

Table 3: Overview of parameter settings and results for monolingual GeoCLEF experiments with the German document collection. The results displayed are the mean average precision (MAP) and the number of relevant and retrieved documents (rel_ret) for a total of 785 documents assessed as relevant.

Run Identifier	Parameters				Results	
	title	description	location elements	query expansion	MAP	rel_ret
FUHo10td	yes	yes	no	no	0.0779	479
FUHo10tdl	yes	yes	yes	no	0.0809	464
FUHo14td	yes	yes	no	yes	0.1053	519
FUHo14tdl	yes	yes	yes	yes	0.1022	530
FUHinstd	yes	yes	no	yes	0.0182	92

lemmata) and compound constituents are extracted from the parsing results as index terms. The representations of 276,579 documents (after duplicate elimination) are indexed with the Zebra database management system[6] (Hammer et al., 1995–2005), which supports a standard relevance ranking (term weighting by *tf-idf*).

Queries (topics) are analyzed with WOCADI to obtain the semantic network representation. The semantic networks are transformed into a Database Independent Query Representation (DIQR) expression. For some experiments (FUHo10tdl and FUHo14tdl), the location elements consisting of the concept, spatial-relation, and place names of a topic are transformed into a corresponding DIQR expression as well. Additional concepts (including toponyms) are added to the query formulation by including semantically related concepts. This approach is described in more detail in Leveling (2004). The fifth experiment employs a modified question answering approach for GIR and is described in Section 4. The five experiments are characterized in the parameter columns of Table 3. It also shows the performance of our experiments with respect to mean average precision (MAP) and number of relevant and retrieved documents.

The experiments with a query expansion based on additional geographic knowledge show a higher performance than the traditional IR approach wrt. MAP (FUHo10td vs. FUHo14td and FUHo10tdl vs. FUHo14tdl). The performance of the experiment employing traditional IR and the experiment with query expansion may be increased by changing to a database supporting a OKAPI/BM25 search for the thematic parts of a query.

4   GIR with Deep Sentence Parses

In addition to the runs described in Section 3, we experimented with an approach based on deep semantic analysis of documents and queries. We tried to turn the InSicht system normally used for question answering (Hartrumpf, 2005) into a GIR system (here abbreviated as GIR-InSicht). To this end, the following modifications were tried:

1. generalizing the central matching algorithm, which is sentence-based,

2. adding geographical background knowledge, and

3. adjusting parameters for network variation scores and limits for generating query network variants.

These three areas are explained in more detail in the following paragraphs.

The base system, InSicht, matches semantic networks derived from a query parse (topic title or topic description[7]) to document sentence networks one by one (whereas sentence boundaries are ignored in traditional IR). In GIR (as in IR), this approach yields high precision, but low recall because often the information contained in a query is distributed across several sentences in a document. To adjust the

---

[6]http://www.indexdata.dk/zebra

[7]GIR-InSicht combines the results for a query from the title field and for a query from the description field. All other topic fields are ignored. The information from the attributes *DE-concept*, *DE-spatialrelation*, and *DE-location* were equally well derived from the parse result of the title or description attribute.

matching approach to such situations, the query network is split if certain graph topologies are encountered. The resulting query network parts are viewed as conjunctively connected. The query network can be split at the following semantic relations: CIRC, CTXT, LOC, TEMP. For example, the LOC edge in Figure 1 can be deleted leading to two separate semantic networks. One corresponds to *"Bericht von Amnesty International über Menschenrechte"* (*'Amnesty International reports on human rights'*) and the other to *"Lateinamerika"* (*'Latin America'*). The greatest positive impact for GeoCLEF comes from splitting at LOC edges.

The geographical knowledge base described in Section 3 is scanned by GIR-InSicht; relations that contain names that do not occur in the document parse results (i.e. the semantic networks of document sentences) are ignored. For simplicity, meronymy edges (PARS) are treated like hyponymy edges so that GIR-InSicht can use all part-whole relations for concept variations (see Hartrumpf (2005)) in query networks. Without the geographical knowledge base, recall is much lower.

Some InSicht parameters had to be adjusted in order to yield more results in GIR-InSicht and/or to keep answer time and RAM consumption acceptable even when working with large background knowledge bases like the one mentioned in Section 2.3. The final steps of InSicht that come after a semantic network match has been found (answer generation and answer selection) can be skipped. Some minor adjustments further reduce run time without losing relevant documents. For example, an apposition for a named entity (like *"Hauptstadt"* (*'capital'*) for *"Sarajevo"*/*"Sarajewo"*) leads to a new query network variant only if this combination occurs at least 2 times in the document collection.

We evaluated about ten different setups of GIR-InSicht on the GeoCLEF 2005 topics. The setups used different extension combinations from the extensions described above and other extensions, like coreference resolution for documents. The performance differences were often marginal. In some cases, this indicates that a specific extension is irrelevant for GIR; in other cases, the number of topics (23 topics have relevant documents) and the number of relevant documents might be too small to draw any conclusions. However, one can see considerable performance improvements for some extensions, e.g. splitting query networks at LOC edges. Larger evaluations are needed to gain more insights about which development directions are most promising for the semantic network matching approach to GIR.

5   Conclusion and Outlook

The semantic network representation with MultiNet offers representational means useful for GIR. We successfully employed semantic networks to uniformly represent queries, documents, and geographical background knowledge and to connect to external resources like GNS data. Three different approaches have been investigated: a baseline corresponding to a traditional IR approach; a variant expanding thematic, temporal, and geographic descriptors from the MultiNet representation of the query; and an adaptation of InSicht, a QA algorithm based on semantic networks. The diversity of our approaches looks promising for a combined system.

Future work also includes completing the topics described in Section 2, namely improving the NER, connecting semantic networks and databases, expanding geographical background knowledge, and investigating the role of semantic relations in geographical queries. It remains to be investigated whether methods that are successful in traditional IR are equally successful for treating polysemy and synonymy for toponyms in GIR.

References

Fonseca, Frederico T.; Max J. Egenhofer; Peggy Agouris; and Gilberto Câmara (2002). Using ontologies for integrated geographic information systems. *Transactions in GIS*, 6(3).

Hammer, Sebastian; Adam Dickmeiss; Heikki Levanto; and Mike Taylor (1995–2005). *Zebra – User's Guide and Reference*. Copenhagen.

Hartrumpf, Sven (2003). *Hybrid Disambiguation in Natural Language Analysis*. Osnabrück, Germany: Der Andere Verlag.

Hartrumpf, Sven (2005). Question answering using sentence parsing and semantic network matching. In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign* (edited by Peters, C.; P. D. Clough; G. J. F. Jones; J. Gonzalo; M. Kluck; and B. Magnini), volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pp. 512–521. Berlin: Springer.

Helbig, Hermann (2005). *Knowledge Representation and the Semantics of Natural Language*. Berlin: Springer.

Jones, Christopher B.; R. Purves; Anne Ruas; Mark Sanderson; Monika Sester; Marc J. van Kreveld; and Robert Weibel (2002). Spatial information retrieval and geographical ontologies – an overview of the SPIRIT project. In *SIGIR 2002*, pp. 387–388.

Kunze, Claudia and Andreas Wagner (2001). Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche. In *Chancen und Perspektiven computergestützter Lexikographie* (edited by Lemberg, Ingrid; Bernhard Schröder; and Angelika Storrer), volume 107 of *Lexicographica Series Maior*, pp. 229–246. Tübingen, Germany: Niemeyer.

Leidner, Jochen L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. In *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR 2004)*. Sheffield, UK.

Leveling, Johannes (2004). University of Hagen at CLEF 2003: Natural language access to the GIRT4 data. In *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Revised Selected Papers* (edited by Peters, Carol; Julio Gonzalo; Martin Braschler; and Michael Kluck), volume 3237 of *Lecture Notes in Computer Science (LNCS)*, pp. 412–424. Berlin: Springer.

Leveling, Johannes and Sven Hartrumpf (2005). University of Hagen at CLEF 2004: Indexing and translating concepts for the GIRT task. In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign* (edited by Peters, C.; P. D. Clough; G. J. F. Jones; J. Gonzalo; M. Kluck; and B. Magnini), volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pp. 271–282. Berlin: Springer.