
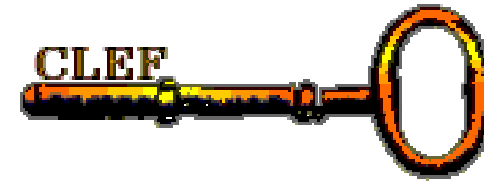


Cross-Language Evaluation Forum (CLEF)



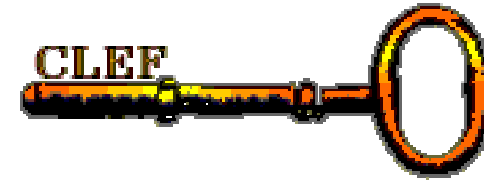
Carol Peters, IEI-CNR, Pisa, Italy

Cross-Language Evaluation Forum



- A collaboration between the DELOS Network of Excellence for Digital Libraries and the US National Institute for Standards and Technology (NIST)
- Extension of CLIR track at TREC (1997-1999)

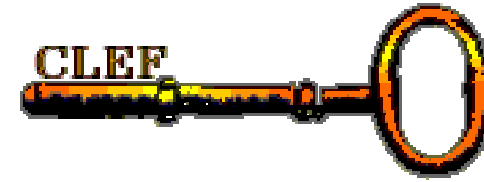
Cross-Language Evaluation Forum



Partners

- English: NIST, Gaithersburg MD
- French: University of Zurich
- German: IZ Sozialwissenschaften, Bonn,
- Italian: IEI-CNR, Pisa, Italy
- Spanish: IEEC-UNED, Madrid
- Test Collections/Processing Results:
Eurospider, Zurich

Cross-Language Evaluation Forum

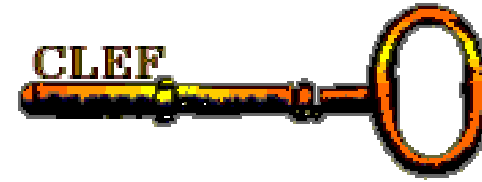


■ Main Goals

Promote research in cross-language system development for European languages by providing an appropriate infrastructure for:

- CLIR system evaluation, testing and tuning
- comparison and discussion of results

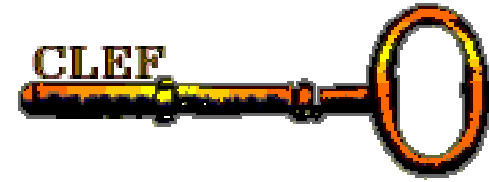
CLEF 2000 Task Description



Four evaluation tracks in CLEF 2000:

- multilingual information retrieval
- bilingual information retrieval
- monolingual (non-English) information retrieval
- domain-specific IR

CLEF 2000 Document Collection



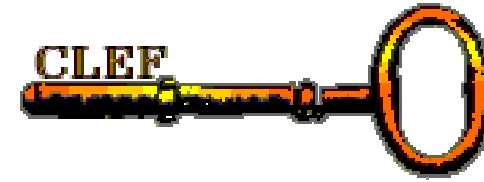
Multilingual Comparable Corpus:

- English: Los Angeles Times
- French: Le Monde
- German: Frankfurter Rundschau + Der Spiegel
- Italian: La Stampa

Similar for genre, content, time

CLEF 2000

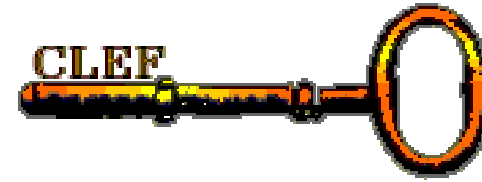
The Queries



- TREC-style topics (title, description, narrative)
 - 40 topics built (on the basis of the document collections)
 - topics translated into all 4 doc. languages
 - translations are from source to target; aim is natural-language renderings
 - additional topics created for NL, FI, SP, SV (mainly by volunteer groups)

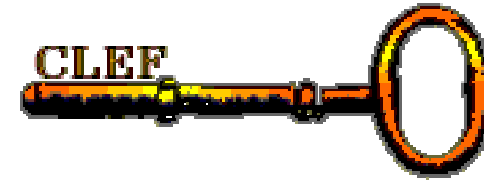
CLEF 2000

A Topic



- **Title:** European Industry
- **Description:** What are the factors that damage the competitiveness of European industry on the world's markets?
- **Narrative:** Relevant documents discuss factors that render European industry and manufactured goods less competitive with respect to the rest of the world, e.g. North America or Asia. Relevant documents must report data for Europe as a whole rather than for single European nations.

CLEF 2000 Task Description



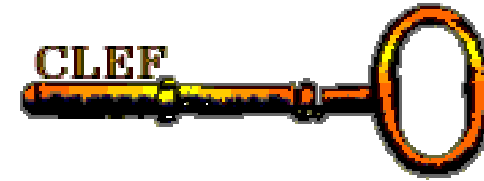
Multilingual Information Retrieval

Task: search multilingual document collection in 4 languages (E,F,G,I) for relevant documents

Goal: retrieve documents for all languages, listing results in a merged, ranked list

Topic languages: EN,FR,DE,IT + NL, FI, SV, SP

CLEF 2000 Task Description



Bilingual Information Retrieval

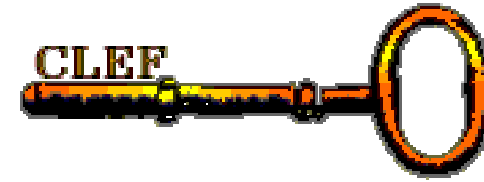
Task: query target document collection in English

Goal: retrieve documents for target language, listing results in ranked list

Topic languages: FR, DE, IT + NL, FI, SV, SP

CLEF 2000

Task Description



Monolingual (non-English) Information Retrieval

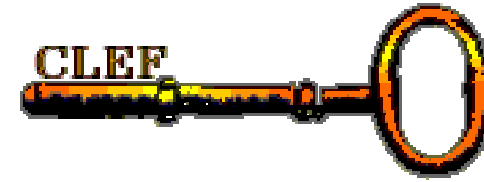
Task: querying document collections in DE, FR, IT

Goal: acquire better understanding of language-dependent retrieval problems

- different languages present different retrieval problems
- issues involved include word order, morphology, diacritic characters, language variants

CLEF 2000

Task Description



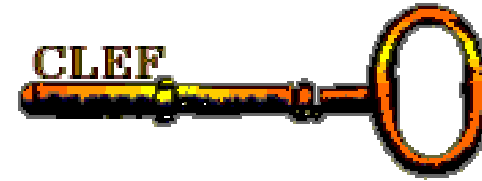
GIRT: Domain-Specific Cross-Language Information Retrieval

Task: querying a structured database from a vertical domain (social sciences) in German

- German/English/Russian thesaurus and English translations of document titles
- Monolingual or cross-language task

Goal: study CLIR in a vertical domain

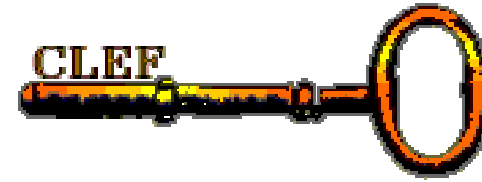
CLEF 2000 Participation



- 20 groups: 8 N.American; 12 European
- Runs were submitted for all tasks:
 - Cross-Language = 16 groups
 - Multilingual = 11 groups
 - Bilingual = 9 groups
 - Monolingual = 12 groups
 - Domain-specific = 3 groups
- A total of 90 runs were submitted

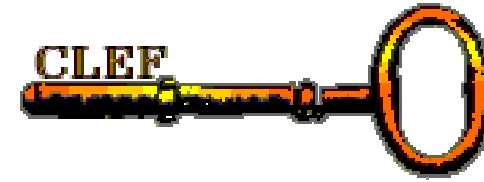
CLEF 2000

Topic languages by Group



- Runs were submitted using all topic languages
 - Multilingual: EN=9; DE=2; NL=2; SV=1
 - Bilingual: DE=4; FR=3; NL=3; SP=3; IT=1; FI=1; SW=1
 - Monolingual: DE=10; FR=9; IT=8

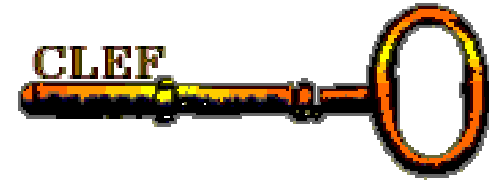
CLEF 2000 Results



- traditional and innovative approaches were tested
- increased participation by European groups
- building of CLIR system development community

CLEF 2000

Approaches to CLIR

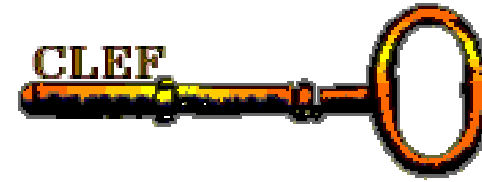


- commercial MT systems (Systran, Lernout and Hauspie Power Translator)
- bilingual dictionary look-up
- aligned parallel corpora (web-derived)
- similarity thesaurus (using comparable corpora)

Different strategies experimented for query expansion and results merging

CLEF 2000

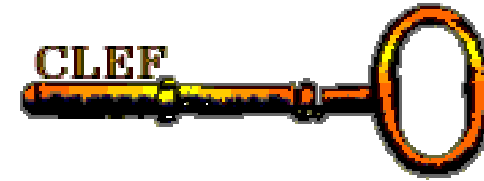
Techniques Tested



Evaluation favours testing

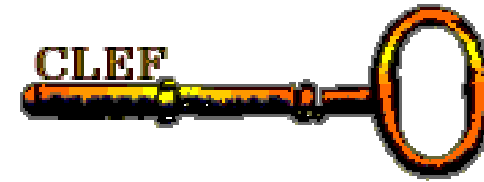
- lexicon derived from parallel corpora vs MT
- use of NLP methods, e.g. phrase identification, compound processing, morphosyntactic analysis
- language independent stemming techniques
- interactive query term disambiguation
- lexical triangulation

CLEF 2000 Workshop



- Results will be presented and discussed in a Workshop, 22 September, Lisbon, Portugal
- Decisions will be taken for future campaigns

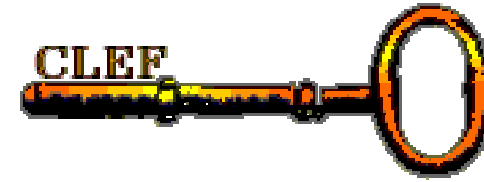
Cross-Language Evaluation Forum



■ Intentions for the Future

- addition of more languages
- addition of new tasks
- C-L evaluation for other document types
- study evaluation methodologies wrt user needs

Cross-Language Evaluation Forum



- For more information:

<http://www.iei.pi.cnr.it/DELOS/CLEF/>

or

carol@iei.pi.cnr.it