Cross-Lingual Information Retrieval

Language Technology I
Termiology

- monolingual, multilingual, cross-lingual
Use Scenarios (I)

• A user has no knowledge of a target language, i.e., she cannot search for documents in that language at all
  • With CLIR she can make use of media data pools that are indexed with captions in that language, for example for picture pools, music databases, etc.
  • With CLIR she can get a pre-selection of documents that can then be passed on to a translator
Use Scenarios (II)

- a user has only passive knowledge of a target language, i.e., she cannot actively search for documents in that language
  - with CLIR she can make use of relevant texts
Use Scenarios (III)

• a document collection has such a large number of languages that it would be impractical to formulate a query in each of these languages
  • with CLIR one could get relevant documents with only a search query in one of these languages
CLIR approaches

- Machine translation:
  - uses NLP tools like PoS-tagger, parser, morphological analyzers, etc.

- Thesaurus-based approaches
  - manual use of thesauri: “controlled vocabulary” systems
  - automatic use of thesauri: “concept retrieval” systems

- Corpus-based methods: work with frequency analysis
  - Implication: aboutness of the two collections should be similar
MT Approach - Architecture

- **CLIR**
  - Query (en) → Index (de) → Documents (de)
  - Index (en) → Documents (en)

- **Document Translation**
  - Query (en) → Index (en) → Documents (en)
  - Index (en) → Documents (en)

- **Index Translation**
  - Query (en) → Index (en) → Documents (en)

- **Query Translation**
  - Query (de) → Index (de) → Documents (de)
Document Translation

• Problem solved by multiplying the texts
  • Make texts available in all languages
  • multilingual (= several monolingual) retrieval

• Feasibility:
  • Required in some applications
    • Patents, multilingual states (EG, Belgium, …)
  • Impossible in other areas (Internet)

• Evaluation:
  • From costly to impossible
  • Results depend on translation quality
    • translation dictionary updates invalidate search on existing document pool (→retranslate everything)
Index Translation

• Idea:
  • multilingual Index
    • Analyze query in query language, translate terms
    • Search with all document language index terms
  • (Problem of retranslation of the hits)

• Feasibility:
  • Not feasible
    • Ambiguity of index terms
    • Multiword terms not in index
    • Context dependency of translations

=> Organize the index as a special resource!
Query Translation

- Approach: Translation of query
  - Analyse and translate the query terms
  - Search in (monolingual) Backend-System
- Evaluation
  - Backend database stays unchanged
  - Translation changes do not affect document base
  - Cross-lingual component as system frontend
    - contains multilingual linguistic resource
    - Which is also usable for re-translation
    - And can be maintained independently
  - Cross-linguality is transparent for the users
  - Fine-tuning between frontend and backend required
MT Approach

• pros:
  • straightforward (if an MT system is available)
  • user can directly use the retrieved documents
  • documents usually have more context which allows more robust MT than for query translation

• cons:
  • translation of document collections may be very time consuming
  • offline translation of document collections may require lots of additional storage
  • inherits most weaknesses of MT and MT system implementations
Thesaurus-Based Approach: “Thesauri”

- thesaurus: a resource which organizes the terminology of a domain of knowledge, i.e., an ontology for terminology
- multilingual thesauri encode
  - usually: cross-linguistic synonymy
  - sometimes: hierarchical relations between terms (hyperonymy, hyponymy, etc.)
  - seldom: associative relations between terms
- the thesaurus-based approach to CLIR
  - uses multilingual thesauri
  - has a rather broad definition of a thesaurus
- examples of multilingual thesauri used for CLIR:
  - simple cross-language synonym lists
  - collection of concepts with attached cross-lingual information
  - “classic” syntax and semantics lexicons
Query translation

Your search will be carried out with the following translations of your query. You can modify the translation by:

1. turning off unwanted translations
2. adding your own translations in the text fields

**English query terms**

- euro
- introduction

**French translations**

- euro
- instauration
- introduction
- présentation

**German translations**

- Euro
- Empfehlungsschreiben
- Einleitung
- Einführung

search | advanced search | help

login | I want to register
email | about mulinex
Thesaurus-Based Approach: “Thesauri”

• pros:
  • very productive, especially for skilled users
  • works transparently for the user
  • unambiguous mapping between the query and the target document

• cons:
  • very expensive to create good thesauri
  • target documents must be labeled with concepts
  • may be difficult to use for unexperienced users (e.g., because of the manual selection of the intended concept)
  • doesn’t scale
  • restricted to certain domains
  • IR queries can only be as precise as the predefined thesaurus concepts
Corpus-Based Approach

- use of statistical information about term usage from parallel corpora
- usually based on two general retrieval principles:
  - target documents with frequent usage of query terms are potentially more relevant than target documents with infrequent query term usage
  - rare query terms are more useful than query terms that are very frequent in the overall target document collection
- pros:
  - usage of recent terminology (as provided by the corpora) is possible
- cons:
  - parallel corpora needed
  - restricted to the domains of the parallel corpora
**Pseudo-Relevance Feedback**

- Enter query terms in French
- Find top French documents in parallel corpus
- Construct a query from English translations
- Perform a monolingual free text search
## Learning From Document Pairs

- Count how often each term occurs in each pair
  - *Treat each pair as a single document*

<table>
<thead>
<tr>
<th></th>
<th>English Terms</th>
<th></th>
<th></th>
<th></th>
<th>Spanish Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E1</td>
<td>E2</td>
<td>E3</td>
<td>E4</td>
<td>E5</td>
</tr>
<tr>
<td>Doc 1</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 2</td>
<td>8</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 3</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Doc 4</td>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 5</td>
<td>4</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Similarity based Dictionaries

• Automatically developed from aligned documents
  • Terms $E_1$ and $E_3$ are used in similar ways
    • Terms $E_1$ & $S_1$ (or $E_3$ & $S_4$) are even more similar
• For each term, find most similar in other language
  • Retain only the top few (5 or so)
CLIR Research Community

  - Arabic, English, Spanish, Chinese, etc.
  - CLIR at TREC: http://www.glue.umd.edu/~dlrg/clir/trec2002/
- Cross-Language Evaluation Forum (CLEF)
  - European languages
  - http://www.clef-campaign.org/
- NTCIR (NII Test Collection for IR Systems)
  - with related workshops
- Information Retrieval for Asian Language (IRAL)
  - international workshop
- and quite a few others