Cross-Language Information Retrieval

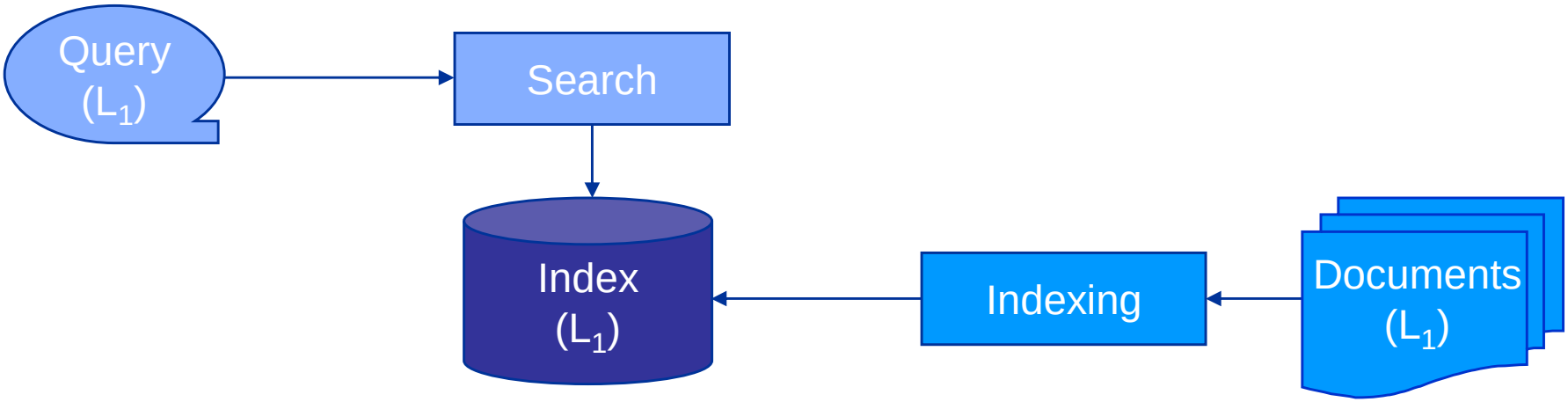
Language Technology I,
December 8, 2008

Christian Spurk
Christian.Spurk@dfki.de
DFKI, Language Technology Lab

(lecture partly based on slides by Hans Uszkoreit and Feiyu Xu)

What is Cross-Language IR (CLIR)? (I)

- Recap: what is IR?
  - LT World (http://www.lt-world.org/) says: “Information Retrieval is the process of locating information that fits a user’s information need, which is usually expressed as a search query.”
- in (monolingual) IR the search query and the located information are encoded in the same (natural) language:



What is Cross-Language IR (CLIR)? (II)

- CLIR is Information Retrieval (IR) on documents with the query being in a language other than the document language.
- terms used in the definition:
  - "language" means natural language (NL) here
  - "documents" can either be text documents or any other media (audio, video, semi-structured data, multimedia etc.) containing NL information
  - a "query" is the user's information need in the form of words
  - the "document language" is the original language of a target document

What is Cross-Language IR (CLIR)? (III)

- not to be confused with:
  - multi-language search engine which allows to query in different languages but which for each language only retrieves documents in the query language
- synonyms for CLIR:
  - MLIR (Multi-Lingual Information Retrieval)
  - TIR (Translingual Information Retrieval)

→ a subset of Multilingual Information Access

Terms Related to CLIR

- source language: the language of the query
- target language(s): the language(s) of the document(s) which are searched

… and some general IR terms:

- relevance: “the fit of the retrieved information with the information need” (LT World)
- precision: the ratio of relevant information in the retrieved data relating to the overall retrieved data
- recall: the ratio of relevant information in the retrieved data relating to the relevant information available overall

Why CLIR? (I)

Top Ten Languages Used in the Web 2008

(Number of Internet Users by Language)

	% of all Internet Users	Internet Users by Language	Internet Penetration by Language	Language Growth in Internet (2000–2008)	2008 Estimated World Population for the Language
English	29.4 %	430,802,172	21.1 %	203.5 %	2,039,114,892
Chinese	18.9 %	276,216,713	20.2 %	755.1 %	1,365,053,177
Spanish	8.5 %	124,714,378	27.6 %	405.3 %	451,910,690
Japanese	6.4 %	94,000,000	73.8 %	99.7 %	127,288,419
French	4.7 %	68,152,447	16.6 %	458.7 %	410,498,144
German	4.2 %	61,213,160	63.5 %	121.0 %	96,402,649
Arabic	4.1 %	59,853,630	16.8 %	2,063.7 %	357,271,398
Portuguese	4.0 %	58,180,960	24.3 %	668.0 %	239,646,701
Korean	2.4 %	34,820,000	47.9 %	82.9 %	72,711,933
Italian	2.4 %	34,708,144	59.7 %	162.9 %	58,175,843
TOP 10 LANGUAGES	84.9 %	1,242,661,604	23.8 %	278.3 %	5,218,073,846
Rest of the Languages	15.1 %	220,970,757	15.2 %	580.4 %	1,458,046,442
WORLD TOTAL	100.0 %	1,463,632,361	21.9 %	305.5 %	6,676,120,288

source: http://www.internetworldstats.com/stats7.htm

Why CLIR? (II)

- in general: better access to more information
- more specificly: (from Douglas W. Oard's keynote at IRAL 99)
  - societal benefits: information exchange to improve understanding
  - economic benefits: information to provide competitive advantage
  - crisis response: language differences can produce costly delays

→ allow anyone to retrieve information that is available in any language

Generic Application Scenarios of CLIR (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 make use of factoid textual data which is language independent, for example registers of names
  - with CLIR she can get a preselection of documents that can then be passed on to a translator

Generic Application Scenarios of CLIR (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
- 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

The Three Main Approaches to CLIR

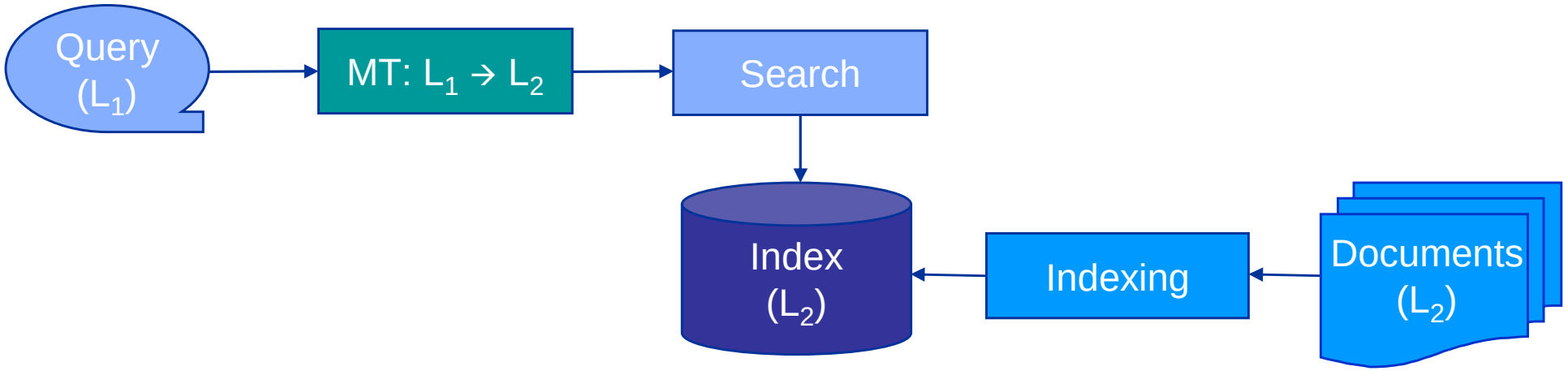
(according to a taxonomy developed by Oard & Dorr, 1996)

- use of Machine Translation (MT)
  - translation of the search query
  - and/or translation of target documents
- thesaurus-based approaches
  - manual use of thesauri: “controlled vocabulary” systems
  - automatic use of thesauri: “concept retrieval” systems
- corpus-based approaches
  - use of statistical information about term usage from parallel corpora

MT Approach: Query Translation (I)

(1) Search query translation

- helps the user formulating or using a query in the target language by automatically translating her query from the source language to the target language

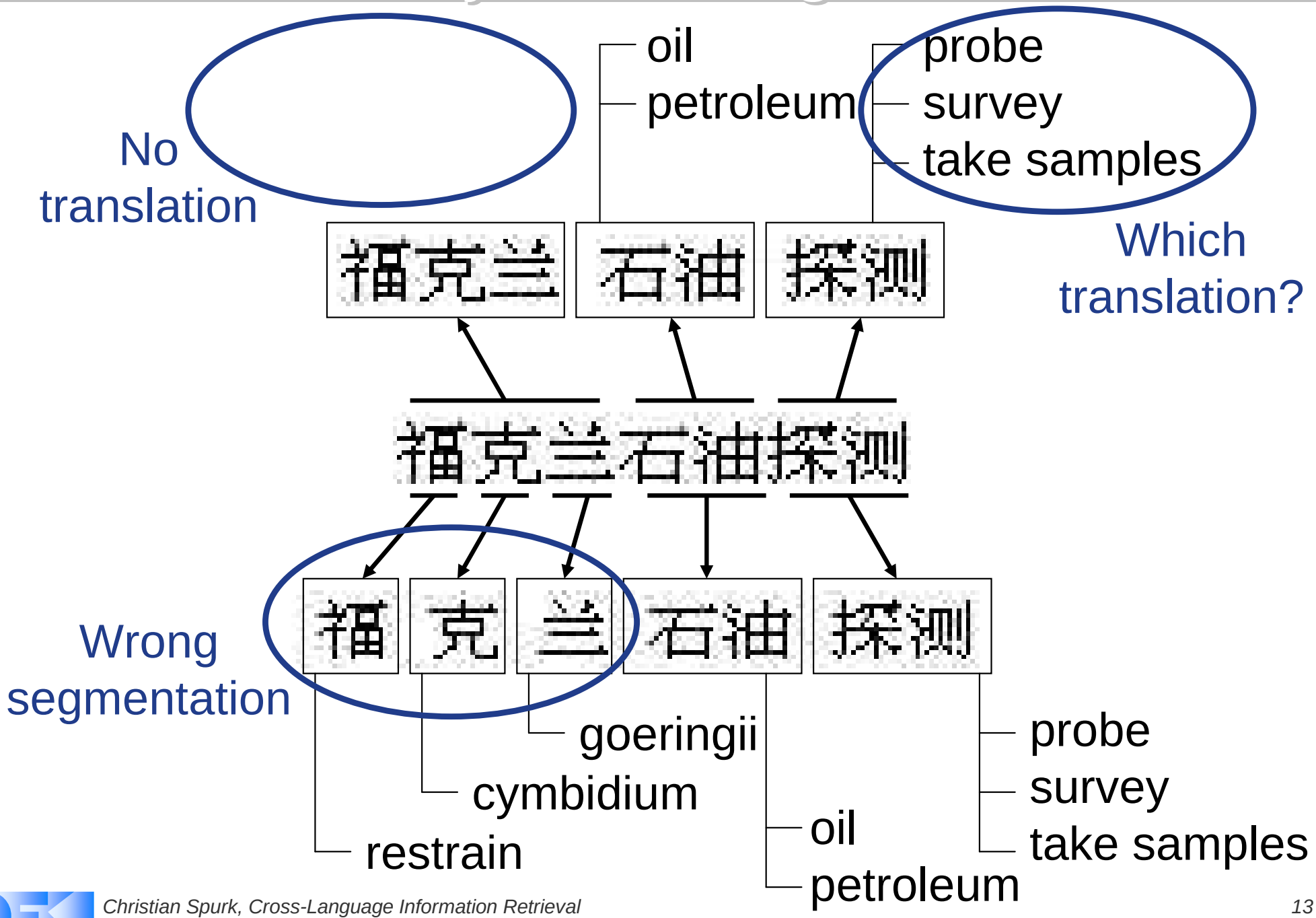


MT Approach: Query Translation (II)

- pros:
  - straightforward (if an MT system is available)
  - once the query is translated, the retrieval is relatively fast
- cons:
  - user may not always be able to make use of the target language documents
  - queries are usually short which makes MT error-prone
  - inherits most weaknesses of MT (cf. three key challenges for MT on the next slide) and MT system implementations

MT: Three Key Challenges

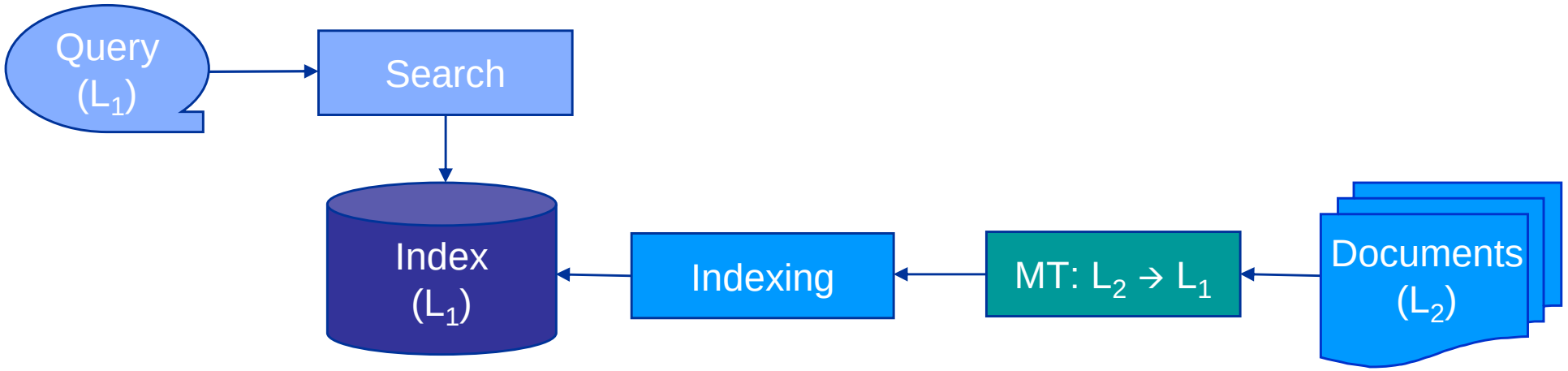
(taken from Douglas W. Oard's IRAL 99 keynote)



MT Approach: Document Translation (I)

(2) Target document translation

- translates target documents before searching through them
- translation is usually done offline and the cached translations are then searched



MT Approach: Document Translation (II)

- 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

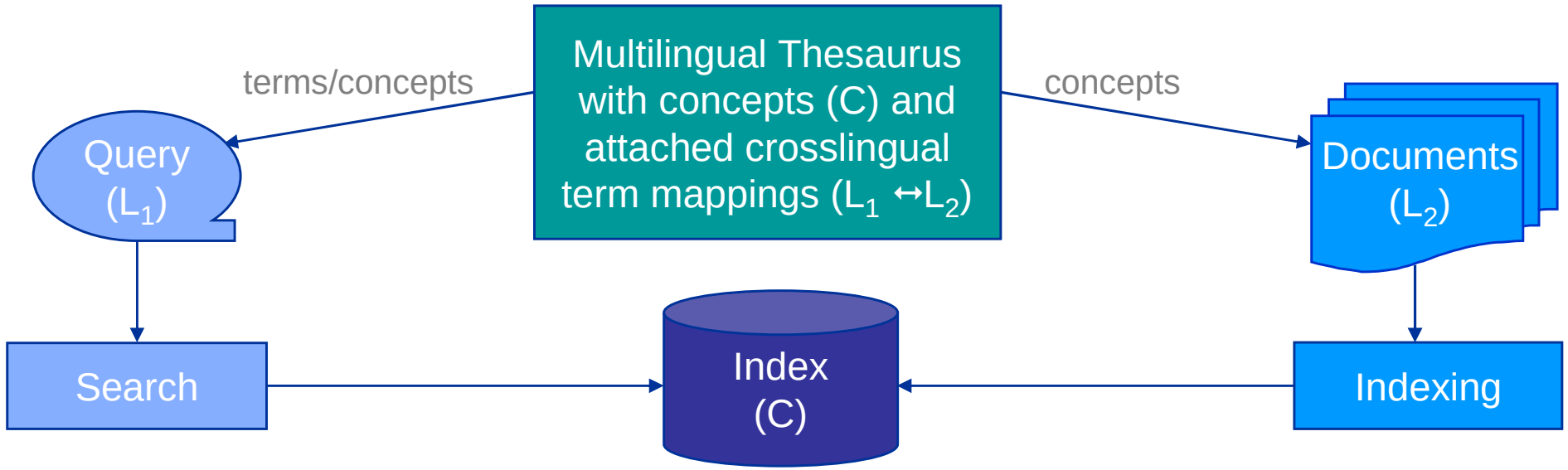
Thesaurus-Based Approach: “Thesauri”

- 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

T.-based CLIR: Controlled Vocabulary (I)

(1) Manual use of thesauri: controlled vocabulary

- each term in the thesaurus uniquely specifies a concept
- target documents are labeled with concepts from the thesaurus
- with the terms from the thesaurus the user manually specifies the concepts she would like to have in the IR query



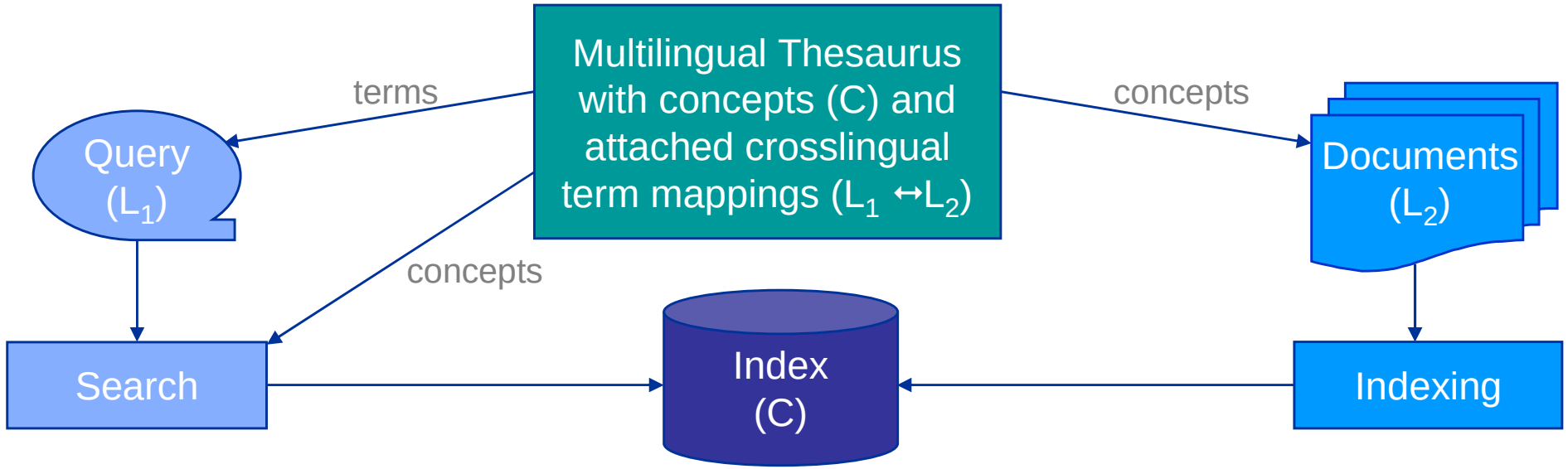
T.-based CLIR: Controlled Vocabulary (II)

- 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

T.-based CLIR: Concept Retrieval (I)

(2) Automatic use of thesauri: concept retrieval

- basically like the controlled vocabulary approach
- terms in the IR query for which there is no unambiguous crosslingual mapping are automatically mapped to concepts with either:
  - concept substitution (simple): ambiguous terms in the query are automatically replaced with a list of all possible concepts
  - query expansion (more sophisticated): concept relations from the thesaurus are used to "intelligently" replace ambiguous terms in the query with possible concepts



T.-based CLIR: Concept Retrieval (II)

- pros:
  - increases recall
- cons:
  - may decrease precision (especially in the case of concept substitution)
  - very expensive to create good thesauri
  - target documents must be labeled with concepts
  - doesn’t scale
  - restricted to certain domains
  - IR queries can only be as precise as the predefined thesaurus concepts

Corpus-Based Approach to CLIR

- 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

Further Information and References

- Douglas Oard's research web page:
  - highly recommended for more detailed information
  - http://terpconnect.umd.edu/~oard/research.html

- a more recent "state of the art" description by Feiyu Xu:
  - http://www.dfki.de/~feiyu/KBIRAF.pdf

- Oard, D. W. and Dorr, B. J. (1996): *A Survey of Multilingual Text Retrieval.* Technical report at the University of Maryland (USA). http://www.glue.umd.edu/~dlrg/filter/papers/mlir.ps

- Fluhr, C. (1995): *Multilingual information retrieval.* In: Cole, R. A.; Mariani, J; Uszkoreit, H.; Zaenen, A. and Zue, V. (eds.): *Survey of the State of the Art in Human Language Technology.* pp. 391–305. Center for Spoken Language Understanding, Oregon Graduate Institute. <http://www.lt-world.org/HLT_Survey/ltw-chapter8-5.pdf>