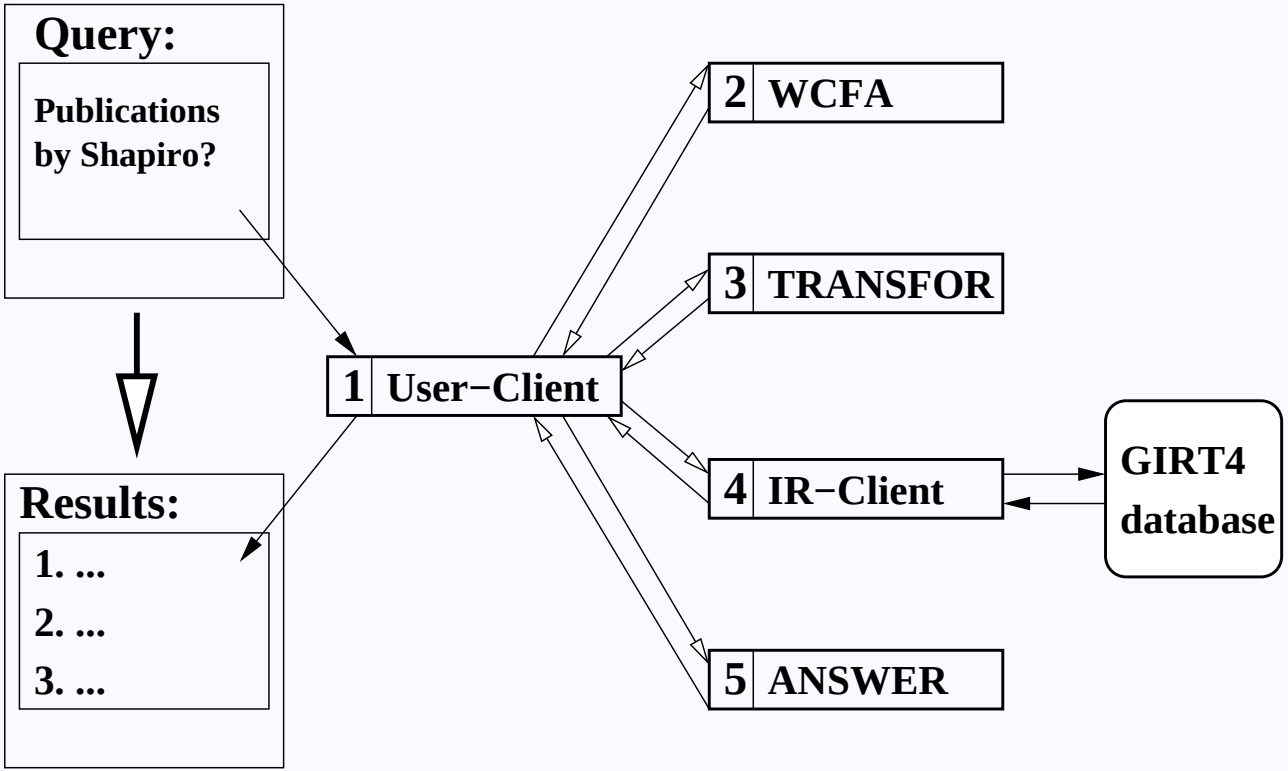
University of Hagen at CLEF 2003: Natural language access to the GIRT4 data

Johannes Leveling
Applied Computer Science VII
Intelligent Information and Communication Systems
FernUniversität Hagen
58084 Hagen, GERMANY
johannes.leveling@fernuni-hagen.de

Motivation and Problem Statement

- **Background:**
  Natural language interface for databases (NLI-DB) accepting unrectricted natural language input (NLI-Z39.50)
- **Major problems:**
  1. Query transformation
  2. Vocabulary mismatches leading to search failures
- **Proposed solution:**
  Automated retrieval strategies using query variants

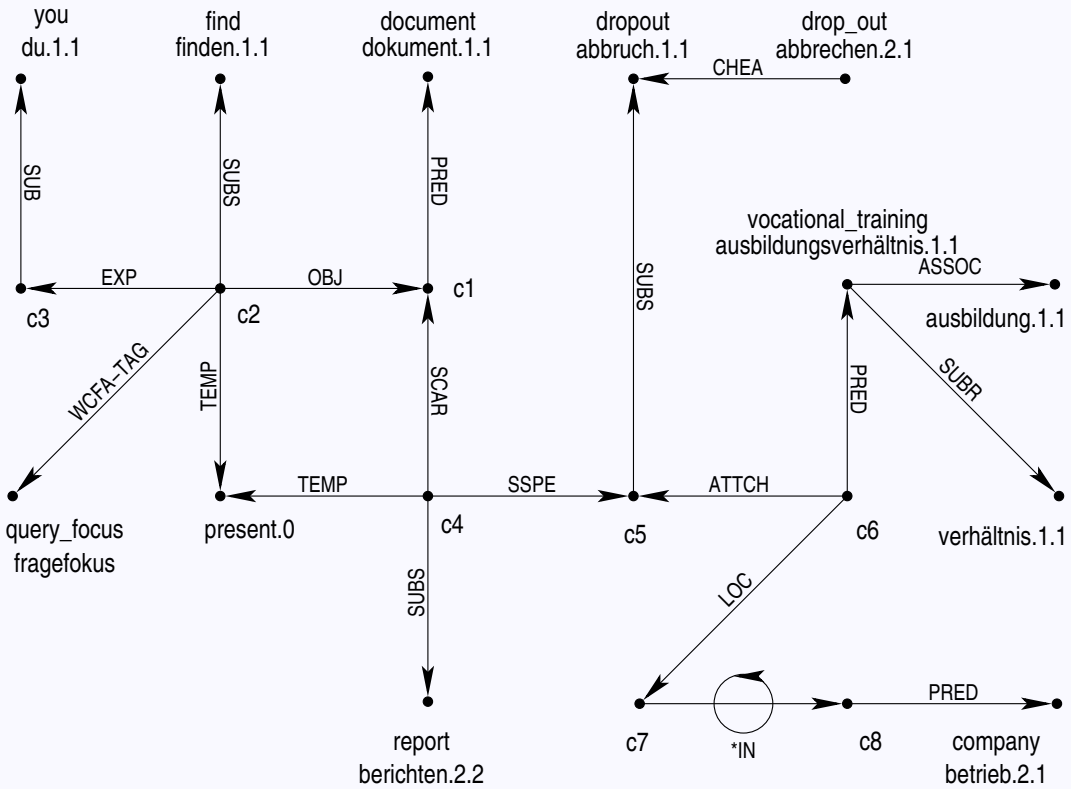
The NLI-Z39.50 Architecture for GIRT4 Experiments



Some Relations defined in the MultiNet Paradigm

MultiNet relation	Short description
ASSOC	relation of association
ATTCH	attachment of object to object
CHEA	change of event to abstractum
LOC	relation specifying the location for a situation
PRED	predicative concept specifying a plurality
SCAR	carrier of a state
SSPE	entity specifying a state
SUB	conceptual subordination for objects
SUBR	metarelation for the description of relations
SUBS	conceptual subordination for situations
TEMP	temporal restriction for a situation
*IN	a location-producing function

MultiNet Representation for GIRT Topic 081



"Finde Dokumente, die über den Abbruch von Ausbildungsverhältnissen in Betrieben berichten."

The Z39.50 GIRT4 Database

- Zebra database software (a free Z39.50 toolkit)
- Default indexing strategy: full word forms
- Default ranking algorithm: *tf-idf* ranking
- Boolean operators / relevance operator employed in IR-Client
- Technical constraints:
  - limited query length
  - restricted number of disjunctions in a query

DIQR for the query *“Kennst du Bücher von Dörner über Komplexitätsmanagement?”*

```
(AND (author = (name ‘dörner.0’))
     (title = (OR  (word ‘komplexitätsmanagement.1.1’)
                   (AND (word ‘komplexität.1.1’)
                        (word ‘management.1.1’)))))
```

Elimination of disjunctions:

```
(AND (author = (name ‘dörner.0’))
     (title  = (word ‘komplexitätsmanagement.1.1’)))

(AND (author = (name ‘dörner.0’))
     (title  = (AND (word ‘komplexität.1.1’)
                    (word ‘management.1.1’))))
```

Search Term Variants

- **Orthographic variants**
  Examples: *“Schiffahrt”* and *“Schifffahrt”*; *“Bänke”* and *“Baenke”*
- **Morphologic variants**
  - Inflectional morphology
    Examples: *“Stadt”* / *“city”* and *“Städte”* / “cities”
  - Derivational morphology
    Examples: *“Abbruch”* / *“dropout”* and *“abbrechen”* / *“drop_out”*
- **Lexical variants**
  Example: (*“ansehen.2.3”* SYNO *“betrachten.1.2”*)
- **Syntactic variation**
  Example: *“Ausbildungsabbrecher”* and *“Ausbildung”*, *“Abbrecher”*

Semantic Similarity (Term Variant Scoring)

$$sim(x, y) = \begin{cases} 1 & : \text{if } (x \text{ EQU } y) \text{ or } (y \text{ EQU } x) \text{ exists (equal variants)} \\ 0.9 & : \text{if } (x \text{ SYNO } y) \text{ or } (y \text{ SYNO } x) \text{ exists (synonymy)} \\ 0.7 & : \text{if } (x \text{ SUB } y) \text{ exists (hyponymy)} \\ 0.5 & : \text{if } (y \text{ SUB } x) \text{ exists (hypernymy)} \\ 0.6 & : \text{if } (x \text{ PARS } y) \text{ exists (meronymy)} \\ 0.4 & : \text{if } (y \text{ PARS } x) \text{ exists (holonymy)} \\ 0.3 & : \text{if } (x \text{ ASSOC } y) \text{ or } (y \text{ ASSOC } x) \text{ exists} \\ \ldots & : \ldots \end{cases}$$

The Automated Retrieval Strategies

1) Obtain sets of concepts linguistically related to original query concepts and compute semantic similarities.
2) Generate a set of query variants and rank them by query score.

3a) *Single database query:*
Collect search terms from top ranked query variants
and retrieve documents up to fixed result set size.

3b) *Multiple database queries:*
Perform searches with top ranked query variants
and insert documents into the result set.

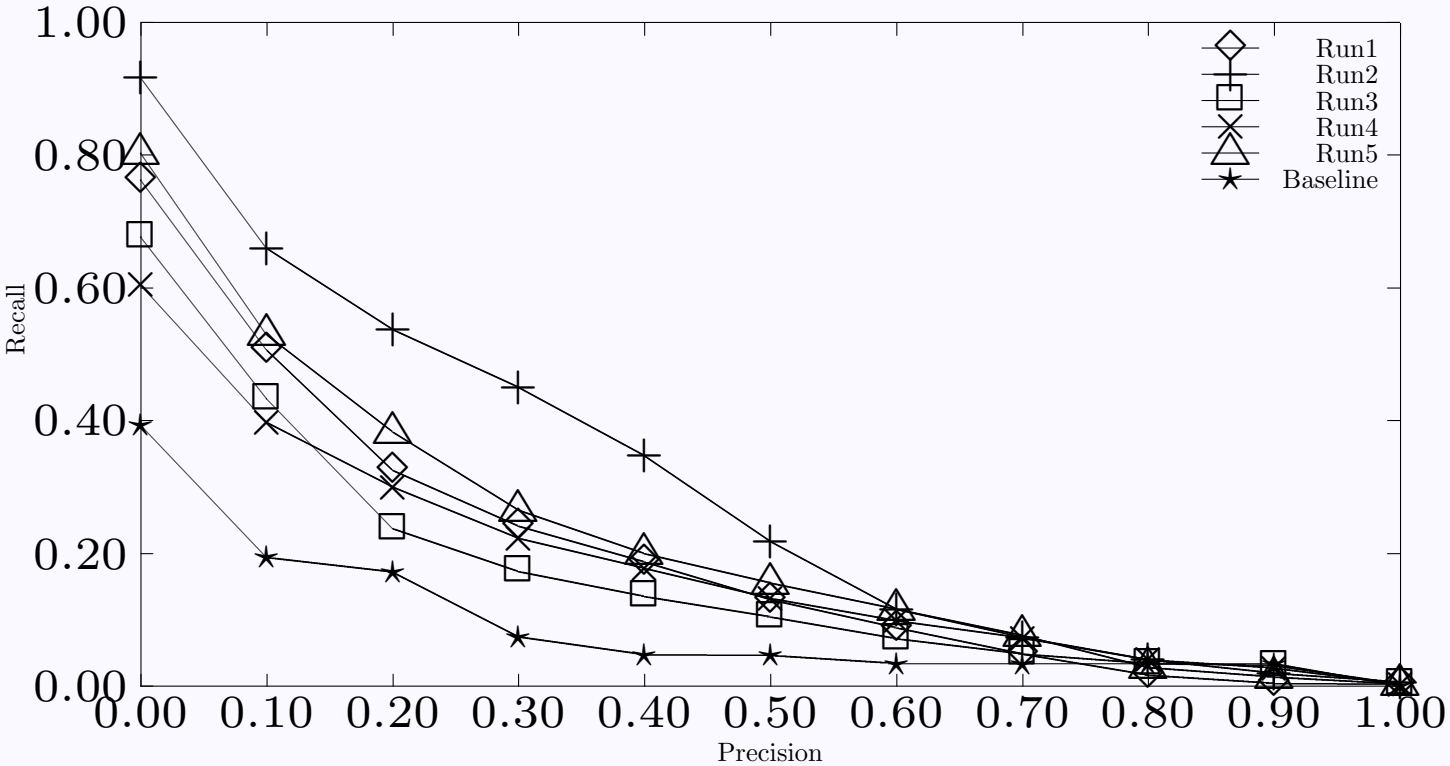
4) Rank result set by document scores.

Parameters for Retrieval Experiments (GIRT German-German)

Run ID	topic fields	background knowledge used?	query type	ranking
Run1:	TD	N	M	QD
Run2:	T	Y	M	QD
Run3:	D	Y	M	QD
Run4:	TD	Y	S	QD
Run5:	TD	Y	M	QD
Baseline:	TD	Y	M (Boolean)	Q

- topic fields: title (T), description (D) or the combination of both (TD)
- lexicon and background knowledge used: yes (Y) or no (N)
- query type: multiple queries (M) or a single query (S)
- ranking: query and database score (QD) or query score alone (Q)

Results for GIRT4 (German-German) Experiments



Example (*good*): Topic 77

T: *“Politische Partizipation von DDR-Frauen”*

D: *“Finde Dokumente, die über die Teilhabe von Frauen in der DDR am politischen Leben des Landes berichten.”*

- derivational information for *“Partizipation”* available
- expansion of abbreviation *“DDR”*
- unlexicalized term *“Teilhabe”* (no variants)
- *unspecific* term *“Leben”* occurs in phrase
- *“Land”* refers to *“DDR”*

Example (*not so good*): Topic 78

T: *“Bildung in der Türkei”*

D: *“Finde Informationen über das Bildungssystem in der Türkei.”*

- decomposition of *“Bildungssystem”*
- disambiguation of terms *“Bildung”* and *“System”*
- only weak semantic connection between *“Türkei”* and *“türkisch”*

Observations

- Boolean queries (Baseline) do not perform well for an IR task, even if background knowledge is used.
- Better performance with **shorter queries** (Run2) than with longer queries (Run3).
- Better performance with **multiple queries** (Run5) than with a single query (Run4).
- **Small improvement** in retrieval performance using **additional lexical information and background knowledge** (Run5 vs. Run1).

Future Work

- Replace, add and omit search terms for a query variant
- Multilingual IR
  (for example: German queries, English documents)
- Semantic network representation for both queries **and** documents