How to be Lazily Productive in Evaluation

Olivier Hamon

ELDA

November 27, 2012
Sommaire

1. Our main issue
2. Solutions
3. Context
4. Evaluation Workflow
5. Methodology
6. Language Resources
7. Evaluation measures
8. Conclusions
Our main issue: task repetition

Why "lazily"?
- Evaluation is nice, but can become really repetitive
- Conversion of the repetitive tasks into generic and sustainable ones

Why "productive"?
- Evaluation generally at the end of the workflow or cycle
- ... often upstream delays...
- ... often impatient system developers...

Quick results, but not to the detriment of quality and reliability
Don’t forget the cost! ⇒ data creation, human judgements, workflow management, etc.
DRY: Don’t repeat yourself!

- Principle in software development
- Can be adapted to evaluation

Don’t repeat evaluation but rather reuse:

- The methodology: protocols, workflows, measures...
- The data: in-domain project transfers, system development, cross-domain project transfers...
- The tools: metrics, interfaces, platforms...
Non-sustainable tasks

- Evaluation set up (find data, judges, etc.)
- Quality quantification
- Results analysis and interpretation

Although experience helps!
ELDA and the evaluation

- Organisation and collaborations in many evaluation campaigns
- Large volume of experimentations
- Metric development and result analysis
- First evaluation platform implementations

(Still) Growing need for evaluation

- Well established in National and European projects
- System development
- Often on similar topics and domains!
Evaluation Workflow

Evaluator / user

Methodology

Language Resources

Input Data

Reference(s)

Measure

Protocol

Run

System A

Results analysis

Final interpretation

Parameter
Evaluation Workflow
Methodology

Evaluator / user

System A

Output A

System B

Output B

System C

Output C

Methodology

Language Resources

Input Data

Reference(s)

Measure

Protocol

Human metric

Automatic metric

Score

Validation / Meta-evaluation

Adjustment

Run

Parameters

Results analysis

Final interpretation
Two evaluation cycles

### Preliminary cycle
- Define the protocol
- Set up the evaluation
- Set up tools (and platform...)

### Practical cycle
- Apply the protocol, run the evaluation workflow
- Measure quality
- Analyse and interpret results
Things to think about...

- Availability or state of a LR
- Reuse
- Intellectual Property Rights (IPR)
- Impact of the LRs on the evaluation results
- Reuse of tools/scripts to build new resources
### Reuse of LRs regarding the technology

<table>
<thead>
<tr>
<th>Technology</th>
<th>Monolingual</th>
<th></th>
<th></th>
<th>Multilingual</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Spell checking</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine Trans.</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Terminological Extr.</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information Extr.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic Summar.</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document Index.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information Retr.</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Speech Recog.</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Synth.</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LR life cycle example

French/English parallel corpus on medical domain

- EQueR/EVALDA project: question-answering evaluation
- CESART/EVALDA project: terminological extraction evaluation
- CESTA/EVALDA project: machine translation evaluation
- Could have been used for alignment evaluation...
Measure choice

Two types
- Human
- (Semi-)automatic

Human part
- Always there
  - Human judgements
  - Reference
- Otherwise, systems could integrate the measure
## Evaluation measures

### Small number of measures

<table>
<thead>
<tr>
<th>Technology</th>
<th>Human measures</th>
<th>automatic measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$ values</td>
<td>binary</td>
</tr>
<tr>
<td>Automatic translation</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Question-answering</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Synthesis</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Speech-to-speech translation</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Terminologic extraction</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Information retrieval</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Speech recognition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alignement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic summary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Olivier Hamon (ELDA)
Human measures

Characteristics

- Observations (judgements) made by human (judges)
- Variable subjectivity
- High cost ($, set up, delays)
- Good reliability if judgements are well supervised

Needs

- Set up the evaluation
- Human judgements
- Interpretation and visualization of the results
- Evaluator follow-up
Interface: Question-answering
### Segments

<table>
<thead>
<tr>
<th>ID</th>
<th>System</th>
<th>Document</th>
<th>Segment</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1801</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>19</td>
<td>no</td>
</tr>
<tr>
<td>1802</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>180</td>
<td>no</td>
</tr>
<tr>
<td>1803</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>181</td>
<td>no</td>
</tr>
<tr>
<td>1804</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>183</td>
<td>no</td>
</tr>
<tr>
<td>1805</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>184</td>
<td>no</td>
</tr>
<tr>
<td>1806</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>185</td>
<td>no</td>
</tr>
<tr>
<td>1807</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>186</td>
<td>no</td>
</tr>
<tr>
<td>1808</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>187</td>
<td>no</td>
</tr>
<tr>
<td>1809</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>188</td>
<td>no</td>
</tr>
<tr>
<td>1810</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>189</td>
<td>no</td>
</tr>
<tr>
<td>1811</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>19</td>
<td>no</td>
</tr>
<tr>
<td>1812</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>190</td>
<td>no</td>
</tr>
<tr>
<td>1813</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>191</td>
<td>no</td>
</tr>
<tr>
<td>1814</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>192</td>
<td>no</td>
</tr>
<tr>
<td>1815</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>193</td>
<td>no</td>
</tr>
<tr>
<td>1816</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>194</td>
<td>no</td>
</tr>
<tr>
<td>1817</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>195</td>
<td>no</td>
</tr>
<tr>
<td>1818</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>196</td>
<td>no</td>
</tr>
<tr>
<td>1819</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>197</td>
<td>no</td>
</tr>
<tr>
<td>1820</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>198</td>
<td>no</td>
</tr>
<tr>
<td>1821</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>199</td>
<td>no</td>
</tr>
<tr>
<td>1822</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>1823</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>20</td>
<td>no</td>
</tr>
<tr>
<td>1824</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>200</td>
<td>no</td>
</tr>
<tr>
<td>1825</td>
<td>cesta-run2-hum-en</td>
<td>TEST-CESTA-DOC-12</td>
<td>201</td>
<td>no</td>
</tr>
</tbody>
</table>

Display from **1801** to **1825** (on **10153** evaluations)
Interface : MT (fluency)

Le texte est-il écrit en bon français ?

Et moyennant abonnement annuel 15 dollars par famille, ont pu bénéficier des résidents sont gratuitement des médicaments essentiels et le transport, le transfert à l’hôpital en cas d’urgence.

- Niveau 5 - Français parfait
- Niveau 4
- Niveau 3
- Niveau 2
- Niveau 1 - Français incompréhensible

Évaluations réalisées : 15 / 96
A quel point le sens exprimé dans la traduction de référence est aussi exprimé dans la traduction cible ?

L’UNICEF est la force motrice qui contribuer à édifier un monde arriver des droits des enfants.

- Niveau 5 - Tous le sens
- Niveau 4
- Niveau 3
- Niveau 2
- Niveau 1 - Aucun sens

La traduction de référence est la suivante:

L’UNICEF est l’élément moteur qui aide à construire un monde où les droits de chaque enfant seront réalisés.

Évaluations réalisées : 29 / 96
Interface: Speech-to-Speech translation

Evaluation measures

Human measures
(Semi-)automatic measures

Characteristics

- Comparison with one or several references
- Objectives: replace human judgements
  - when they are not "possible"
  - when they are too costly
- Advantages: execution speed, reproductibility, *objectivity*, *cost*, workflow integration
Intelligence and automation: various ways...

Scripts
- Simple and fast implementation
- Task merging, reproducibility at wish
Evaluation measures

Integration and automation: various ways...

```
mylaptop$ perl 01_CHECK_SUBMISSIONS.PL
[...]
mylaptop$ perl 02_LIST_SUBMISSIONS.PL
[...]
mylaptop$ perl 03_EVALUATION.PL
[...]
mylaptop$ perl 04_BUILD_RESULTS_TABLES.PL
[...]
```
Evaluation platforms

- Evaluation results tracking
- Genericity
- Easy evaluation access
- Need some programming
- Users do (most of) the job
Example at Elda

- Speech-to-Speech translation
  - TC-STAR project
  - UIMA usage
  - 3 technical components, evaluation components
Speech-to-Speech evaluation - TC-STAR
Speech-to-Speech evaluation - TC-STAR
Integration and automation: various ways...

Example at Elda

- Parsing
  - PASSAGE project
  - Open access to the server during the system development cycle
  - Two evaluation campaigns run
Parsinig evaluation within PASSAGE
PASSAGE server

Evaluation server

Submission uploading
- XML validation
- Evaluation

Result queries
- Result tables

Application
- File storage
- Database
- Results storage

Participant
- Participant rights
- Evaluation report

Registration
- Evaluation report
- https / Organisation rights

Organisation
- https / Participant rights
Evaluation measures
Integration and automation: various ways...

**PASSAGE**: Results visualization

![Graphs and data visualizations](image)

---

Olivier Hamon (ELDA)

November 27, 2012
PASSAGE: follow-up for the evaluator

<table>
<thead>
<tr>
<th>#Eval Primaire</th>
<th>Date</th>
<th>Heure</th>
<th>Constituants</th>
<th>Relations</th>
<th>Détails</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>F-mesure</td>
<td>Précision</td>
<td>Rappel</td>
</tr>
<tr>
<td>0</td>
<td>2007-12-22</td>
<td>00:46:44</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2008-01-03</td>
<td>16:07:57</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Résumé de vos évaluations de développement:

<table>
<thead>
<tr>
<th>Evaluation id</th>
<th>Description</th>
<th>Date</th>
<th>Heure</th>
<th>Constituants</th>
<th>Relations</th>
<th>Détails</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2007-11-29</td>
<td>18:23:34</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>2007-12-10</td>
<td>14:15:46</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>2007-12-10</td>
<td>14:25:53</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>2007-12-10</td>
<td>14:38:02</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>2007-12-13</td>
<td>10:17:49</td>
<td>0.0303321</td>
<td>0.552605</td>
<td>0.015594</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>2007-12-13</td>
<td>10:54:51</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>2007-12-13</td>
<td>10:58:25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>2008-01-09</td>
<td>13:11:08</td>
<td>0.887493</td>
<td>0.880327</td>
<td>0.894776</td>
</tr>
</tbody>
</table>
Web services / automatic workflow

- Not that hard to implement
- When available, easy use
- Flexible
- Allows a free access to the tools
Example

- PANACEA project
- Tools available through web services (registry.elda.org)
- Build workflows from the available web services (myexperiment.elda.org)
PANACEA: build a service with an ACD file

Using a Tomcat server (http://tomcat.apache.org)

```
appl: BLEU_Evaluation [
  documentation: "BLEU Evaluation within MEDAR"
  groups: "MEDAR"
  nonemboss: "Y"
  executable: "perl"
]

string: script [
  standard: "Y"
  parameter: "Y"
  default: "/home/olivier/[...]/mteval-v11b.pl"
  comment: "display false"
  comment: defaults
]

boolean: bool_env [
  additional: "Y"
  information: "case sensitive evaluation"
  qualifier: "c"
  default: false
]

infile: reference_file [
  standard: "Y"
  qualifier: "r"
  comment: "data direct"
]

infile: source_file [
  standard: "Y"
  qualifier: "s"
  comment: "data direct"
]

infile: target_file [
  standard: "Y"
  qualifier: "t"
  comment: "data direct"
]

outfile: output [
  additional: "Y"
  default: "stdout"
]
```
PANACEA: run a workflow

Using Taverna (http://www.taverna.org.uk)

(demo)
What shall I implement? It depends on:

- the size of the evaluation (versions of one system, a whole evaluation campaign, etc.)
- the usage (by the evaluator vs by the system developers...)
- the repetitivity of the evaluation (3-year project, a development evaluation once a week, etc.)
- my knowledge
- my available time
Validation and meta-evaluation

Diagram:
- Evaluator / user
- Methodology
- Language Resources
  - Input Data
  - Reference(s)
- Measure
- Protocol
- System A
  - Output A
  - Human metric
  - Automatic metric
  - Score
- System B
  - Output B
- System C
  - Output C
- Validation / Meta-evaluation
- Adjustment
- Results analysis
- Final interpretation

Run
Parameters
Validation and meta-evaluation

Automation

- Human judgements: validation
- Automatic metrics: meta-evaluation
- (Don’t forget to meta-evaluate to check metrics!)
Validation of human measures

- Measure agreement (Kappa coefficient, inter-judge and intra-judge agreements)
- Allow to interpret results
- Identify diverging judges
Meta-evaluation of metrics

**Fiability : relevance of scores**
- Comparison with another *reference* measure
- Correlation coefficient (Pearson, Spearman, Kendall)

**Robustness : production of similar scores for data of similar quality**
- Data samples (*bootstrapping*)
- Difference with the samples’ mean
Conclusions

How to be lazy?
- Maximum reuse of the existing
- Do not reinvent the wheel
- Avoid duplicated tasks/tools (DRY)

How to be productive
- Build as many as possible generic things
- Use fast methodologies and tools
- For new metrics, be creative!
- Do not forget: quality of the results is the final objective