MUCHMORE

Monolingual and cross-language information retrieval

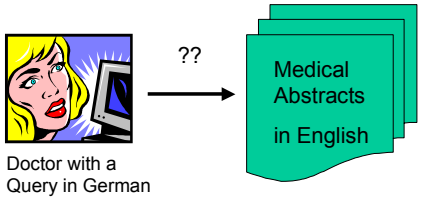
Martin Volk
Bärbel Ripplinger

Project Partners

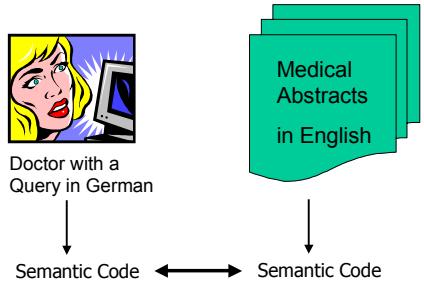
- European Partners
- Xerox Grenoble
- DFKI Saarbrücken
- ZInfo Frankfurt
- Eurospider Zürich

- American Partners
- Carnegie Mellon University
- CSLI at Stanford University

The idea in CLIR



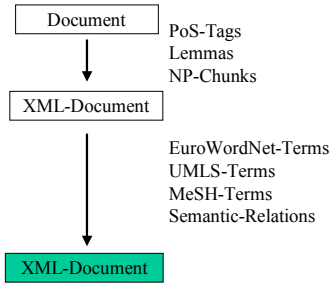
The idea in CLIR



Example of a Medical Abstract

- HIV remains infectious at room temperature for about six to twelve hours under sterile conditions, the time is shorter in the presence of bacteria and when clotting occurs. HIV infectivity is easily destroyed by disinfectants. The human infectious dose of HIV is around 100 to 1000 particles, dependent on the virus and the site of entry into the host.

The offline annotation process



Example of a Medical Abstract

```
<token id="w1" pos="NN">HIV</token>
<token id="w2" pos="NN"
   lemma="infectivity">infectivity</
<token id="w3" pos="VBZ" lemma="be">is</token>
<token id="w4" pos="RB"
   lemma="easy">easily</token>
<token id="w5" pos="VBN"
   lemma="destroy">destroyed</
<token id="w6" pos="IN" lemma="by">by</token>
<token id="w7" pos="NNS" lemma="disinfectant">
   disinfectants
<token id="w8" pos="PUNCT">.</token>
```


Example of a Medical Abstract

```
<token id="w1" pos="NN">HIV</token>
<token id="w2" pos="NN" lemma="infectivity">infectivity</
<token id="w3" pos="VBZ" lemma="be">is</token>
<token id="w4" pos="RB" lemma="easy">easily</token>
<token id="w5" pos="VBN" lemma="destroy">destroyed</
<token id="w6" pos="IN" lemma="by">by</token>
<token id="w7" pos="NNS" lemma="disinfectant">
disinfectants
<token id="w8" pos="PUNCT">.</token>

<umlsterm id="t1" from="w1" to="w1">
      <concept id="t1.1" cui="C0019682" preferred="HIV"/>
</umlsterm>
<umlsterm id="t2" from="w2" to="w2">
      <concept id="t2.1" cui="C0030657"
preferred="pathogenicity"
```


Scope of the annotation

- parallel corpus: in both DE and EN
- assymmetrical: annotation based only on the particular language
- in documents and queries

Starting point

- 7809 medical abstracts in both DE and EN
- 25 queries with human relevance assessments
- Example query:
  - *Arthroskopische Behandlung bei Kreuzbandverletzungen*
  - *Arthroscopic treatment of cruciate ligament injuries*
- each query has between 7 and 104 relevant documents
- the total number of relevant documents is 959

How to obtain human relevance assessments

- **Pooling technique**: Use a number of 'sufficiently' different IR systems and run the queries on all of them. Use X number of documents retrieved by 'most' systems for manual inspection.
- Example: For a query Q IR system 1 delivers 750 documents, IR system 2 delivers 900 docs and IR system 3 delivers 1000 docs. And we do not want to inspect more than 500 docs per query.
- Let's assume 200 docs are in all 3 result lists and 400 are in 2 result lists. Then we might accept for manual inspection the 200 common docs and those 300 that have the highest precision values.

Indexing information

- **Token (with part-of-speech)**
  - ***DE: Kreuzbandes***
  - ***EN: ligaments***
- **Lemma (or sequence of lemmas)**
  - ***DE: Faserknorpel → Faser + Knorpel***
  - ***EN: ligament***
- **UMLS term and code**
  - ***Kniegelenk: C0022745_T030***
- **MeSH code**
  - ***A2.513***
- **EuroWordNet code**
  - ***2747860***
- **Semantic relation (over a pair of UMLS terms)**
  - ***Kniegelenk is_connected_to Hinteres Kreuzband***

Clinical Terminologies

- are organized in **synonym sets**.
- Example: *Myocardial infarction* is synonym to
  - *heart attack* and
  - *coronary thrombosis and*
  - *M.I.*

  →*all of them get the same code*

→are organized in a **hierarchy**

→Example: *Myocardial infarction* is a type of

→*Ischaemic heart disease* which is a type of

→*Disorder of the heart* which is a type of

→*Cardiovascular disorder*

Clinical Terminologies

- **Uses**:
- direct patient care
- statistical reporting
- automated decision support (= expert systems)
- clinical research

UMLS

- The Unified Medical Language System
- A project of the U.S. National Library of Medicine
- "The 2002AC edition of the Metathesaurus includes 870,853 concepts and 2.27 million concept names in its source vocabularies."

UMLS Concept Names (MRCON)

BAQ	695
DAN	723
DUT	36491
ENG	1753789
FIN	21086
FRE	36556
GER	67987
HEB	485
HUN	718
ITA	23602
NOR	722
POR	45711
RUS	42346
SPA	51469
SWE	723
Sum of all entries	2083103

Monolingual retrieval for German

	Relevant retrieved docs	Average precision	Precision at 0.1 recall
Token	322	0.1600	0.5622
Token & Lemma	516	0.2180	0.5967
Token & Lemma & UMLS term	509	0.2236	0.5895
Token & Lemma & EuroWordNet	500	0.1980	0.5571
Token & Lemma & Mesh	526	0.2462	0.6356
Token & Lemma & Sem-relation	516	0.2224	0.5841

Evaluation measures according to TREC

1. **Number of relevant documents retrieved** for a query Q

- Caution: For a collection of 7000 documents it is relatively easy to obtain a high number of relevant documents retrieved!!
- In our experiments the maximal number of documents to be retrieved per query is 1000. So, even by chance we could get 1/7 (14.3% of 956 is 136) of the relevant documents.
- → The precision figures are more important than the number of relevant documents retrieved.

Evaluation measures according to TREC

- **2. Average precision** is the average of the precision value obtained after each relevant document is retrieved.
- It rewards systems that retrieve relevant docs highly ranked.
- Example: Let's assume there are 4 relevant documents to a query Q. And they are found in the ranked list on positions 1, 2, 4 und 7.
- Then:

Position	Precision
1	1
2	1
4	0.75
7	0.57
Average	**0.83**

Evaluation measures according to TREC

- **3. Recall Level Precision Average at 0.1 Recall (P0.1)**
- How many documents do I have go through in order to find 10% of the relevant documents?
- Example: Let us assume there are 100 relevant documents to a query Q.

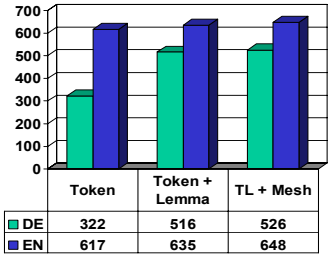
→How many documents do I have go through in order to find 10 relevant documents to Q?

→If P0.1 = 75% then #docs = 13.3

→If P0.1 = 50% then #docs = 20

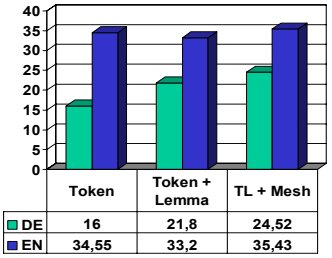
→If P0.1 = 25% then #docs = 40

Monolingual retrieval for German vs. English: # of Rel. Retr.



	Token	Token + Lemma	TL + Mesh
DE	322	516	526
EN	617	635	648

Monolingual retrieval for German vs. English: Precision



	Token	Token + Lemma	TL + Mesh
DE	16	21,8	24,52
EN	34,55	33,2	35,43

Heuristic morphology for German

•For all adjectives, nouns, verbs without lemma create lemmas by:

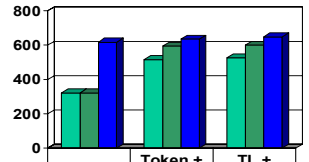
- remove inflectional suffix for all adjectives
- (e.g. *arthroskopischen* → *arthroskopisch*)
- separate all hyphenated compounds
- (e.g. *HWS-Distorsion* → *HWS + Distorsion*)
- separate all Ns, Adjs, Vs into two components if both components occur stand-alone in the corpus

(e.g. *Kardiomyopathie* → *Kardio + Myopathie*)

•Over all documents we generated:

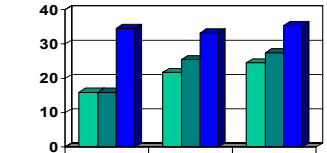
- 28'341 new adjective lemmas
- 22'260 new lemmas from hyphenated compounds
- 20'876 new lemmas from decompounding

Monolingual retrieval for German vs. English II: # Rel. Retr.



	Token	Token + Lemma	TL + Mesh
DE	322	516	526
DE +	322	594	600
EN	617	635	648

Monolingual retrieval for German vs. English II: Precision

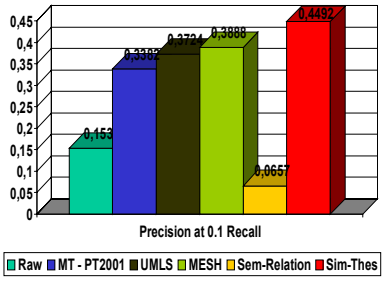


	Token	Token + Lemma	TL + Mesh
DE	16	21,8	24,52
DE +	16	25,47	27,46
EN	34,55	33,2	35,43

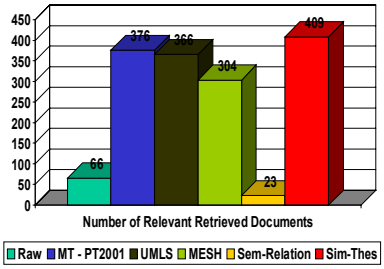
Cross Language Information Retrieval (CLIR)

- Task: Queries in DE and documents in EN (or vice versa)
- Methods:
  1. Use Tokens / Lemmas from source language queries and search in target language documents (rely on cross language overlap in the technical language)
  2. Machine Translation of queries with standard MT system (Personal Translator 2001; linguatec-Munich)
  3. Translation of queries with bilingual similarity thesaurus
  4. Use of semantic terms, codes and relations (UMLS, Mesh, EWN etc.)

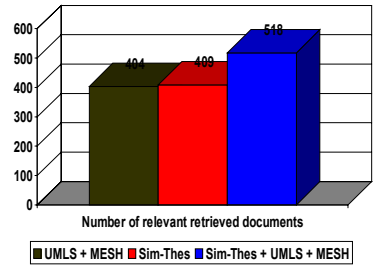
CLIR from German to English



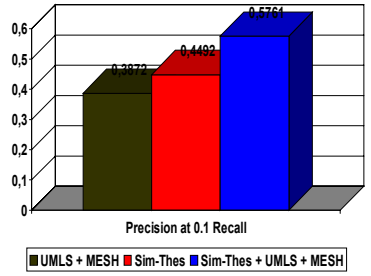
CLIR from German to English



CLIR from German to English (Combinations)



CLIR from German to English (Combinations)



Conclusions

- Lemmatization is the most important step in German monolingual retrieval.
- MESH is most reliable among semantic codes.
- In CLIR
  - MT is surprisingly good in terms of recall.
  - Combination of semantic codes (UMLS + MESH) outperforms MT.
  - Best results: Combination of semantic codes and similarity thesaurus.

Problems with EuroWordNet

- **German EWN codes are too general.**
- Examples from the Queries:
- EWN-Code: **Grenze**
- ← Query 9: *Patientengesteuerte Analgesie, Indikationen und Grenzen*

→EWN-Code: **Behandlung, Therapie**

← Query 91: *Behandlung von Plattenepithel-karzinomen*

→EWN-Code: **Ursache**

← Query 108: *Ursachen von Schluckstörungen*

Problems with EuroWordNet

- **English EWN codes are much more detailed.**
- Examples from the Queries:
- EWN-Codes: **analgesia; indication; limit**
- ← Query 9: *Patient-controlled analgesia, indications and limits*

→EWN-Codes: **treatment; cell; carcinoma**

← Query 91: *Treatment of squamous cell carcinoma*

→EWN-Code: **cause**

← Query 108: *Cause of dysphagia*

Problems with EuroWordNet

→English EWN codes are much more detailed than German EWN codes.

→This helps in English monolingual retrieval.

	mAvP	Recall	Prec at 0.1	Prec at 10 docs
DE	0.0025	86	0.0116	0.0120
EN	0.1178	462	0.3058	0.2440

BUT this does not help in CLIR since both the German queries and documents have only very general EWN codes.