An Evaluation Infrastructure for HLT in Europe

Khalid CHOUKRI
ELRA/ELDA
55 Rue Brillat-Savarin, F-75013 Paris, France
Tel. +33 1 43 13 33 33 -- Fax. +33 1 43 13 33 30
Email: choukri@elda.org
Web: http://www.elra.info & www.elda.org/
Presentation Outline

1. Quick (over)view of/on ELRA/ELDA Evaluation Activities
2. Evaluation Projects & Campaigns & Packages
3. ELRA Evaluation Services
4. ELRA’s HLT Evaluation Portal
5. ELRA Evaluation Services
6. What’s Next?
Role of ELRA

- Carrying out evaluation campaigns: further involvement at:
  - European (CLEF, TC-STAR, CHIL, etc.)
  - International (CHIL-NIST-AMI Cooperation) and
  - French (Technolangue/Evalda) levels.

- Producing or commissioning the production of needed LRs for evaluation, Validation of the “reference data

- Distributing evaluation packages
Role of ELRA

Permanent infrastructure that focus on:

- R&D on (all) Evaluation issues
- Elaborations of Evaluation protocols, assessment tools,
- Production of Language Resources and Validation
- Coordination team for the management and supervision of projects
- Logistics and support
- Capitalization on the outcome of each and every project (evaluation resources, tools, methodologies, protocols, best-practices AND Results)
Need for a Permanent Infrastructure

• Problem with Call for Proposals mechanism
  – Limited duration (FPs) / Share of cost by participants

• Permanent organization
  – General policy / Strategy / Ethical aspects
  – Label attribution / Quality insurance & control
  – Production of Language Resources (dev,test)
  – Distribution of Language Resources
  – Capitalization on a large number of specialized centers (Evaluation Network)
Levels of Evaluation (ELSE)

- Basic Research Evaluation (validate research direction)
- Technology Evaluation (assessment of solution for well defined problem)
- Usage Evaluation (end-users in the field)
- Impact Evaluation (socio-economic consequences)
- Programme Evaluation (funding agencies)
Examples of Evaluation Campaigns – Capitalization

- Speech & Audio/sound
  - ASR: TC-STAR, CHIL, ESTER
  - TTS: TC-STAR, EVASY
  - Speaker identification (CHIL)
  - Speech 2 Speech Translation
  - Speech Understanding (Media)
  - Acoustic Person tracking
  - Speech activity detection, …..
  - ……..
Examples of Evaluation Campaigns – Capitalization

- Multimodal --- Video – Vision technologies
  - ✔ Face Detection
  - ✔ Visual Person Tracking
  - ✔ Visual Speaker Identification
  - ✔ Head Pose Estimation
  - ✔ Hand Tracking
Examples of Evaluation Campaigns – Capitalization

- Written NLP & Content
  - IR, CLIR, QA, (Amaryllis, EQUER, Shortly CLEF)
  - Text analysers (Grace, EASY)
  - MT (CESTA, TC-STAR)
  - Corpus alignement & processing (Arcade, Arcade-2, Romanseval/Senseval, …)
  - Term & Terminology extraction
  - Summarisation
Evaluation Projects with ELDA

- **Technolangue/Evalda**: the Evalda platform that ELDA coordinates consists of 8 evaluation campaigns with a focus on the spoken and written language technologies for the French language:
  - **ARCADE II**: evaluation of bilingual corpora alignment systems.
  - **CESART**: evaluation of terminology extraction systems.
  - **CESTA**: evaluation of machine translation systems.
  - **EASY**: evaluation of parsers.
  - **ESTER**: evaluation of broadcast news automatic transcribing systems.
  - **EQUER**: evaluation of question answering systems.
  - **EVASY**: evaluation of speech synthesis systems.
  - **MEDIA**: evaluation of in and out-of context dialog systems.
Evaluation Tasks
Illustration via TC-STAR Project

TC-STAR: Speech to speech translation: Packages with Speech recognition, speech translation, and speech synthesis
Development and Testdata, metrics & results. e.g.

2 categories of transcribing and translating tasks

– European Parliament Plenary Sessions: (EPPS): English (En) and Spanish (Es),

– Broadcast News (Voice of America VoA): Mandarin Chinese (Zh) and English (En)
ASR Tasks

✓ 2 Tasks
  – EPPS:
    • English 3 hours (~34 K words)
    • Spanish 3 h (~32 K words)
  – BN
    • Zh : 3 hours of VoA recorded in Dec 1998 (~42 K characters)

✓ 3 Conditions
  – Restricted training condition (ie TC-Star data)
  – Public data condition (ie data available through ELDA and LDC)
  – Open condition (any data before the cutoff date)
SLT Tasks

• 3 Tasks

  – **ASR:** translate automatic transcripts from a combination of ASR engines (ROVERed). No case, no punct.

  **ASR-limsi:** translate output of Limsi’s ASR, with case. (UPC)

  – **Verbatim:** translate manual transcripts, with case, no punct.

  – **Text:** translate Final Text Edition (FTE) documents,

    » with case and punct.
Language Resources

- Development and test resources taken from the same original sets for ASR and SLT

- **EPPS (English, Spanish):** data from European Parliament satellite broadcast, usage rights negotiated
  - ASR: audio files + manual transcripts ~4h
  - SLT: subset of manual transcripts ~25000 words + 25000 words taken from corresponding FTE documents

- **VOA (Mandarin):** original data available at LDC, packaged by ELDA
  - ASR: audio files + resegmented and corrected manual transcripts ~3h
  - SLT: subset of resegmented manual transcripts ~15000 words
Off-the-shelf comparison

• MT system available for the language pairs of TC-STAR

• All acknowledge the good performance of Systran
  • a large number of Languages,
  • used within SdT?
  • Available off-the-shelf

• Use of Systran Premium 5.0 Global Pack (800€),
  • Acquired via Internet
  • No customization
To sum up … Results

• Evaluation packages to be made available
  – ASR
    • EPPS train, dev and test set
    • Scoring packages for En, Es and Zh
  – SLT
    • EPPS English-to-Spanish: EPPS training, development and test (incl. ref. translations) sets
    • EPPS Spanish-to-English: EPPS training, development and test (incl. ref. translations) sets
    • VOA Mandarin-to-English: VOA development and test sets (incl. resegmented transcripts and ref. translations)
  – TTS: See the project web for details
# How to Increase Synergies

## Synergies within EVALDA

<table>
<thead>
<tr>
<th>ARCADE-II</th>
<th>CESART</th>
<th>CESTA</th>
<th>EASY</th>
<th>EQUER</th>
<th>ESTER</th>
<th>EVASY</th>
<th>MEDIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCADE-II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CESART</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CESTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EASY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQUER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESTER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVASY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDIA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Re-use of LRs
How to Increase Synergies

Synergies between EVALDA, TC-STAR, CHIL, …. 

<table>
<thead>
<tr>
<th>TC-STAR</th>
<th>CHIL</th>
<th>Amaryllis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCADE-II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CESART</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CESTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EASY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQUER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESTER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVASY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDIA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Use of same LRsar
Re-use of LRs
Re-use of tools, metrics, etc.
ELRA Evaluation services

• Capitalize on its expertise to extend the evaluation activities towards:
  – new technologies
  – new languages
  – new environments

• Cooperation with a network of specialized centers

• Services on demand
ELRA Evaluation services

- Services on demand

Bring in our expertise
(Observer)

Produce data

Run Evaluation

Turn key
ELRA Evaluation services

- Do it yourself .. Evaluation à la TC-STAR (WP5)..
- Web-service Evaluation
HLT Evaluation Portal
http://www.hlt-evaluation.org/

• Overview
• HLT Evaluations
• Activities by technology
• Activities by geographical region
• Players
• Evaluation resources
• Evaluation Services
• Site map
Questions?

- Is Evaluation a business … a Niche Business?
- Profitability ROI ? 50%