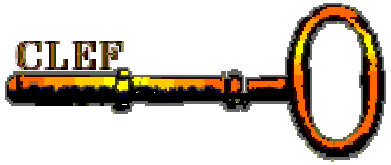
Cross-Language Evaluation Forum - CLEF

Carol Peters
IEI-CNR, Pisa, Italy

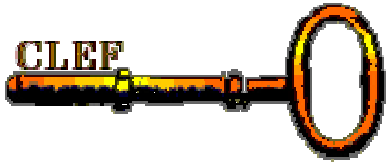
IST-2000-31002
Kick-off: October 2001

Outline



- Project Objectives
- Background
- CLIR System Evaluation
- CLEF Infrastructure
- Results so far

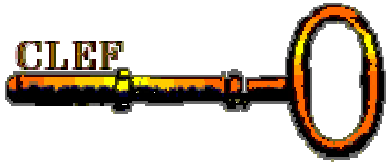
CLEF - Objectives



Promote CLIR research by providing an appropriate infrastructure for:

- system evaluation, testing and tuning
- comparison and discussion of approaches
- building of reusable test-suites for system developers

CLEF Partners



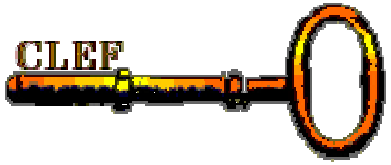
Consortium

- IEI-CNR, Pisa, Italy (Coordinator)
- IZ Sozialwissenschaften, Bonn, Germany
- IEEC-UNED, Madrid, Spain
- Eurospider, Zurich, Switzerland
- ELRA/ELDA, Paris, France
- NIST, Gaithersburg MD, USA

Associated Partners

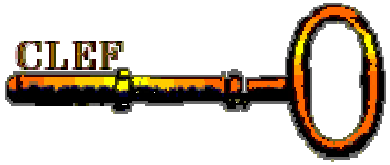
- University of Hildesheim, Germany
- University of Twente, The Netherlands
- University of Tampere, Finland
- INIST, CNRS, France
- University of Maryland, USA

CLEF - Background



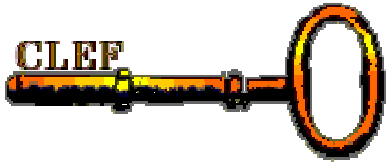
- Extension of CLIR track at TREC (1997-1999)
- CLEF2000 and CLEF2001 funded by DELOS Network of Excellence for Digital Libraries in collaboration with US National Institute for Standards and Technology

Why DELOS?



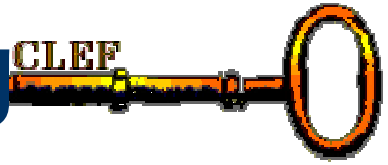
- Cross-language search and retrieval is a key issue in digital libraries
- Big gap between research and application communities

Survey of DL Projects in 5FP



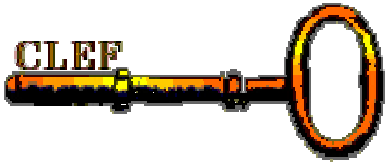
- 14 projects contained collections in multiple languages
- 4 had not considered any kind of multiple language processing
- 10 monolingual retrieval functionality for all languages
- 1 had implemented cross-browsing of collections using common metadata schema
- 6 had some kind of basic cross-language functionality:
  - 5 used multilingual controlled vocabulary / thesaurus
  - 1 used bilingual dictionary search
  - 1 used pseudo relevance feedback (in addition to thesaurus)
  - 1 proposed using similarity search (in addition to controlled vocab)

Why an Accompanying Measure?



- Encourage CLIR system development for European languages
- Disseminate research results to application community

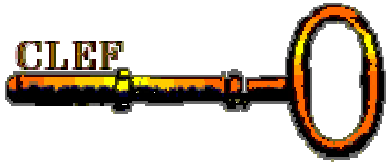
IR System Evaluation



Cranfield Methodology

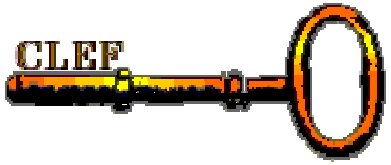
- Laboratory activity which tests system performance on a given task (or set of tasks) under standard conditions
- Permits contrastive analysis of approaches/technologies

Organising an Evaluation Activity



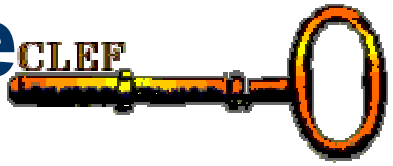
- Select control task(s)
- Provide data to test and tune systems (the test collection)
- Define metrics to be used in results analysis

Test Collection

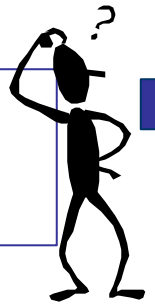


- Set of documents - must be representative of task of interest; must be large
- Set of “topics” - statement of user needs from which system data structure (query) is extracted
- Relevance judgments – judgments vary by assessor but no evidence that differences affect comparative evaluation of systems

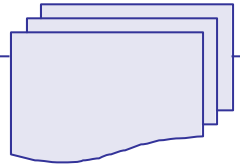
Using Pooling to Create Large Test Collections



Assessors create topics.



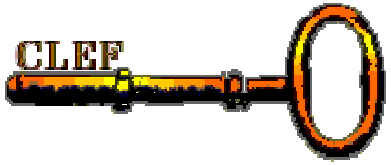
A variety of different systems retrieve the top 1000 documents for each topic.



Form pools of unique documents from all submissions which the assessors judge for relevance.

Systems are evaluated using relevance judgments.

Evaluation Measures



- Recall: measures ability of system to find **all** relevant items

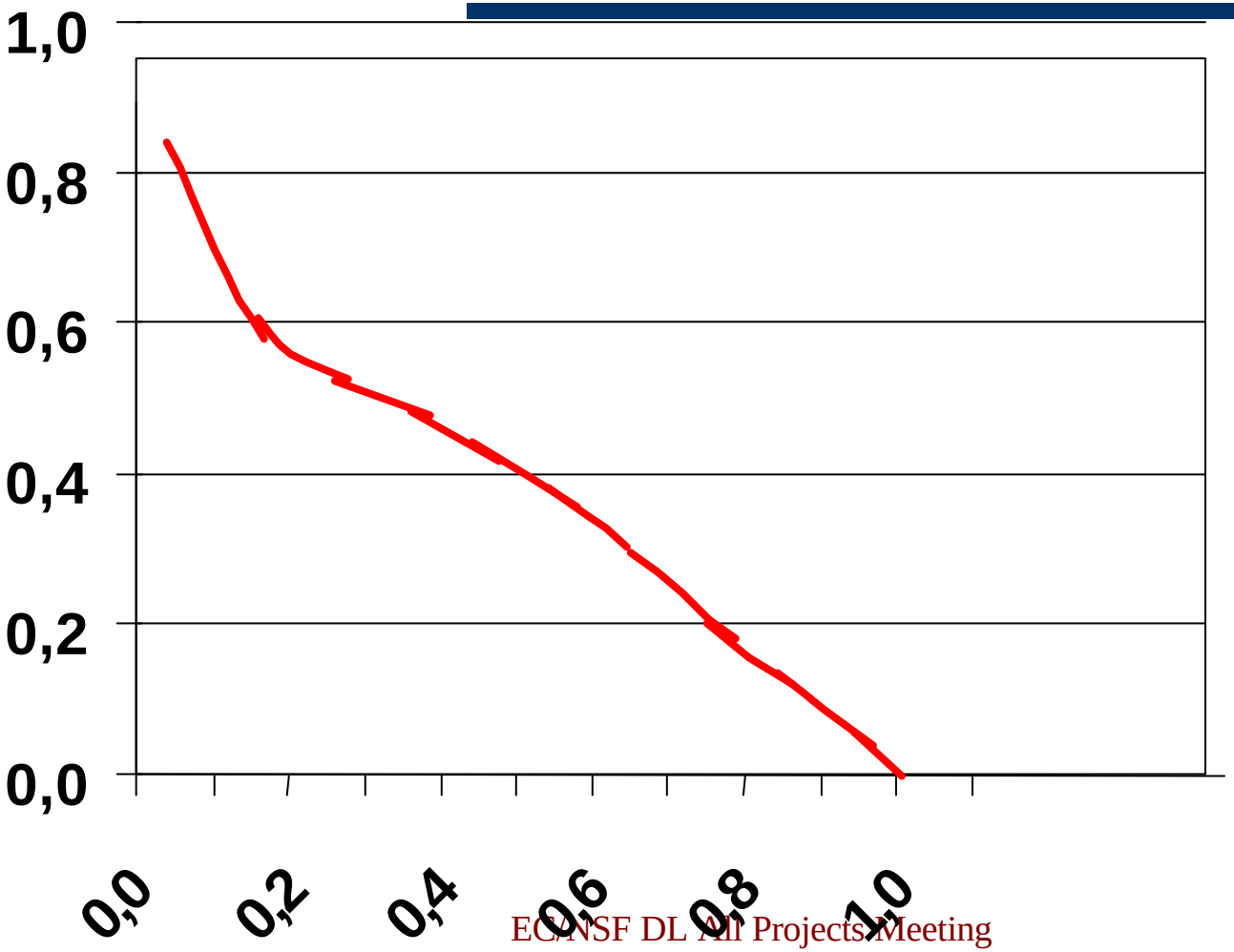
$$\text{recall} = \frac{\text{no. of rel. items retrieved}}{\text{no. of rel. items in collection}}$$

- Precision: measures ability of system to find **only** relevant items

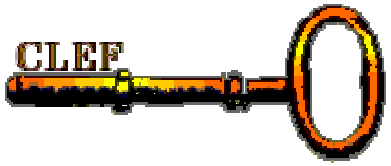
$$\text{precision} = \frac{\text{no. of rel. items retrieved}}{\text{total no. of items retrieved}}$$

Recall-Precision Graph is used to compare systems

Precision-Recall Curve



Cross-language Test Collections



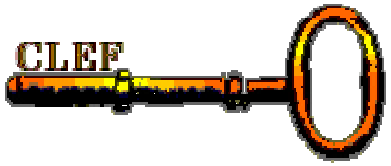
Consistency of data harder to obtain than for monolingual

- parallel or comparable document collections
- multiple assessors per topic creation and relevance assessment (for each language)
- must take care when comparing different language evaluations (e.g., cross run to mono baseline)

Pooling harder to coordinate

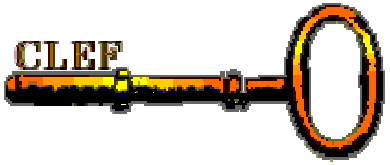
- need to have large, diverse pools for all languages
- retrieval results are not balanced across languages

Main CLIR Evaluation Programs



- **TIDES:** sponsors TREC (Text REtrieval Conferences) and TDT (Topic Detection and Tracking) - Chinese-English tracks in 2000; TREC focussing on English/French - Arabic in 2001/02
- **NTCIR:** Nat.Inst. for Informatics, Tokyo. Chinese-English; Japanese-English C-L tracks
- **CLEF:** Cross Language Evaluation Forum - C-L evaluation for European languages

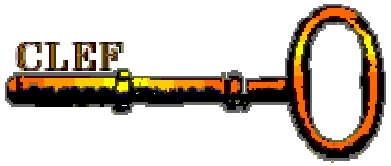
CLEF 2002 Task Description



- Multilingual information retrieval
- Bilingual IR
- Monolingual (non-English) IR
- Mono- and cross-language IR for scientific collections
- Interactive track

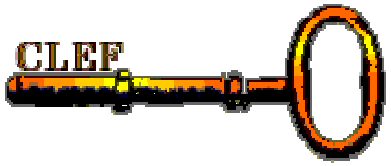
Plus feasibility study for spoken document track (within DELOS – results reported at CLEF)

CLEF 2002 Data Collection



- Multilingual comparable corpus of news agencies and newspaper documents for seven languages (DE,EN,FI,FR,IT,NL,SP).
- Scientific document collections
  - GIRT: German social science docs plus German/English/Russian thesaurus
  - AMARYLLIS: French bibliographic docs plus English/French controlled vocabulary
- Common set of 50 topics (from which queries are extracted) in 10 European (DE,EN,FR,IT,NL,SP, FI + PO, RU,SV) and 2 Asian languages (JP,ZH)

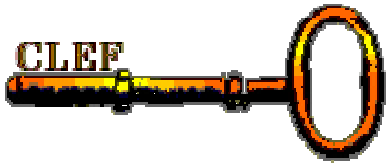
CLEF 2002
Creating the Topics



- Title: European Industry
- Description: What factors 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.

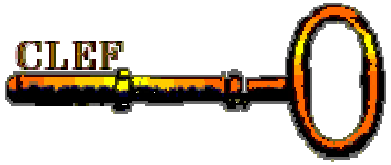
Queries are extracted from topics: 1 or more fields

CLEF 2002
Creating the Queries



- Distributed activity (Bonn, Gaithersburg, Paris, Pisa,Tampere, Twente, Madrid)
- Each group produced 13-15 topics, 1/3 local, 1/3 European, 1/3 international
- Topic selection at meeting in Berlin (50 topics)
- Topics are created in DE, EN,FR,IT,NL,SP and additionally translated to SV,RU,FI and TH,JP,ZH
- Cleanup after topic translation (Hildesheim)

CLEF 2002
Multilingual IR



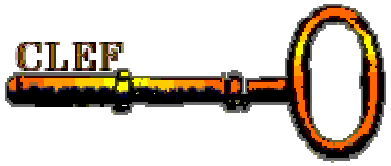
Topics either
DE,EN,FR,IT
FI,NL,SP,PO,
SV,RU,ZH,JP

documents
English German French Italian Spanish

Participant's Cross-Language Information Retrieval System

One result list of DE, EN, FR,IT and SP documents ranked in decreasing order of estimated relevance

CLEF 2002
Interactive CLIR

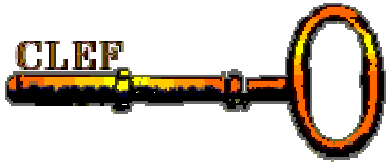


Task: interactive query or interactive document selection in an “unknown” target language

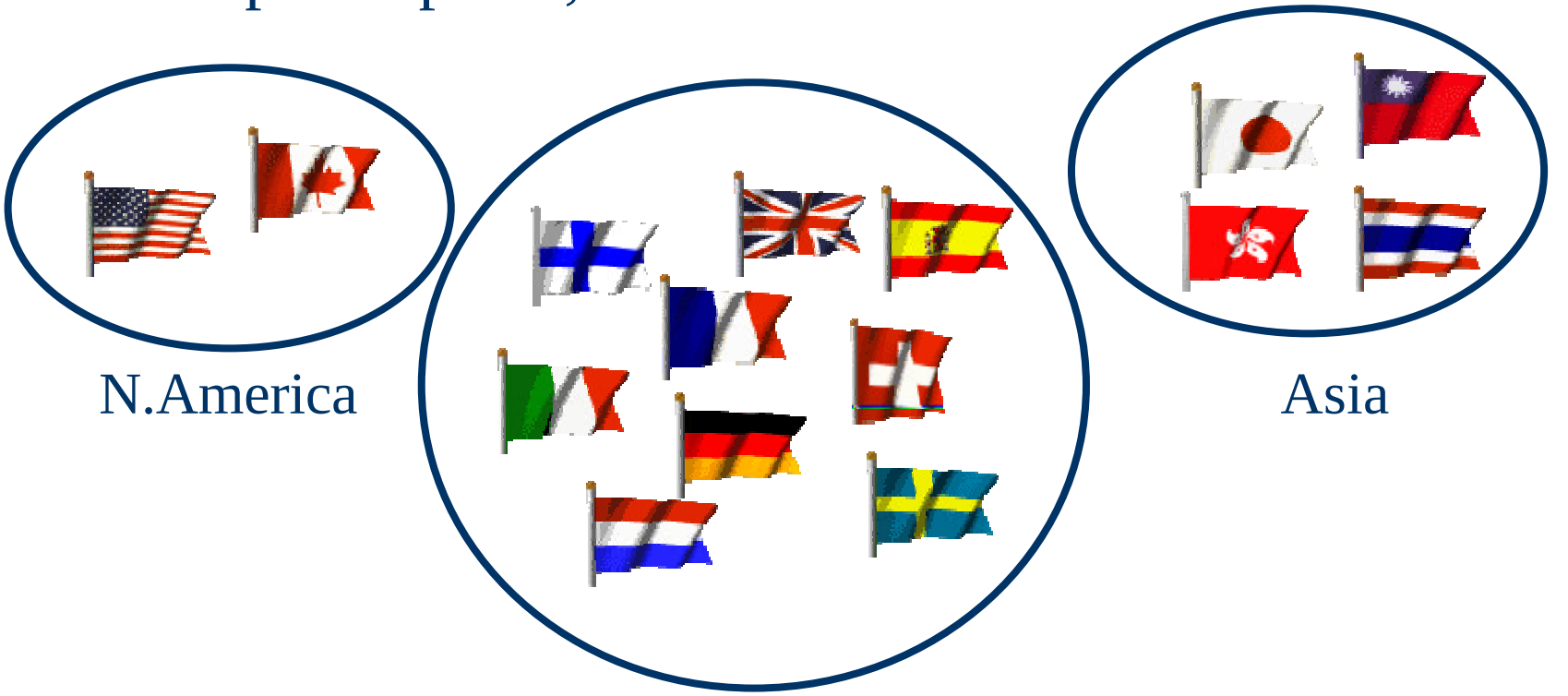
Focus: searcher with passive language abilities in target language / searchers with no language abilities in target language

Goals: explore different approaches to common tasks; evaluate how system assists user/ meets user needs

CLEF 2001: Participation



34 participants, 15 different countries

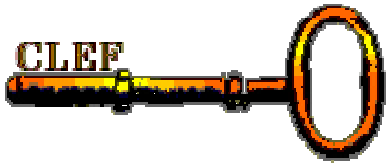


N.America

Asia

Europe

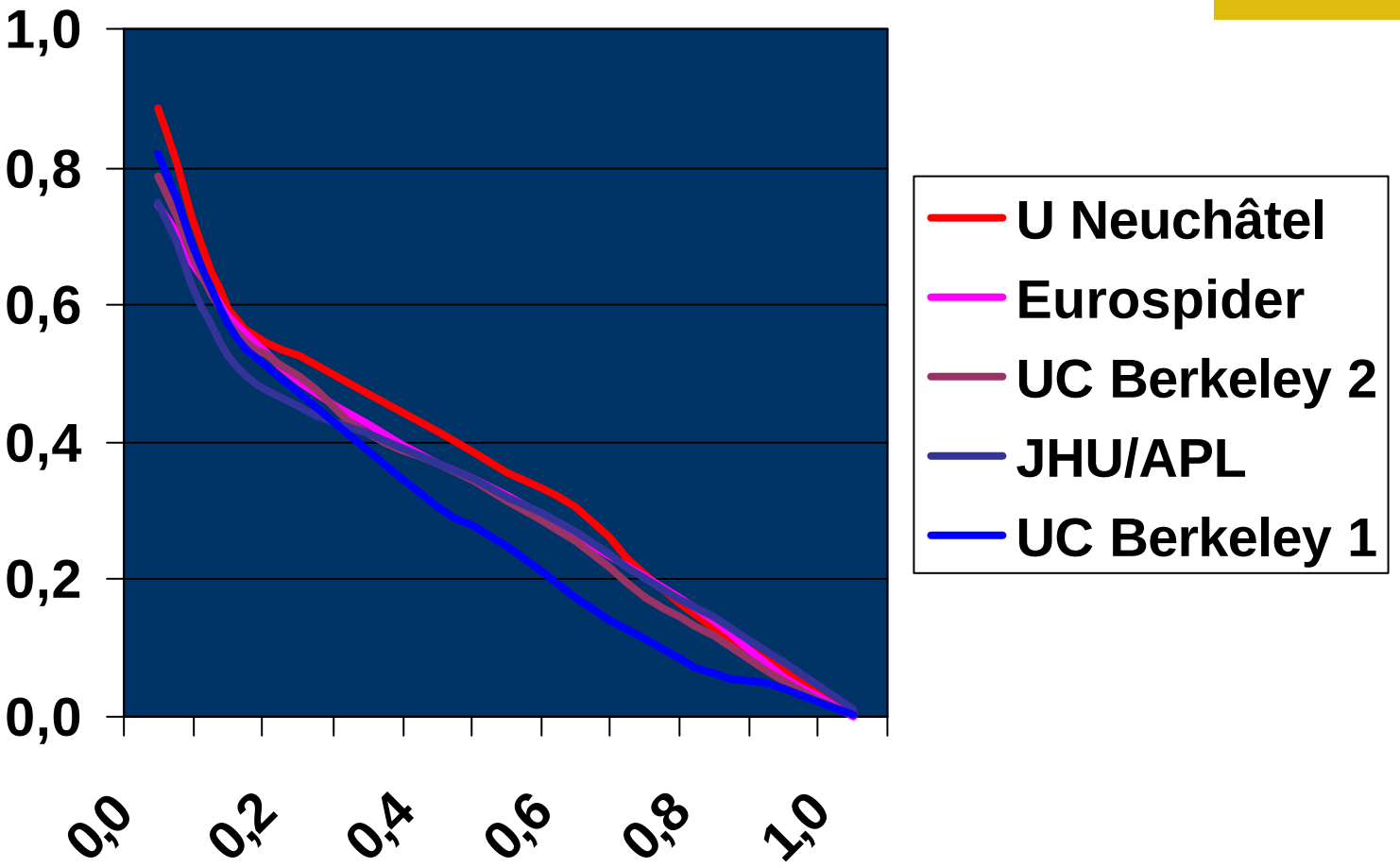
CLEF 2001 Participation



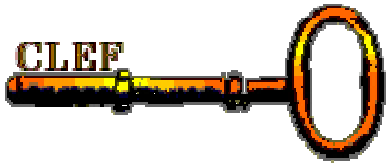
- CMU
- Eidetica
- Eurospider *
- Greenwich U
- HKUST
- Hummingbird
- IAI *
- IRIT *
- ITC-irst *
- JHU-APL *
- Kasetsart U
- KCSL Inc.
- Medialab
- Nara Inst. of Tech.
- National Taiwan U
- OCE Tech. BV
- SICS/Conexor
- SINAI/U Jaen
- Thomson Legal *
- TNO TPD *
- U Alicante
- U Amsterdam
- U Exeter
- U Glasgow *
- U Maryland * (interactive only)
- U Montreal/RALI *
- U Neuchâtel
- U Salamanca *
- U Sheffield * (interactive only)
- U Tampere *
- U Twente (*)
- UC Berkeley (2 groups) *
- UNED (interactive only)

* = also participated in 2000
8 = industry; 26 = academic

CLEF 2001
Multilingual Results



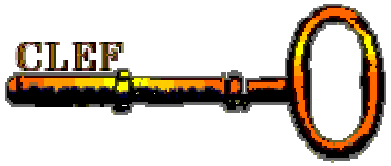
CLEF 2001 Approaches



All traditional approaches used:

- commercial MT systems (Systran, Babelfish, Globalink Power Translator,)
  - both query and document translation tried
- bilingual dictionary look-up (on-line and in-house tools)
- aligned parallel corpora (web-derived)
- comparable corpora (similarity thesaurus)
- conceptual networks (Eurowordnet, ZH-EN wordnet)
- multilingual thesaurus (domain-specific task)

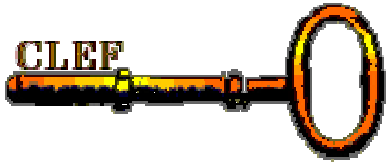
CLEF 2001 Techniques Tested



Text processing for multiple languages:

- Porter stemmer, Inxight commercial stemmer, on-site tools
  - simple generic “quick&dirty” stemming
  - language independent stemming
- separate stopword lists vs single list
- morphological analysis
- n-gram indexing, word segmentation, decompounding (e.g. Chinese, German)
- use of NLP methods, e.g. phrase identification, morphosyntactic analysis

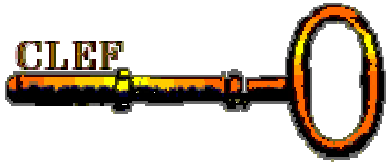
CLEF 2001 Techniques Tested



Cross-language strategies included:

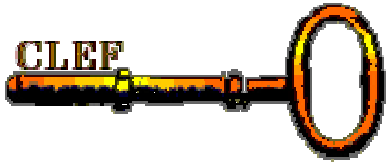
- integration of methods (MT, corpora and MRDs)
- pivot language to translate from L1 -> L2 (DE -> FR,SP,IT via EN)
- N-gram based technique to match untranslatable words
- prior and post-translation pseudo-relevance feedback (query expanded by associating frequent cooccurrences)
- vector-based semantic analysis (query expanded by associating semantically similar terms)

CLEF 2001 Techniques Tested



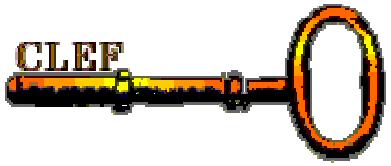
- Different strategies experimented for results merging
- This remains still an unsolved problem

Plans for the Future



- Increase the size of the test collection
- More languages in the test collection
- Provide the possibility to test on different text types
- Provide more task variety (Q&A, Web queries, Text categorization)
- Work with multimedia
- Provide standard resources to permit objective comparison of individual system components
- Focus more on user satisfaction issues (e.g. query formulation, results presentation)

Cross-Language Evaluation Forum



For further information see:

http://www.clef-campaign.org

or contact:

Carol Peters - IEI-CNR

E-mail: carol@iei.pi.cnr.it