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
Doctor with a Query in German

Medical Abstracts in English

The offline annotation process

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.
Example of a Medical Abstract

HIV infectivity is easily destroyed by disinfectants.

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: Kreuzbandverletzungen - EN: ligamens
- Lemma (or sequence of lemmas) - DE: Faserknorpel -> Faser + Knoepf - EN: ligament
- UMLS term and code - Kniegelenk: C0022745_T030
- MeSH code - A2.513
- EuroWordNet code - A2.513
- 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)

<table>
<thead>
<tr>
<th>Language</th>
<th>Total Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAQ</td>
<td>695</td>
</tr>
<tr>
<td>DAN</td>
<td>723</td>
</tr>
<tr>
<td>DUT</td>
<td>36491</td>
</tr>
<tr>
<td>ENG</td>
<td>1753789</td>
</tr>
<tr>
<td>FIN</td>
<td>21086</td>
</tr>
<tr>
<td>FRE</td>
<td>36556</td>
</tr>
<tr>
<td>GER</td>
<td>67987</td>
</tr>
<tr>
<td>HEB</td>
<td>485</td>
</tr>
<tr>
<td>HUN</td>
<td>718</td>
</tr>
<tr>
<td>ITA</td>
<td>23602</td>
</tr>
<tr>
<td>NOR</td>
<td>722</td>
</tr>
<tr>
<td>POR</td>
<td>45711</td>
</tr>
<tr>
<td>RUS</td>
<td>42346</td>
</tr>
<tr>
<td>SPA</td>
<td>51469</td>
</tr>
<tr>
<td>SWE</td>
<td>723</td>
</tr>
<tr>
<td>Total</td>
<td>2083103</td>
</tr>
</tbody>
</table>

Monolingual retrieval for German

<table>
<thead>
<tr>
<th>Retrieval Method</th>
<th>Relevant Retrieved Docs</th>
<th>Average Precision</th>
<th>Precision at 0.1 Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>522</td>
<td>0.1688</td>
<td>0.7622</td>
</tr>
<tr>
<td>Tokens &amp; Lemmas</td>
<td>516</td>
<td>0.1988</td>
<td>0.7967</td>
</tr>
<tr>
<td>Tokens &amp; Lemmas &amp; UMLSSem</td>
<td>509</td>
<td>0.2236</td>
<td>0.3595</td>
</tr>
<tr>
<td>Tokens &amp; Lemmas &amp; EuroWordNet</td>
<td>500</td>
<td>0.1980</td>
<td>0.5371</td>
</tr>
<tr>
<td>Tokens &amp; Lemmas &amp; NL</td>
<td>526</td>
<td>0.2462</td>
<td>0.6556</td>
</tr>
<tr>
<td>Tokens &amp; Lemmas &amp; Non-related</td>
<td>516</td>
<td>0.2254</td>
<td>0.5841</td>
</tr>
</tbody>
</table>

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 and 7.
  - Then:

<table>
<thead>
<tr>
<th>Position</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
</tr>
<tr>
<td>7</td>
<td>0.37</td>
</tr>
<tr>
<td>Average</td>
<td>0.83</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th># of Documents</th>
<th>Token</th>
<th>Token + Lemma</th>
<th>TL + Mesh</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>322</td>
<td>516</td>
<td>526</td>
</tr>
<tr>
<td>EN</td>
<td>517</td>
<td>635</td>
<td>648</td>
</tr>
</tbody>
</table>

Monolingual retrieval for German vs. English: Precision

<table>
<thead>
<tr>
<th># of Documents</th>
<th>Token</th>
<th>Token + Lemma</th>
<th>TL + Mesh</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>16</td>
<td>21.8</td>
<td>24.52</td>
</tr>
<tr>
<td>EN</td>
<td>34.55</td>
<td>33.2</td>
<td>35.43</td>
</tr>
</tbody>
</table>

Heuristic morphology for German

- For all adjectives, nouns, verbs without lemma create lemmas by:
  - remove inflectional suffix for all adjectives (e.g. *arthroskopen* → *arthroskopisch*)
  - separate all hyphenated compounds (e.g. *HWS-Distorsion* → *HWS + Distorsion*)
  - separate all Ns, Adj, 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: # Rel. Retr.

<table>
<thead>
<tr>
<th># of Documents</th>
<th>Token</th>
<th>Token + Lemma</th>
<th>TL + Mesh</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>322</td>
<td>516</td>
<td>526</td>
</tr>
<tr>
<td>DE +</td>
<td>322</td>
<td>594</td>
<td>600</td>
</tr>
<tr>
<td>EN</td>
<td>617</td>
<td>635</td>
<td>648</td>
</tr>
</tbody>
</table>
Monolingual retrieval for German vs. English II: Precision

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.)
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
      Δ Query 9: Patientengesteuerte Analgesie, Indikationen und Grenzen
    • EWN-Code: Behandlung, Therapie
      Δ Query 91: Behandlung von Plattenepithel-karzinomen
    • EWN-Code: Ursache
      Δ Query 108: Ursachen von Schluckstörungen

• English EWN codes are much more detailed.
  • Examples from the Queries:
    • EWN-Codes: analgesia; indication; limit
      Δ Query 9: Patient-controlled analgesia, indications and limits
    • EWN-Codes: treatment; cell; carcinoma
      Δ Query 91: Treatment of squamous cell carcinoma
    • EWN-Code: cause
      Δ Query 108: Cause of dysphagia

<table>
<thead>
<tr>
<th></th>
<th>mAvP</th>
<th>Recall</th>
<th>Prec at 0.1</th>
<th>Prec at 10 docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>0.0025</td>
<td>86</td>
<td>0.0116</td>
<td>0.0120</td>
</tr>
<tr>
<td>EN</td>
<td>0.1178</td>
<td>462</td>
<td>0.3058</td>
<td>0.2440</td>
</tr>
</tbody>
</table>

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