University of Hagen at CLEF 2004: Indexing and Translating Concepts for the GIRT Task

Johannes Leveling, Sven Hartrumpf
Intelligent Information and Communication Systems
Computer Science Department
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany
http://pi7.fernuni-hagen.de/

CLEF 2004 Workshop, Bath, UK

Objectives for our experiments for the GIRT task

Monolingual experiments (DE–DE):

- Compare approaches for indexing and matching full word forms, concepts, and semantic networks

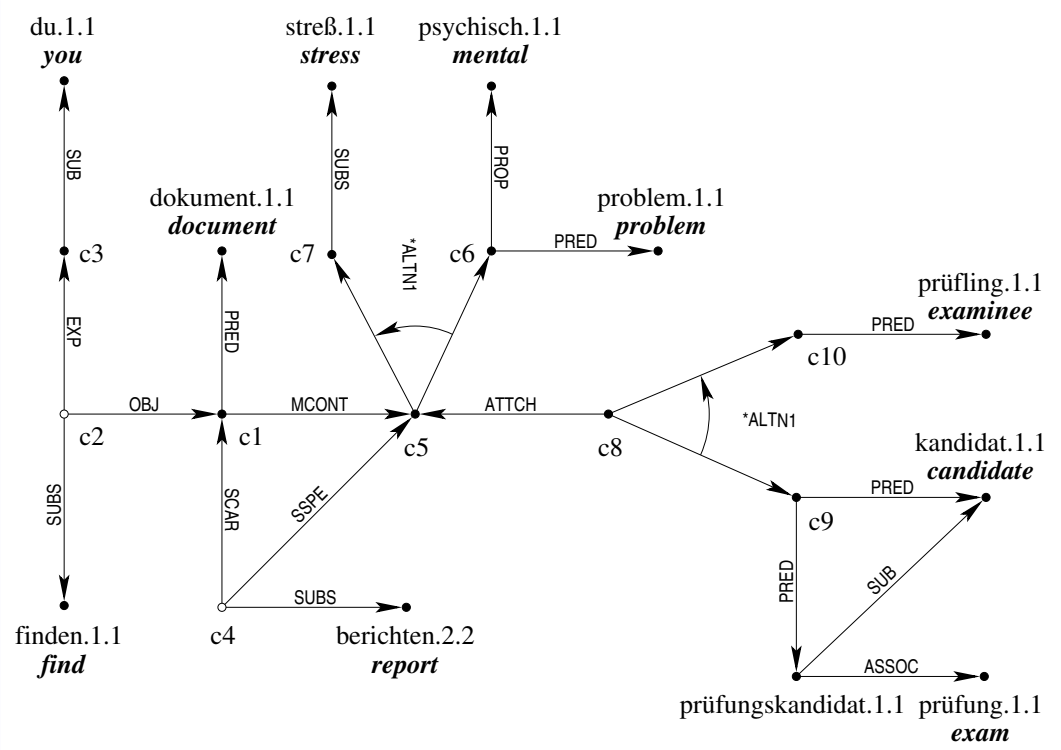
Bilingual experiments (DE–EN):

- Find parameters for an approach to automatically translate and expand a query

Basics and background

- Multi-Layered Extended Semantic Networks (MultiNet)
  [Hel01, HG02]
- Word Class Functional Analysis (WCFA) /
  WOrd ClAss based DIsambiguating parser (WOCADI)
  [Har03]
- Hagen German Lexicon (HaGenLex)
  [HHO03]
- Natural Language Interface for the Z39.50 protocol (NLI-Z39.50)
  [LH02, Lev03]

Core MultiNet representation for GIRT topic 116



“Finde Dokumente, die über psychische Probleme oder Stress von Prüfungskandidaten oder Prüflingen berichten.”

‘Find documents reporting on mental problems or stress of examination candidates or examinees.’

Database Independent Query Representation (DIQR) for GIRT topic 116

((OR *title abstract*) = (AND (OR (phrase *“psychologisch .1.1” “problem.1.1”*)
(word *“stress.1.1”*))
(OR (word *“prüfungskandidat.1.1 ”*)
(wordlist *“prüfung.1.1” “ kandidat.1.1”*)
(word *“prüfling.1.1”*))))

- search terms are expanded with semantically related terms
- normalization into disjunctive normal form (DNF)
- DNF components, written as conjunctions, are interpreted as query variants

Semantic similarity between two concepts x and y

$$sim(x,y) = \begin{cases} 1.0 & \text{if } x \text{ and } y \text{ are identical:} \\ & x\ rel\ y \text{ and } rel \in \{EQU\} \\ 0.95 & \text{if } x \text{ is a synonym of } y: \\ & x\ rel\ y \text{ and } rel \in \{SYNO\} \\ 0.7 & \text{if } x \text{ is a narrower term than } y: \\ & x\ rel\ y \text{ and } rel \in \{SUB, SUBS, PARS, ...\} \\ 0.6 & \text{if } x \text{ and } y \text{ are morphologically derived:} \\ & x\ rel\ y \text{ and } rel \in \{CHPA, CHPE, ...\} \\ 0.5 & \text{if } x \text{ is a broader term than } y: \\ & y\ rel\ x \text{ and } rel \in \{SUB, SUBS, PARS, ...\} \\ 0.35 & \text{if } y \text{ is a term otherwise related to } x: \\ & x\ rel\ y \text{ and } rel \in \{ASSOC, ...\} \end{cases}$$

Retrieval strategy

1. **Monolingual:** expand query terms with disjunction of search term variants (optionally weighted by semantic similarity)

   Bilingual: expand query terms with disjunction of semantically related translations the original query.
2. Normalize query into DNF and interpret components as query variants

   Rank query variants by their score (semantic similarity between a query variant and the original query)
3. **Single query:** collect all search terms in the top ranked 250 query variants

   Multiple queries: use the 250 top ranked query variants for retrieval

   Documents are retrieved until the result set exceeds a fixed size (here: 1000 documents)
4. Score documents by a weighted sum of the database score (d_{score}) (a standard *tf-idf* score as determined by the database ranking schema) and a query score (q_{score})

Parameters for monolingual GIRT experiments

- **Q-S**: create a single query from all query variants
  Q-M: process query variants separately
- **I-W**: search terms and index terms are words
  I-C: search terms are HaGenLex concepts
  I-N: search terms are concepts and relations from semantic networks
- **R-E**: perform an exact search for search terms
  R-T: use a so-called right truncation or prefix match
- a document score (doc_{score}) is a weighted sum of database document score and query score:

  D-1: $doc_{score} = 0.7 \cdot d_{score} + 0.3 \cdot q_{score}$

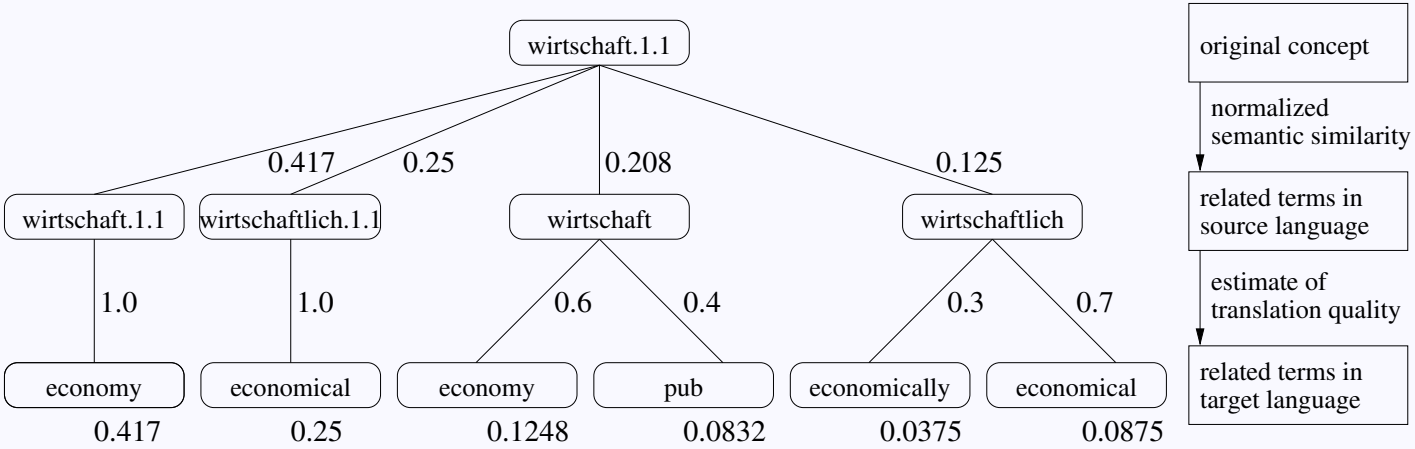
  D-2: $doc_{score} = 0.7 \cdot d_{score} + 0.3 \cdot \frac{q_{score}}{query\_length}$

  D-3: $doc_{score} = q_{score}$

Parameter settings and results for monolingual GIRT experiments

Run Identifier	Database Name	Parameters				MAP
FUHds1	GIRT4DE	Q-S	I-W	R-T	D-1	0.2446
FUHdw1	GIRT4DE	Q-M	I-W	R-T	D-1	0.2482
FUHdw2	GIRT4DE	Q-M	I-W	R-T	D-2	0.2276
FUHdrw	GIRT4RDG	Q-M	I-C	R-E	D-1	0.1162
FUHdm	InSicht	Q-M	I-N	R-E	D-3	0.1126

Concept translation



The tree representation for translating the concept *wirtschaft.1.1*. A translation score for a concept is computed as the sum of products of edge markings on the path to the leaf nodes containing this concept. The resulting order of concept translations is

- *'economy'* (score: $0.417 \cdot 1.0 + 0.208 \cdot 0.6 = 0.5418$),
- *'economical'* (0.3375),
- *'pub'* (0.0832), and
- *'economically'* (0.0375).

Parameters for bilingual experiments

- **Q-S**: create a single query
  Q-M: process multiple query variants
- **G-all**: all semantically related terms are used as term variants
  G-5: the best five semantically related terms are used
- **E-all**: all translations found are used as term translations in a query
  E-5: the best five translations are used
- **W-T**: translation scores are used to weight query search terms
  W-U: query terms are not weighted

Parameter settings and results for re-runs of bilingual GIRT experiments

Run Identifier	Database Name	Parameters				MAP
FUHe1 re-run	GIRT4EN	G-5	E-5	W-U	Q-M	0.1117
FUHe2 re-run	GIRT4EN	G-all	E-5	W-T	Q-M	0.1135
FUHe3 re-run	GIRT4EN	G-5	E-all	W-T	Q-M	0.1288
FUHe4 re-run	GIRT4EN	G-5	E-5	W-T	Q-M	0.1275
FUHe5 re-run	GIRT4EN	G-5	E-5	W-T	Q-S	0.1104

Results

Monolingual experiments:

- methods perform best in the order of
  - word form indexing (FUHds1, FUHdw1, FUHdw2),
  - concept indexing (FUHdrw), and
  - semantic network matching (FUHdm)

  with respect to mean average precision (MAP)
- performance of indexing and matching concepts and semantic networks may be specific to GIRT corpus:
  only 34.3% of the 1,111,121 sentences parsed successfully;
  parses for newspaper articles are significantly better

Bilingual experiments:

- not much difference for experiments with different parameter settings

Perspectives for improvements

For monolingual experiments:

- improve coverage of the WCFA
- match across several sentences / semantic networks (with intersentential resolution)
- match partial semantic networks

For bilingual experiments:

- complete the mapping of HaGenLex concepts to readings of EuroWordNet
- translate multi-word expressions and compounds

References

[Har03] Sven Hartrumpf. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany, 2003. ISBN 3-89959-080-5.

[Hel01] Hermann Helbig. *Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet*. Springer, Berlin, 2001.

[HG02] Hermann Helbig and Carsten Gnörlich. Multilayered extended semantic networks as a language for meaning representation in NLP systems. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, volume 2276 of *LNCS*, pages 69–85, Berlin, 2002. Springer.

[HHO03] Sven Hartrumpf, Hermann Helbig, and Rainer Osswald. The semantically based computer lexicon HaGenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105, 2003.

[Lev03] Johannes Leveling. University of Hagen at CLEF 2003: Natural language access to the GIRT4 data. In Carol Peters, editor, *Results of the CLEF 2003 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2003 Workshop*, pages 253–262, Trondheim, Norway, August 2003.

[LH02] Johannes Leveling and Hermann Helbig. A robust natural language interface for access to bibliographic databases. In Nagib Callaos, Maurice Margenstern, and Belkis Sanchez, editors, *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002)*, volume XI, pages 133–138, Orlando, Florida, July 2002. International Institute of Informatics and Systemics (IIIS).