

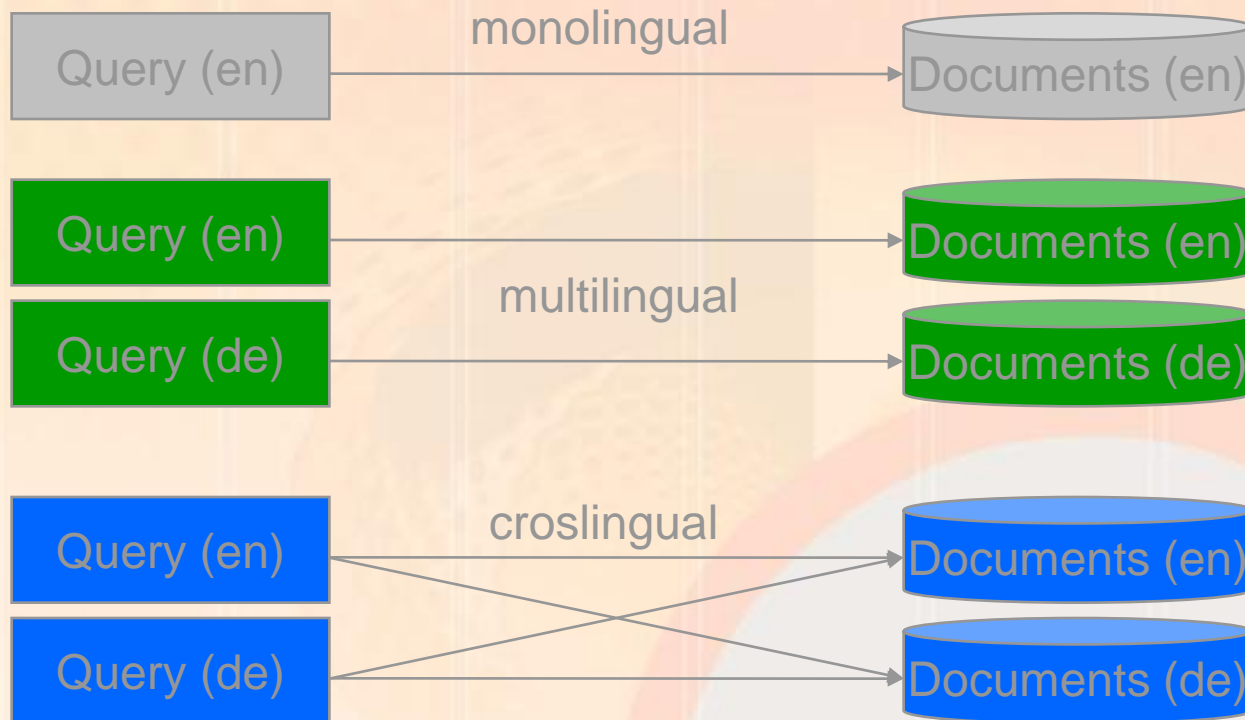
Cross-Lingual Information Retrieval

Language Technology I



Terminology

- 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

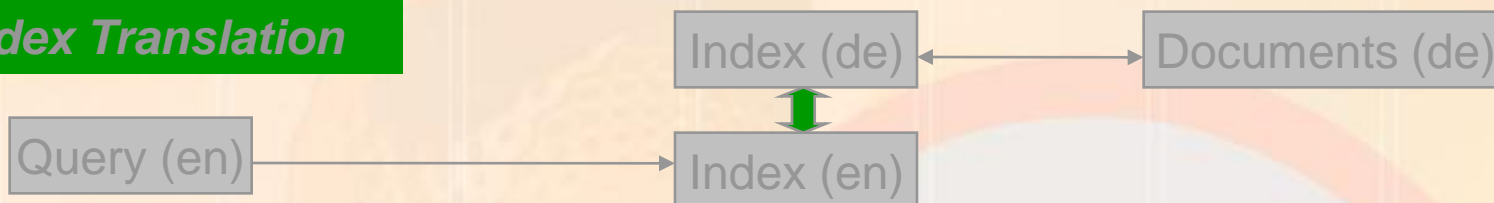
CLIR



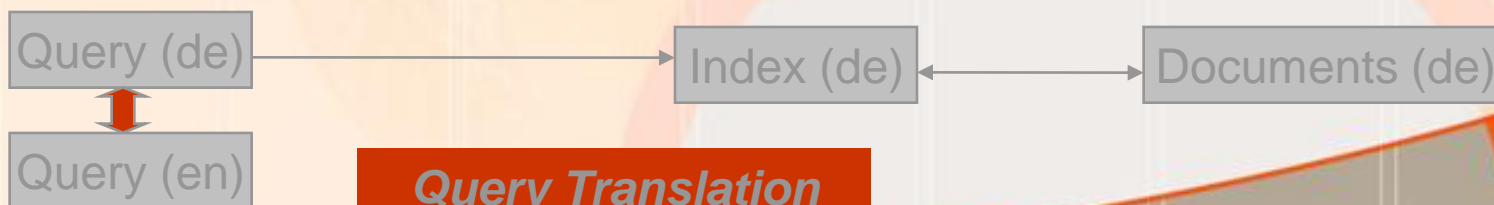
Document Translation



Index Translation



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*



MULINEX Search - Netscape

File Edit View Go Communicator Help

MULINEX Hemm The Force Wi THE Presenta

Location: <http://mulinex.dfi.de/mulinexproto/cgi-bin/mulinex.sh>

Back Forward Reload Home Search Guide Print Security Stop



english deutsch français

search advanced search help

Search

Tailor the search engine to fit your needs and preferences.

- ▶ [About Mulinex](#)
- ▶ [I want to register](#)
- ▶ [login](#)

Search for

The language of the query is

Find documents in

- English
- French
- German

[search](#) | [advanced search](#) | [help](#)
[login](#) | [I want to register](#)
[e-mail](#) | [about mulinex](#)

Document: Done


MULINEX Query Assistant - Netscape

File Edit View Go Communicator Help

MULINEX Hemm The Force Wi THE Presenta

Location: http://mulinex.dfki.de/mulinexproto/cgi-bin/mulinex.sh

Back Forward Reload Home Search Guide Print Security Stop



english deutsch français
search advanced search help

Query translation

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

- turning off unwanted translations
- adding your own translations in the text fields

English query terms	French translations	German translations
<input checked="" type="checkbox"/> euro	<input checked="" type="checkbox"/> euro <input type="text"/>	<input checked="" type="checkbox"/> Euro <input type="text"/>
<input checked="" type="checkbox"/> introduction	<input type="checkbox"/> instauration <input checked="" type="checkbox"/> introduction <input checked="" type="checkbox"/> présentation <input type="text"/>	<input type="checkbox"/> Empfehlungsschreiben <input checked="" type="checkbox"/> Einleitung <input checked="" type="checkbox"/> Einführung <input type="text"/>
<input type="text" value="Germany"/>		

[search](#) | [advanced search](#) | [help](#)
[login](#) | [I want to register](#)
[e-mail](#) | [about mulinex](#)

Document: Done





MULINEX Search Results - Netscape

File Edit View Go Communicator Help

MULINEX Hemm THE Presenta

Location: http://mulinex.dfki.de/mulinexproto/cgi-bin/mulinex.sh

Back Forward Reload Home Search Guide Print Security Stop

mulinex

english deutsch français

search advanced search help

Personal search
Tailor the search engine to fit your needs and preferences.

- Feedback
- I want to register
- login

Search for: euro introduction Germany

The language of the query is: English

Find documents in:

- English
- French
- German

search query assistant

You searched for the English query **euro introduction Germany** in German, French, English.

- all documents
- deutsch
- français
- english

100 documents 22 documents 57 documents 21 documents

hide summaries

French

1 [Charte PME OEC](#)

Category: Politics, Legal, Macintosh, Finance

Summary: CHARTE DE LA PROPARATION DES PME □ L'EURO. Questionnaire PME-EURO. Ventas, politique commerciale, marketing. Achats, politique diapprovisionnement, logistique. Gestion financière. Ressources humaines. Système d'information et informatique. Comptabilité, comptes annuels et information de gestion. '

Summary in: English German

<http://www.finances.gouv.fr/euro/charte/98-111-2.htm> Size 18 K

German

2 [Europarl: der Euro - VIERTE WAHLPERIODE \(1994-1999\)](#)

Category: Legal, Politics, Finance, Travel

Summary: VIERTE WAHLPERIODE (1994-1999). 1. Legislativberichte. 2. Nichtlegislative Berichte. 3.LAUFENDE ARBEITEN. Europarl: der Euro - VIERTE WAHLPERIODE. VIERTE WAHLPERIODE Legislativberichte Nichtlegislative Berichte Laufende Arbeiten 1. Legislativberichte A4-0379/96 - PV 28 11 96 - ARI

Document: Done



Thesaurus-Based Approach: “Thesauri”

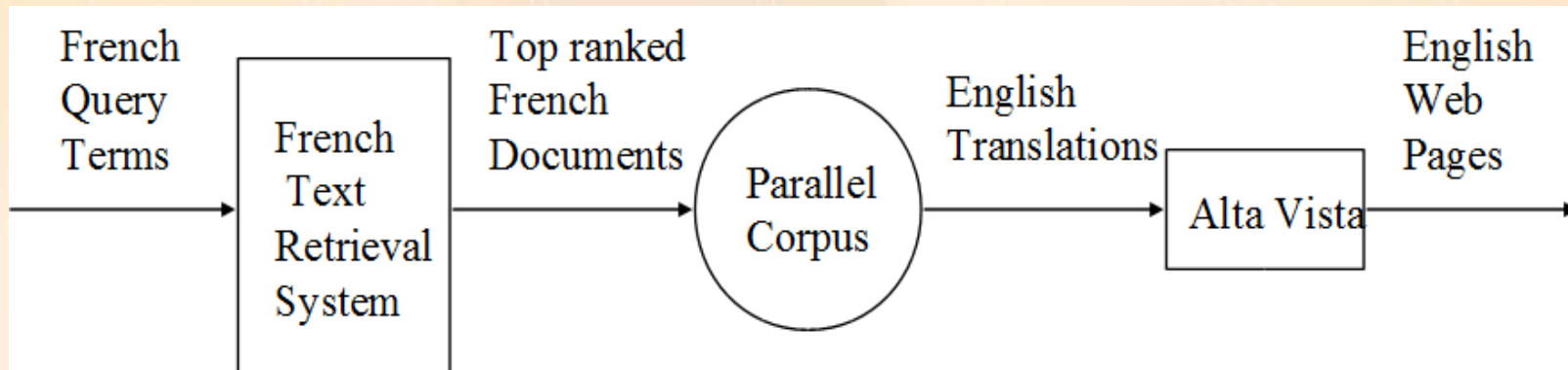
- pros:
 - *very productive, especially for skilled users*
 - *works transparently for the user*
 - *unambiguous mapping between the query and the target document*
- cons:
 - *very expensive to create good thesauri*
 - *target documents must be labeled with concepts*
 - *may be difficult to use for unexperienced 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*

	English Terms					Spanish Terms			
	E1	E2	E3	E4	E5	S1	S2	S3	S4
Doc 1	4		2			2			1
Doc 2	8		4			4			2
Doc 3		2		2			2	1	
Doc 4		2	1				2		1
Doc 5	4				1	2		1	

Similarity based Dictionaries

- Automatically developed from aligned documents
 - *Terms E1 and E3 are used in similar ways*
 - *Terms E1 & S1 (or E3 & S4) are even more similar*
- For each term, find most similar in other language
 - *Retain only the top few (5 or so)*



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)
 - *international workshop*
- and quite a few others