Data and Evaluation: Critical Resources for Research in Knowledge Processing

Edouard Geoffrois
French National Research Agency (ANR/STIC) & French National Defence Procurement Agency (DGA/DS/MRIS)

CHIST-ERA Conference 2011
Cork, Ireland, Sept 6th
Questions

- From data to new knowledge: what do we mean?
- How to evaluate systems and measure progress?
- How to best support progress?
From data to new knowledge

| Explicit code for semantics of data and functions | Partial code for semantics |
| Structured information | Partial code for semantics |
| Analytic function \( o = f(i) \) | Examples from the real world |
| Structured information | Learning |
| Unstructured information | Partial code for semantics |
| Parameteric model \( o = f_M(i) \) | Partial code for semantics |
| New knowledge |

The data express the semantics through an **explicit** code.

The data are **transformed** using an explicit mathematical function (rules, etc.).

**Theoretical** approach (model is the mathematical proof).

**Experimental** approach (model is natural science).

| Trigger keywords: data processing, computing |
| Trigger keywords: intelligent / semantic processing of digital / multimedia content / knowledge |

Examples of domains: **formal languages**, traditional **signal processing**.

Examples of domains: **natural language and speech processing**, **scanned documents**, **image and video processing**, **information fusion**.

---

06/09/2011
Need n°1: Manually annotated data

A task is defined by a representative sample data set
A good model should agree well with the observed data
Data is also important for training models
Example of metric (for speech transcription)

“I would like to go to London tomorrow morning hum”
I will like to go to lone done tomorrow morning

Error rate = \((2+1)/10 = 30\%\)

Error rate = edit distance between an hypothesis and a reference or a set of references
Evaluation data flow

- Corpus provider
- Researchers
- System models
- Human experts
- Reference
- Output
- Comparison
- Measure
- Evaluator
Need n°2: Synchronized evaluations

Data should be shared for the sake of reproducibility
Tests should occur almost simultaneously to avoid bias
Evaluation design should serve the community

→ Evaluation campaigns
Benefits of evaluation

1. Explicit problems
2. Validate new ideas
3. Identify missing science
4. Compare approaches and systems
5. Determine maturity for a given application
6. Facilitate technology transfer
7. Incite innovation
8. Organise the community
9. Support competitiveness
10. Assess public funding efficiency
History

Late 70's  **NATO** Research Study Group on Automatic **Speech Recognition** (ASR) produces a common benchmark database in several languages

Mid 80's  After failure of earlier programs, the **US** (DARPA ans NIST) introduce systematic objective performance measurement in ASR programs

Early 90's  DARPA and NIST extend evaluation to automatic **Textual information processing** (TIPSTER program, then TREC, MUC, DUC, ...) and opens its evaluation campaign to non-US participants

Mid 90's  First **European** program including evaluation (SQALE program on ASR)

Late 90's  First **French** evaluation program on speech and language processing, followed by a larger one in the early 2000's (Technolangle)  First **Japanese** evaluation on information retrieval (NTCIR)

2001  DARPA and NIST extend evaluation to **Machine Translation**

2003  The major European programs on language processing (TC-STAR, CHIL) include evaluation

Mid 2000's  Evaluation methodology gradually extends to **Image processing** (TRECVid, US-EU CLEAR evaluations, French Techno-Vision program, ...)

06/09/2011  E. Geoffrois
Examples of evaluation campaigns today

<table>
<thead>
<tr>
<th>Funding</th>
<th>Organisers</th>
<th>Name</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>DARPA, DoC</td>
<td>NIST</td>
<td>Rich Transcription</td>
<td>Speech transcription</td>
</tr>
<tr>
<td>DARPA, DoC</td>
<td>NIST</td>
<td>Text REtrieval Conference</td>
<td>Documents retrieval</td>
</tr>
<tr>
<td>DARPA, DoC</td>
<td>NIST</td>
<td>OpenMT</td>
<td>Translation</td>
</tr>
<tr>
<td>DoC, ...</td>
<td>NIST, ...</td>
<td>TREC Vid</td>
<td>Video analysis</td>
</tr>
<tr>
<td>DoC, IARPA, FBI</td>
<td>NIST</td>
<td>SRE, LRE</td>
<td>Speaker and language recognition</td>
</tr>
<tr>
<td>DoD</td>
<td>NIST</td>
<td>Text Analysis Conference</td>
<td>Natural language</td>
</tr>
<tr>
<td>NII, NICT, U. Tokyo</td>
<td>NII, NICT, U. Tokyo</td>
<td>NTCIR</td>
<td>Information retrieval</td>
</tr>
<tr>
<td>EU</td>
<td>U. Pisa, Delft, ...</td>
<td>CLEF, MultiMediaEval</td>
<td>Crosslingual, ...</td>
</tr>
<tr>
<td>OSEO</td>
<td>DGA, LNE, IRIT, UJF, LIPN, GREYC</td>
<td>Quaero</td>
<td>Multimedia document processing</td>
</tr>
<tr>
<td>DGA</td>
<td>DGA</td>
<td>RIMES, ICDAR</td>
<td>Handwriting recognition</td>
</tr>
<tr>
<td>Trento</td>
<td>CELCT, ...</td>
<td>Evalita</td>
<td>Natural language</td>
</tr>
</tbody>
</table>
Impact on the evolution of performances (example of spoken language recognition)

LR Performance History 1996 - 2007

Source: NIST

Evolution of the error rate of the best system over the years
Impact on the evolution of performances (example of speech transcription)

Source: NIST

When a problem (one colored curve) is considered as solved, move on to a more difficult one.
The transformative power of evaluation

Before

After
Issues

- Why evaluate?
  - “We did without it until now. Why change?”
  - “It is not a research activity. Why bother?”
  - “It creates additional constraints...”
- How to evaluate?
  - “It works on the examples shown in the demonstration.”
  - “The algorithm is mathematically proven. Isn't that enough?”
  - “We conducted user tests. Isn't that enough?”
  - “There are publications. Isn't that enough?”
- Why so much debate?
  - A relatively young science with an even younger metrology
  - A relatively unknown economic model
Technology evaluation vs. usage studies

Objectives:
- **Technology evaluation**
  - Reproduce results, measure progress, determine maturity

- **Usage studies**
  - Measure user perception, refine the needs

Methods:
- **Evaluation through publications**
  - Interpret results, share knowledge

- **Theoretical**
  - Reproduce results, measure progress, determine maturity

- **Experimental**
  - Interpret results, share knowledge

- **Subjective (user panels)**
  - Reproduce results, measure progress, determine maturity
Technology performance vs. satisfaction of user need

Performance level

Usability threshold for need 2

Usability threshold for need 1
Need for a strong incentive

- A critical component...
  - It represents only a few % of the investments
  - It dramatically increases the return on these investments
- … which must be funded by those who want to see the field make progress as a whole...
  - Campaigns must be organized regularly to measure progress
  - Most of the costs are fixed ones
  - The infrastructure must be open to all to support scientific progress
  - There is no direct return on investment for the party doing the measurements
- … and must be prepared early in project design
  - Data, evaluation and R&D activities are tightly linked and should be jointly designed in integrated projects
Conclusions

- A relatively large but homogeneous domain
  - characterised by the interpretation of data using a model of the world to create new knowledge,
- with a need for manually annotated data
  - representative of the task under study
- and for synchronised evaluations
  - in the form of evaluation campaigns,
- both deserving special attention
  - to really happen and serve the research needs
Thank you for your attention.