Cross-Lingual Information Retrieval
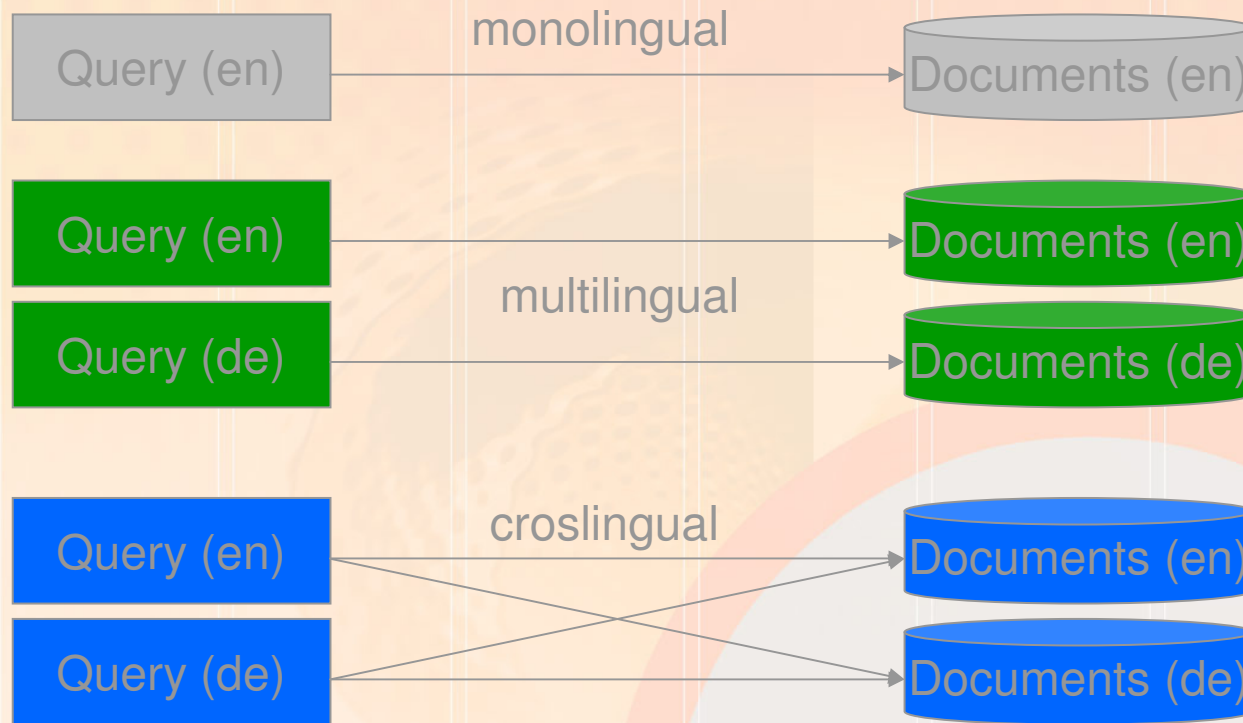
Language Technology I

Terminology

- monolingual, multilingual, cross-lingual

| Query (en) | monolingual → | Documents (en) |

| Query (en) | → | Documents (en) |
| Query (de) | multilingual → | Documents (de) |

| Query (en) | croslingual | Documents (en) |
| Query (de) | | Documents (de) |

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*
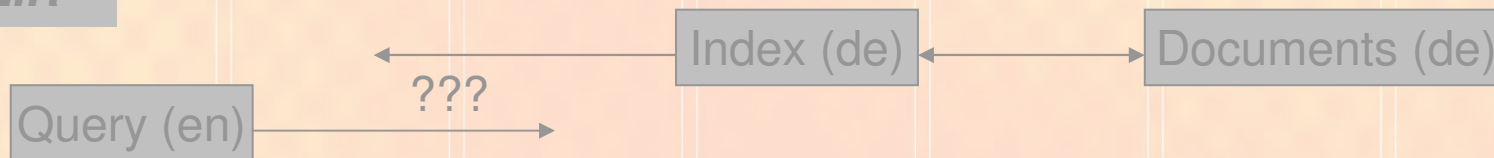
CLIR approaches

- Machine translation: uses morphological parser 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
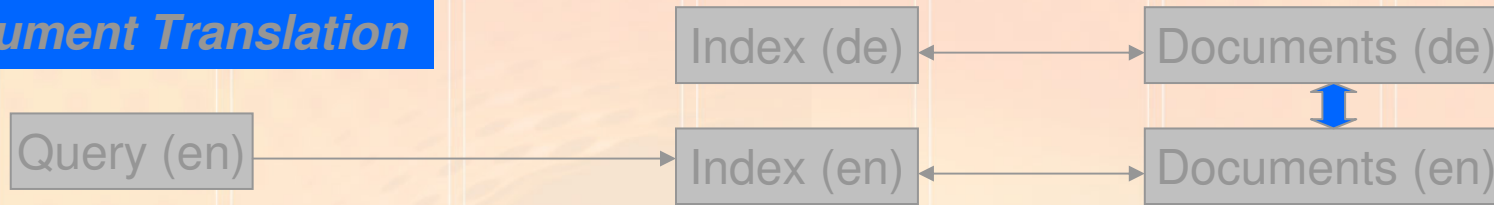 - *Implication: aboutness of the two collections should be similar*
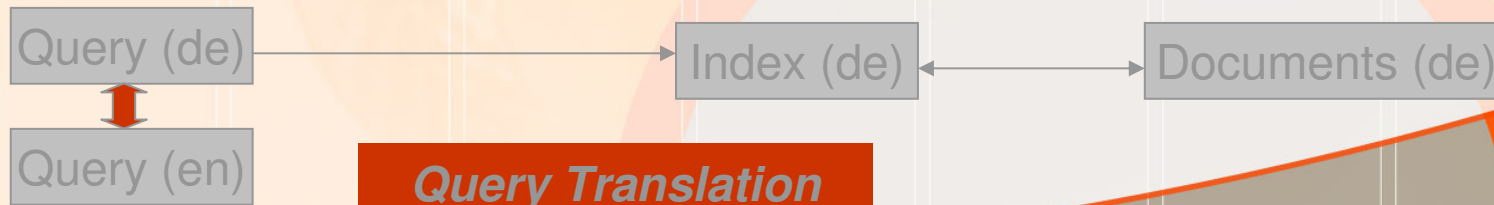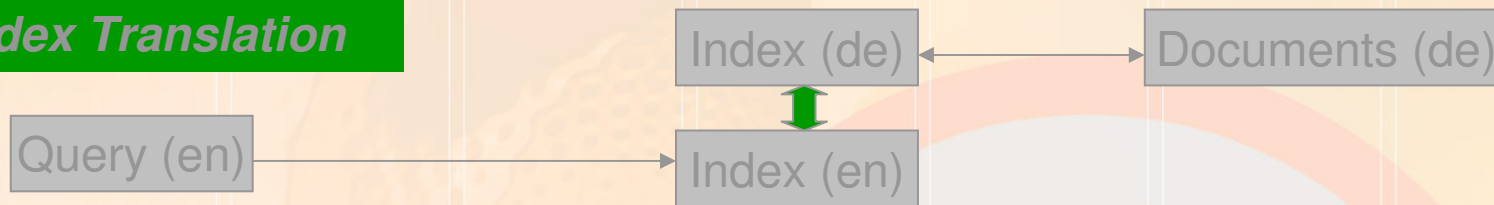
MT Approach - Architecture

CLIR

Index (de) ← Documents (de)

Query (en) → ??? →

Document Translation

Index (de) ↔ Documents (de)

Query (en) → Index (en) ↔ Documents (en)

Index Translation

Index (de) ↔ Documents (de)

Query (en) → Index (en)

Query (de) → Index (de) ↔ Documents (de)

Query (en)

Query Translation

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*

Fehler: mistake, fault, error, bug
nuclear: Kern~, zentral, nuklear
power: Macht, Kraft, Strom
plant: Pflanze, Unternehmen

=> 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*

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*

CLIR Research Community

- Text REtrieval Conference (TREC, http://trec.nist.gov/)
  - *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)
  - *http://research.nii.ac.jp/ntcir/index-en.html*
  - *with related workshops*
- Information Retrieval for Asian Language (IRAL)
  - *internaltional workshop*
- and quite a few others