Evaluating Information Retrieval Systems

Focusing on systems for Cross Language Information Retrieval

Carol Peters
ISTI-CNR, Pisa, Italy

Outline

- IR System Evaluation
- Cross-Language System Evaluation (CLEF)
- Evaluation for Question Answering Systems (with input from Bernardo Magnini)
- Challenges

What is an IR System Evaluation Campaign?

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

Why we need IR Evaluation

- evaluation saves developers time and money
- evaluation permits hypotheses to be validated and progress assessed
- evaluation creates reusable test collections
- evaluation helps to identify areas where more R&D is needed
- evaluation campaigns can promote research

What should IR System Evaluation Campaign measure?

How well does the system meet the information need?

- System-based evaluation:
  how good are document rankings?
- User-based evaluation:
  how satisfied is the user?

Organising an Evaluation Activity

- select control task(s)
- provide data to test and tune systems
- define protocol and metrics to be used in results assessment

Aim is to produce test collections that permit an objective comparison between systems and approaches

Test Collections

- 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

Test collections must be appropriate for the task; must be stable and valid in order to be reusable

Cranfield Tradition

- Laboratory testing of retrieval systems first done in Cranfield II experiment (1963)
  - fixed document and query sets
  - evaluation based on relevance judgments
  - relevance abstracted to topical similarity
- Laboratory tests less expensive
- Laboratory tests more diagnostic
- Laboratory tests are effective

Cranfield Tradition Assumptions

- Relevance can be approximated by topical similarity
  - relevance of one doc is independent of others
  - all relevant documents equally desirable
  - user information need doesn't change
- Single set of judgments is representative of user population
- Complete judgments (i.e., recall is knowable)

The Case Against the Cranfield Tradition

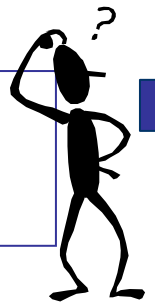
- Relevance judgments
  - vary too much to be the basis of evaluation
  - topical similarity is not utility
  - static set of judgments cannot reflect user's changing information need
- Results on test collections are not representative of operational retrieval systems

Response to Criticism

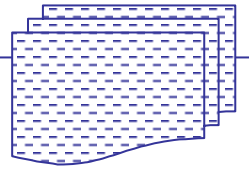
- Goal in Cranfield tradition is to compare systems
  - gives *relative* scores of evaluation measures, not absolute
  - differences in relevance judgments matter only if relative measures based on those judgments change
- Cranfield tests used small collections and assessed relevance for whole collections
- TREC, NTCIR and CLEF have very big collections - thus adopt pooling methodology

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 Graphs used to compare systems

Main IR Evaluation Programs

- **TREC:** Text REtrieval Conference, co-sponsored by NIST and ARDA
- **NTCIR:** Evaluation of Information Access Technologies: IR, QA and C-L Information Access, NII, Tokyo
- **CLEF:** Cross Language Evaluation Forum - C-L evaluation for multilingual IR systems operating on European languages, sponsored by DELOS

TREC 2005

- **Enterprise Track**
  - **satisfy user searching the data of an organization to complete some task**
- Genomics Track
  - domain-specific retrieval
- HARD Track
  - achieve high accuracy retrieval by targeted interaction with user
- Question Answering Track
  - information extraction
- Robust Retrieval Track
  - ad hoc retrieval with focus on individual topic effectiveness
- **SPAM Track**
  - **evaluation of current and proposed e-mail filtering approaches**
- Terabyte Track
  - investigating scalability of IR systems with very large web collection

5th NTCIR Workshop (2004/2005)

- 1. Cross-Lingual Information Retrieval Task (CLIR)
  - Multilingual CLIR
  - Bilingual CLIR and Pivot Bilingual CLIR, and
  - Single language IR
  - Languages: Traditional Chinese, Korean, English and Japanese
- 2. Cross-Language Question Answering Task (CLQA)
  - CLQA is a new task based on CLIR and QAC
  - 5 subtasks (C->C, E->C, C->E, E->J, J->E) are provided (J->J will is provided in QAC).
- 3. **Patent Retrieval Task (PATENT)**
  - **Retrieval task: "Invalidity search"**
  - **Classification task: The purpose is to categorize target patent applications based on the F-term classification system.**
- 4. Question Answering Task (QAC)
  - In a simulated interactive situation, all and only correct answers should be listed as response for each question.
  - Task for reference evaluation: all references are resolved manually and the questions are made understandable in isolation.
- 5. Web Task (WEB)
  - Navigational Retrieval Task: known item search
  - pilot task: Query Term Expansion Task

Cross Language Evaluation Forum

Objectives of CLEF

Promote research and stimulate development of multilingual IR systems for European languages, through

- Creation of evaluation infrastructure and organisation of regular evaluation campaigns for system testing
- Building of an MLIA/CLIR research community
- Construction of publicly available test-suites

CLIR System Evaluation is Complex

CLIR systems consist of integration of components and technologies

- need to evaluate single components
- need to evaluate overall system performance
- need to distinguish methodological aspects from linguistic knowledge

Influence of language and culture on usability of technology needs to be understood

Cross-language Test Collections

Consistency 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

Cross-Language Evaluation Forum

Backgound

- Extension of CLIR track at TREC (1997-1999)
- Partly sponsored by DELOS Network of Excellence for Digital Libraries under FP6 – IST programme
- Mainly dependent on voluntary efforts
- Coordination is distributed for language and for task

Main Institutions involved in Coordination for 2005

- Bulgarian Acad. Sci.
- CELCT, Trento, Italy
- DCU, Ireland
- DFKI, Germany
- ELRA/ELDA, France
- ISTI-CNR, Pisa, Italy
- ITC-irst, Trento, Italy
- Hungarian Acad. Science
- IZ- Bonn, Germany
- Linguateca Sintef, Norway
- LSI-UNED, Spain
- Moscow State U. Russia
- NIST, USA
- Oregon Health & Sci U.
- U.Amsterdam, NL
- UC Berkeley, USA
- U. Hospitals Geneva
- U.Limerick, Ireland
- U. Padova, Italy
- U. Maryland, USA
- U. Sheffield, UK

CLEF 2000-2005: Evaluation Tracks

CLEF 2000

- mono-, bi- and multilingual textual document retrieval on news collections (Ad Hoc)
- mono- and cross-language information on structured scientific data (Domain-Specific)

CLEF 2001

- interactive cross-language retrieval (iCLEF)

CLEF 2002

- cross-language speech retrieval (CL-SR)

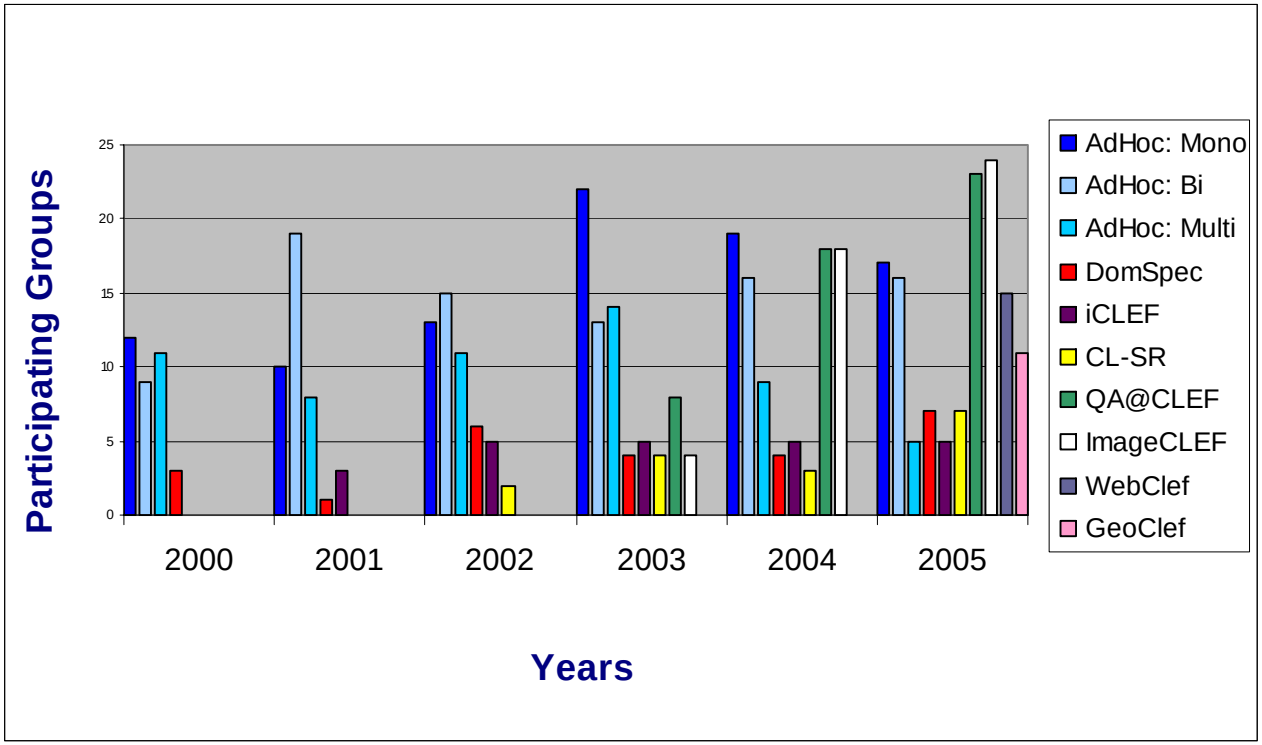
CLEF 2003

- multiple language question answering (QA@CLEF)
- cross-language retrieval in image collections (ImageCLEF)

CLEF 2005

- multilingual retrieval of Web documents (WebCLEF)
- cross-language geographical retrieval (GeoCLEF)

CLEF 2000 – 2005 Shift in Focus



From cross-language textual document retrieval towards all aspects of multilingual multimedia information access

CLEF 2005 Document Collections

Ad Hoc, QA@CLEF, iCLEF, GeoCLEF

- CLEF multilingual comparable corpus of more than 2M news docs in 12 languages: DE,EN,ES,FI,FR,IT,NL,RU,SV, PT, **BG** and **HU** (new in 2005)

Domain-Specific

- The GIRT-4 social science database in EN and DE: more that 300,000 docs
- **The Russian Social Science Corpus:** almost 100,000 docs

ImageCLEF

- St Andrews historical photographic archive: 28,000 images
- CasImage radiological medical database with case notes in FR and EN: 9,000
- **PEIR** 33,000 images, **MIR** 2,000, **PathoPic** 9,000
- **IRMA** collection in EN and DE for automatic medical image annotation: 10,000

CL-SR

- **Malach collection** of spontaneous conversational speech derived from the Shoah archives: 589 hours

WebCLEF

- **EuroGOV**, a multilingual collection of more than 2M webpages crawled from European governmental sites

CLEF 2005: Research in New Directions I

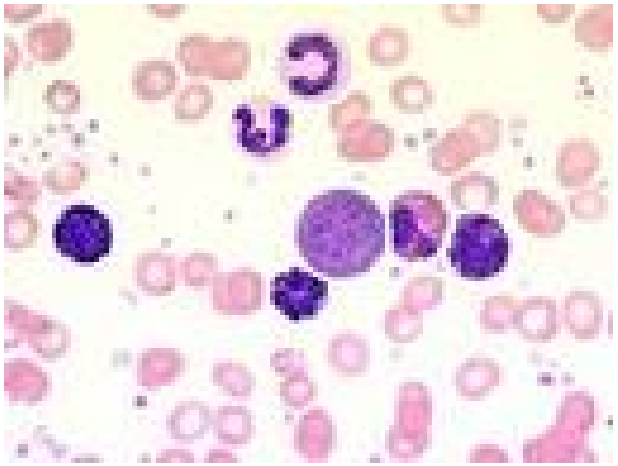
- the ad-hoc track offered a task aimed at studying in-depth the problem of results merging over collections/over languages and at measuring progress in multilingual IR system development over time
- the interactive track was devoted to the comparative study of user-inclusive cross-language search strategies in two contexts: cross-language question answering and retrieval of annotated images
- the question-answering track focused on building up a common and replicable evaluation framework to test both mono- and cross-language QA systems. New types of natural language questions and new evaluation measures – namely the K1 value and r coefficient – were introduced in order to build more challenging test sets and to explore system self-scoring ability
- the image retrieval track explored the use of both text and content-based retrieval methods for cross-language image retrieval; a major goal was to investigate the effectiveness of combining text and image for retrieval

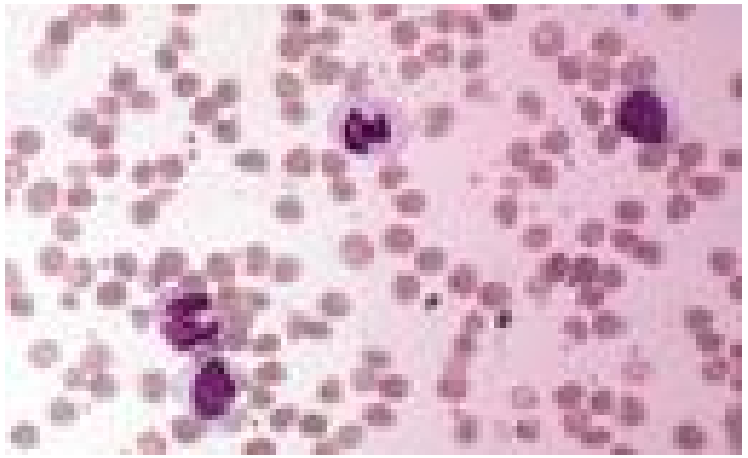
ImageCLEF: An example (topic # 20)

Show me microscopic pathologies of cases with chronic myelogenous leukemia.

Zeige mir mikroskopische Pathologiebilder von chronischer Leukämie.

Montre-moi des images de la leucémie chronique myélogène.





CLEF 2005: Research in New Directions II

- the speech track focused on searching spontaneous speech from oral historical interviews (Shoah archives). Aim is to encourage development of technologies facilitating access to spontaneous speech
- the web track constructed a multilingual web corpus as an important first step towards a cross-lingual web retrieval test collection. This will serve as an important resource to better understand the challenges of multilingual web retrieval
- the geographic information retrieval track was run as a pilot task with the aim of building an evaluation infrastructure to evaluate the retrieval of multilingual documents with an emphasis on geographic search; this was the first time that GIR systems have been evaluated in a multilingual context. The interest in this initial work, especially from industry, was encouraging

CLEF Results in 5 (or 9) years of activity

- Stimulation of research activity in new, previously unexplored areas, such as cross-language question answering, image and geographic information retrieval
- Study and implementation of evaluation methodologies for diverse types of cross-language IR systems
- Creation of a large set of empirical data about multilingual information access from the user perspective;
- Quantitative and qualitative evidence with respect to best practice in cross-language system development
- Creation of important, reusable test collections for system benchmarking building of a strong, multidisciplinary research community
- Cross-language textual document retrieval now considered an „understood“ problem – CLEF has created blueprints for successful truly multilingual systems (L1 -> Ln)
- Documented improvement in system performance for cross-language text retrieval systems

Improvements in System Performance

As test collections and tasks vary over the years it is not easy to document improvements in system performance. In CLIR system evaluation a common method is to compare results against the monolingual baseline

At TREC-6 (1997) the best CLIR systems using queries in English:

- EN -> FR: 49% of best monolingual French IR system
- EN -> DE: 64% of best monolingual German IR system

But to encourage research into multilingual system development, CLEF enforces use of „unusual“ language pairs

- CLEF 2003 Bilingual
  - IT -> ES: 83%
  - DE -> IT: 87%
  - FR -> NL: 82%
- CLEF 2004 Bilingual
  - DE/NL/FI/SV -> FR: 76%
  - IT/FR/ES/RU -> FI: 47%
  - X -> RU: 90%
  - X -> PT: 70%
- CLEF 2004 Bilingual
  - X -> FR: 85%
  - X -> PT: 88%
  - X -> BG: 74%
  - X -> HU: 73%

(performance enhanced in 2004 even with restrictions in topic languages!)

Question Answering at TREC 2005

- Goal: return answers, not document lists
- Tasks:
  - define a target by answering a series of factoid and list questions about that target, plus returning other info not covered by previous questions
  - document ranking task
  - “relationship” question task
- ACQUAINT document collection source of answers for all tasks
  - 3GB text; approx. 1 M newswire articles

Question Answering at NTCIR

- Monolingual QA: interactive task in Japanese
- Cross-language QA (J/E; C/E) + Chinese monolingual: find named entities

Multilingual Question Answering at CLEF

Objective: promote development of multilingual QA systems

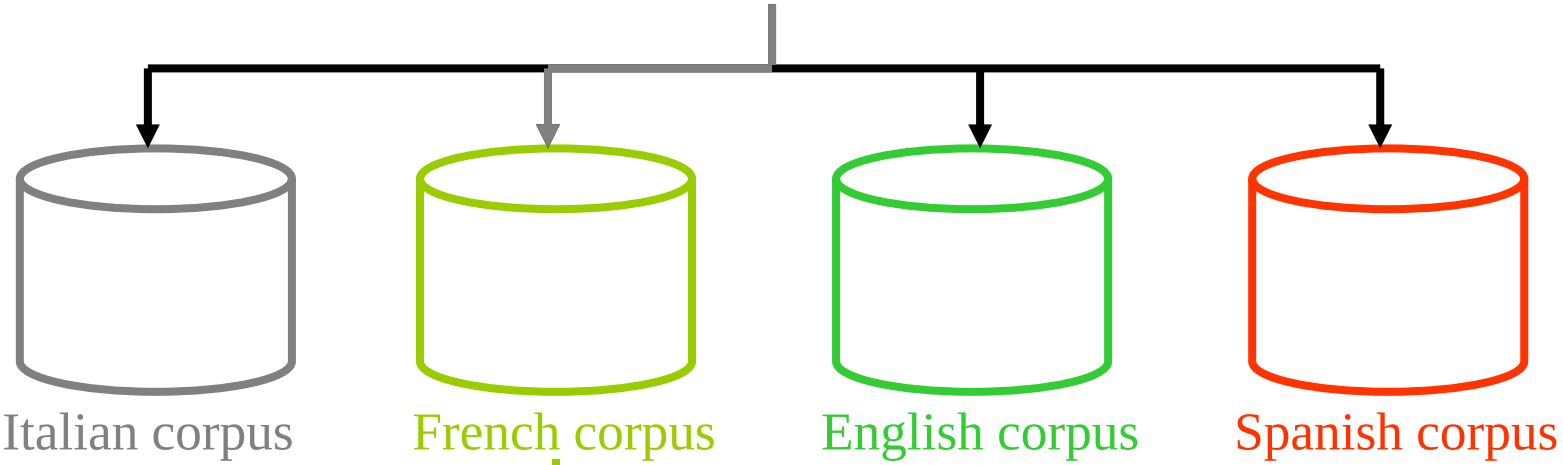
- Monolingual tasks in languages other than English
- Cross-language QA: questions in source language, answers in target language
- Force the QA community to design real multilingual systems
- Check/improve the portability of technologies implemented in current English QA systems

Slides on QA@CLEF provided by Bernardo Magnini

Cross-Language QA

Quanto è alto il Mont Ventoux?

(How tall is Mont Ventoux?)



“Le Mont Ventoux, impérial avec ses *1909 mètres* et sa tour blanche telle un étendard, règne de toutes …”

1909 metri

QA@CLEF

- TREC QA style
  - Prevalence of Factoid questions
  - Exact answers + document id
- Use CLEF corpora (news, 12 langs, most 1994-95)
- Return the answer in the language of the text collection in which it has been found (i.e. no translation of the answer)
- QA-CLEF started as a pilot in 2003

QA@CLEF: Organisation

Nine groups coordinated the QA track:

- **CELCT / ITC-irst** (A. Vallin, D. Giampiccolo): Italian
- **DFKI** (G. Erbach, B. Sacalenau): German
- **ELDA/ELRA** (C. Ayache): French
- **Linguateca** (D. Santos): Portuguese
- **UNED** (A. Penas): Spanish
- **U. Amsterdam** (M. De Rijke): Dutch
- **U. Limerick** (R. Sutcliff): English
- **Bulgarian Academy of Sciences** (P. Osenova): Bulgarian
- **U. Helsinki** (I. Aunimo): Finnish
- plus University of Indonesia: Indonesian as source

Overall coordination: Bernardo Magnini: ITC-irst, Italy

QA@CLEF: Task

Questions: 200 open domain questions; three kinds:

- **Factoid** (ca. 50%): *Which Arab country produces the most oil?*
- **Definition** (ca. 25%): *Who is Josef Paul Kleihues?*
- **Temporally restricted** (ca. 15%): by period, event and date
- **NIL** (ca. 10%):

Answers: exact answer in the target language

- **Document collections:** open domain news corpora (on average 230MB)

Evaluation:

- Each answer: Right, Wrong, ineXact, Unsupported
- Two runs per participant

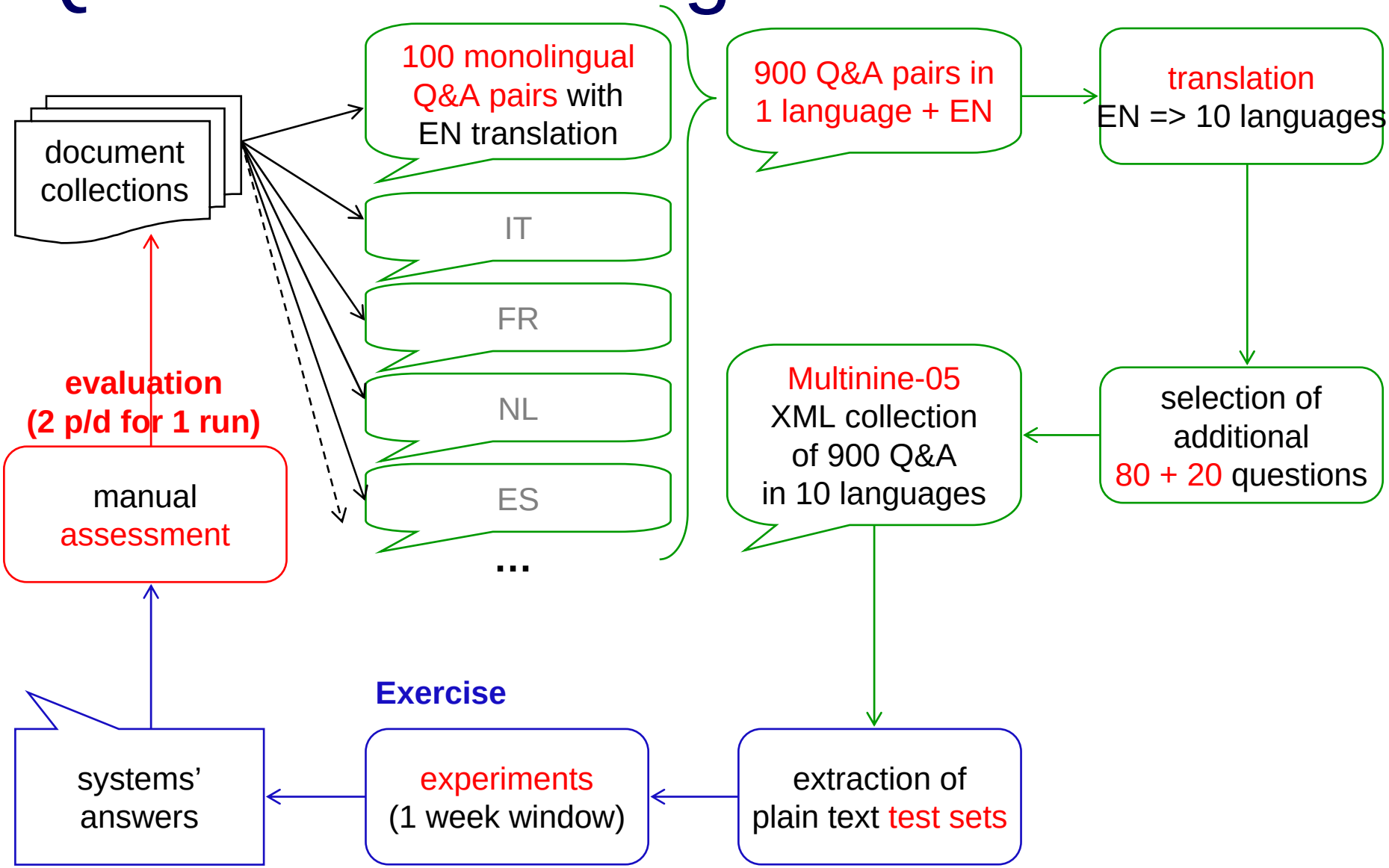
QA-CLEF-05: Evaluation measures

- Main evaluation measure was **accuracy** (fraction of Right responses).
- Whenever possible, a **Confidence-Weighted Score** was calculated:

$$\mathbf{CWS} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{\text{number of correct responses in first i ranks}}{i}$$

- Beside CWS, two new measures were introduced, namely K1 and r value, to take into account both accuracy and confidence

QA-CLEF-05: Organization



QA-CLEF-05: Activated Tasks

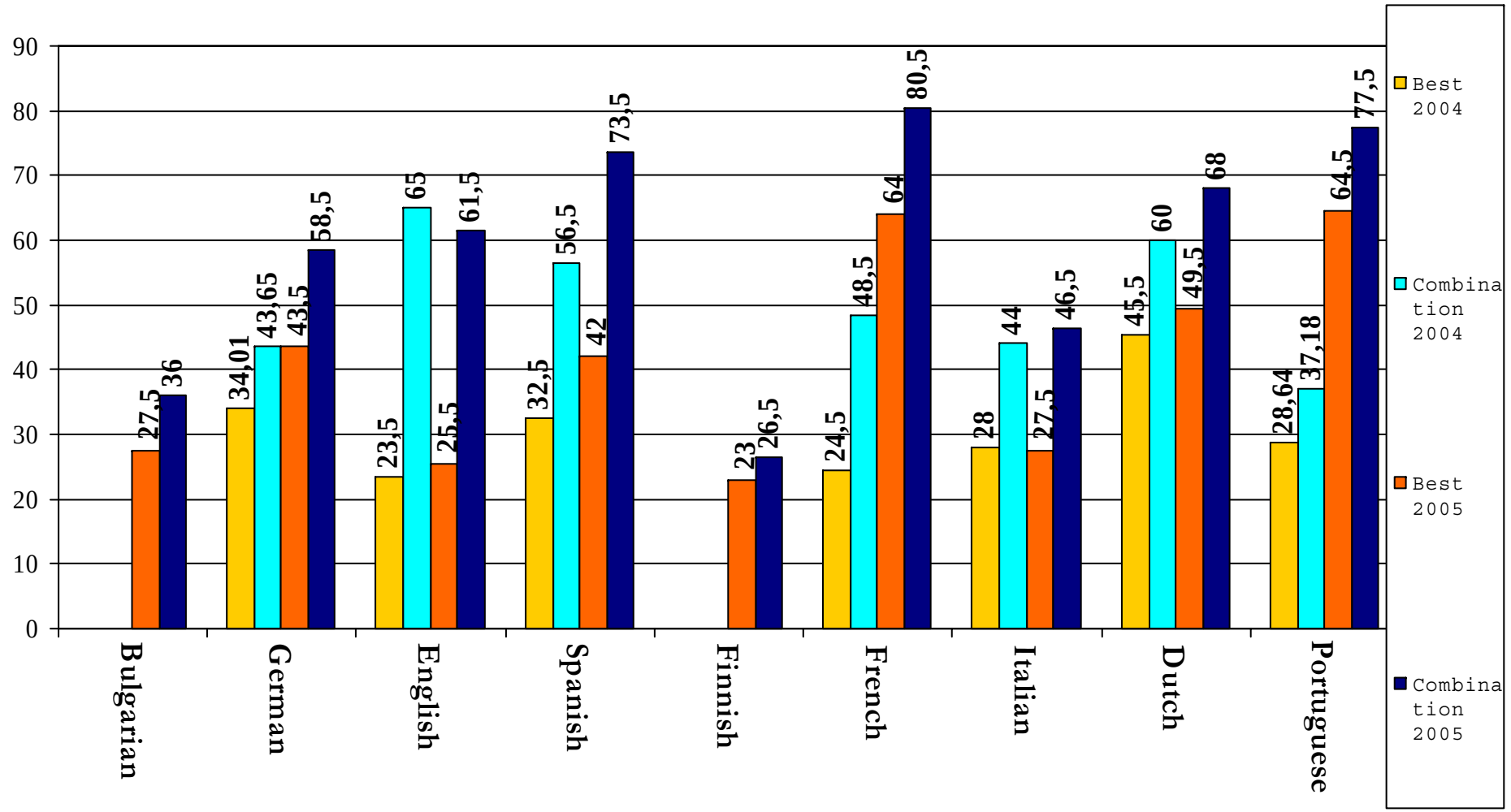
8 Monolingual and 15 bilingual tasks; (6 + 13 in 2004)

Tasks / participant = 1.8; (1.6 in 2004)

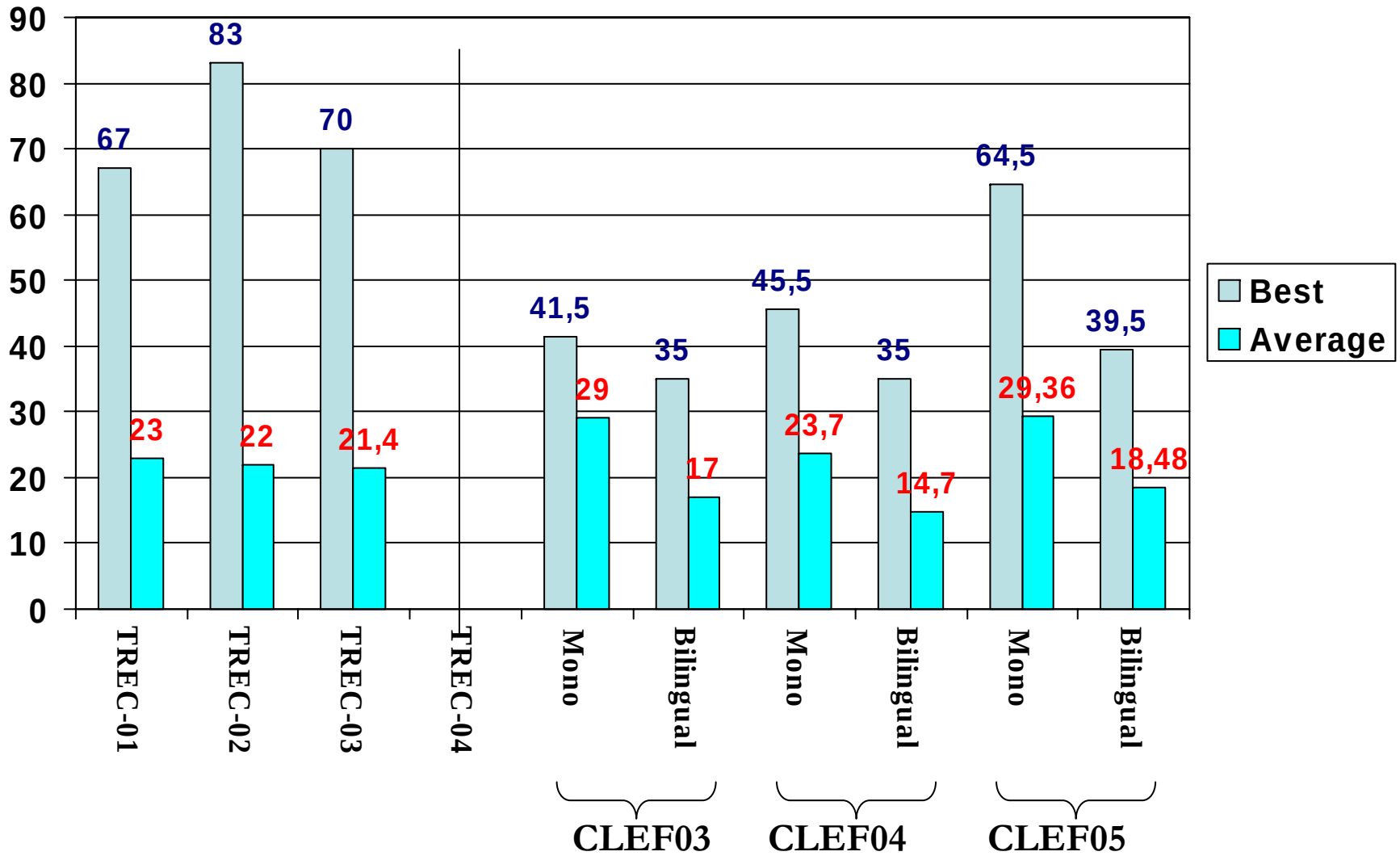
Comparability (tasks > 2 part. / tasks) = 0.2 (0.2 in 2004)

S / T	BG	DE	EN	ES	FI	FR	IN	IT	NL	PT
BG	2		1							
DE		2	1							
EN	1	1		1	1	2	1	1		
ES			2	7				1		
FI					1					
FR			1			7		1		1
IT								3		
NL									2	
PT			1							3

QA-CLEF-05: Results



QA-CLEF-05: Results



QA@CLEF: Approaches I

Linguistic processors and resources are used by most of the systems.

- POS-tagging, Named Entities Recognition, WordNet, Gazzetters, partial parsing (chunking).
- Deep parsing is adopted by many systems
- Semantics (logical representation) is used by few systems

Answer patterns:

- superficial patterns (regular expressions)
- deep (dependency trees): pre-processing the document collection, matching dependency trees, off-line answer pattern retrieval.

QA@CLEF: Approaches II

- Few systems use some form of **“semantic” indexing** based on syntactic information or named entities
- Few systems consult the **Web** at run-time
  - to find answers in specialized portals
  - to validate a candidate answer
- **Architecture** of QA modules: e.g. XML based
- **Cross-language approaches**
  - commercial translators, word by word translation
  - keyword translation

QA@CLEF: Conclusions

Increasing interest in multilingual QA

- More participants (+ 33%)
- Two new languages as target (Bulgarian and Finnish) and one as source (Indonesian)
- More activated tasks (44, they were 29 in 2004, + 51%)
- Monolingual is the prevalent interest (61%)
- Comparability among tasks is 20% (as in 2004)
- Interesting results: 6 systems above 40% accuracy

QA@CLEF 2006

Agenda under discussion:

- New Pilot Task: QA on Wikepedia
- Changes to Main Task
- New Source Languages: Rumanian
- Answer Validation Exercise
- Estimation of Question Difficulty

Participate in the discussion

see http://clef-qa.itc.it/

Evaluation - Summing up

- IR system evaluation is not a competition to find the best
- evaluation provides opportunity to test, tune, and compare approaches in order to improve system performance
- an evaluation campaign creates a community interested in examining the same issues and comparing ideas and experiences
- evaluation creates valuable reusable resources
- evaluation promotes research in new directions

Points for Discussion

- How can we better involve the user?

  (resolve laboratory-based – user-centred dichotomy)

- How can we bridge the gap between research and application communities?

  (evaluation activities should promote Tech. Transfer)

- How can we convince the Commission of the importance of investing in large-scale evaluation initiatives?

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