MT EVALUATION & TC-STAR

Khalid CHOUKRI, Olivier Hamon, Djamel Mostefa

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.elda.org/  or http://www.elra.info/
Presentation Outline  Key themes

• Part 1: general views on Evaluation(s)

• Part 2: TC-STAR evaluations approaches & lessons

• Part 3 (1 slide): some (of my) open issues for
TC-STAR evaluations

...... 3 Consecutives annual evaluations

1. SLT in the following directions
   i. Chinese-to-English (Broadcast News)
   ii. Spanish-to-English (European Parliament plenary speeches)
   iii. English-to-Spanish (European Parliament plenary speeches)

2. ASR in the following languages
   i. English (European Parliament plenary speeches)
   ii. Spanish (European Parliament plenary speeches)
   iii. Mandarin Chinese (Broadcast News)

3. TTS in Chinese, English, and Spanish under the following conditions:
   i. Complete system
   ii. Voice conversion intralingual and crosslingual, expressive speech:
   iii. Component evaluation
SLT Tasks

• 3 Inputs
  – **ASR**: translate automatic transcripts from ASR engines (ROVERed), with case and punctuation, **no manual segmentation**
  – **Verbatim**: translate manual transcripts, with case and punctuation
  – **Text**: translate Final Text Edition (FTE) documents, with case and punctuation

• 2 Conditions
  – **Primary**: use single-best hypo from ASR output, use only for training:
    • EPPS: EPPS training set
    • CORTES: Spanish Parliament training set
    • VOA: LDC Large Data
  – **Secondary**: like primary plus ASR word graphs or any other optional input and publicly available data, and use any publicly available data for training
**Verbatim**
I'm I'm I'm starting to know what Frank Sinatra must have felt like.

**ASR output**
and I'm times and starting to know what Frank Sinatra must have felt like.

**Text**
I am starting to know what Frank Sinatra must have felt like,
Resources Production

Development data:

- 2005 development data
- 2005 test data
- 2006 development data
- 2006 test data

<table>
<thead>
<tr>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPPS</td>
<td>manual transcripts (ELDA) taken from ASR test ~25000 words + reference translations (ELDA) 200 kwords (En → Es + Es → En)</td>
</tr>
<tr>
<td>CORTES</td>
<td>manual transcripts (ELDA) taken from ASR test ~25000 words + reference translations (ELDA) 100 kwords (Es → En)</td>
</tr>
<tr>
<td>VOA</td>
<td>manual transcripts (ELDA) of 3h excerpt from ASR test ~25000 words + English reference translations (ELDA) 50 kwords (Zh → En)</td>
</tr>
</tbody>
</table>
Validation procedure for SLT data sets

- Task: assess quality of given translations (references),
- Per set: 1200 contingent words (5%) are selected (from different texts, from source text, except Mandarin; there from target text)
- Two translations per text from different agencies, same samples, checked by a professional translator
- Procedure and criteria are adopted from LDC/NIST
- Max. 40 penalty points per translation allowed
- Validators were unaware of the scoring, only of categories

<table>
<thead>
<tr>
<th>Error</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactical</td>
<td>3</td>
</tr>
<tr>
<td>Lexical</td>
<td>3</td>
</tr>
<tr>
<td>Poor usage</td>
<td>1</td>
</tr>
<tr>
<td>Capitalisation</td>
<td>1</td>
</tr>
<tr>
<td>Punctuation - spelling errors</td>
<td>0.5 (max 10)</td>
</tr>
</tbody>
</table>
## Validation results for SLT data sets (should be < 40)

<table>
<thead>
<tr>
<th>Direction</th>
<th>Data</th>
<th>Task</th>
<th>Agency 1</th>
<th>Agency 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English-to-Spanish</strong></td>
<td><strong>EPPS</strong></td>
<td>FTE</td>
<td>35</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td><strong>EPPS</strong></td>
<td>Verbatim</td>
<td>40</td>
<td>17</td>
</tr>
<tr>
<td><strong>Spanish-to-English</strong></td>
<td><strong>EPPS</strong></td>
<td>FTE</td>
<td>18 R</td>
<td>38 R</td>
</tr>
<tr>
<td></td>
<td><strong>EPPS</strong></td>
<td>Verbatim</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td><strong>PARL</strong></td>
<td>FTE</td>
<td>34 R</td>
<td>35 R</td>
</tr>
<tr>
<td></td>
<td><strong>PARL</strong></td>
<td>Verbatim</td>
<td>26.5 R</td>
<td>22.5 R</td>
</tr>
<tr>
<td><strong>Chinese-to-English</strong></td>
<td><strong>VOA</strong></td>
<td>Verbatim</td>
<td>27 R</td>
<td>37 R</td>
</tr>
</tbody>
</table>
SLT Scoring

• Automatic metrics:
  – BLEU, NIST, IBM(BLEU), mWER, mPER, WNM

• Human evaluation
  – English-to-Spanish direction
  – 100 evaluators (Spanish native speakers, university level education)
  – Around 350 segments by primary system + Softissimo + Systran + one reference translation (45 systems)
  – Evaluation of adequacy and fluency
    • Evaluation of adequacy: the target segment is compared to a reference segment
    • Evaluation of fluency: only the quality of “grammar” is evaluated.
  – Each segment assessed twice by two different judges
SLT Scoring

- Human evaluation
  - 15,835 segments are evaluated, which correspond to 317 segments per evaluator.
  - Re-use of a specific web interface which has already been used for the human evaluation of the French CESTA project.
  - The evaluation is done online
Human Evaluation Interface

El texto está escrito en buen español?

"Es un logro magnífico que deberíamos continuar a celebrar."

Bandeja de evaluación:
- Nivel 5 - Español impecable
- Nivel 4
- Nivel 3
- Nivel 2
- Nivel 1 - Español incomprehensible

Número de evaluaciones realizadas: 12 / 100
SLT Participants

Number of submissions all types of input and all training conditions included:

- **Final Text Edition, Verbatim, ASR**
- **(primary, secondary)**
- **One combined TC-STAR « system » and Softissimo and Systran**

<table>
<thead>
<tr>
<th>Languages</th>
<th>2007</th>
<th>2006</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-&gt;Es</td>
<td>57</td>
<td>40</td>
<td>28</td>
</tr>
<tr>
<td>Es-&gt;En</td>
<td>64</td>
<td>48</td>
<td>38</td>
</tr>
<tr>
<td>Zh-&gt;En</td>
<td>55</td>
<td>33</td>
<td>31</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>176</td>
<td>121</td>
<td>97</td>
</tr>
</tbody>
</table>
SLT Results … Summary

- See details in the MT paper by O. Hamon et. al. (tomorrow)

- Here just a summary to connect with Human evaluation(s)

- Use of **Systran Premium 5.0 Global Pack (800€)**,
  - Acquired via Internet
  - No customization
BLEU/NIST Results – EnEs
(scale: 0-100)
Human Evaluation Translations … EnEs adequacy (1-5)

Combinations

Commercial
Human Eval Results
fluency (0-100)

Diagram showing fluency ratings for various systems, including Human, ROVER, IBM, IRST, LIMSI, RWTH, UKA, UPC, UDS, Reverso, Systran, IBM-2006, IRST-2006, and RWTH-2006, with ratings ranging from 1 to 5.
### Human Eval Results – subset correlation

<table>
<thead>
<tr>
<th></th>
<th>NIST</th>
<th>BLEU</th>
<th>IBM</th>
<th>mWER</th>
<th>mPER</th>
<th>WNM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTE</td>
<td>99.88</td>
<td>99.77</td>
<td>99.76</td>
<td>99.89</td>
<td>99.93</td>
<td>99.72</td>
</tr>
<tr>
<td>Verbatim</td>
<td>99.92</td>
<td>99.89</td>
<td>99.84</td>
<td>99.82</td>
<td>99.90</td>
<td>99.84</td>
</tr>
</tbody>
</table>
SLT Automatic Results … BLEU measures… TC-STAR versus ….
### Automatic Metrics Comparison

<table>
<thead>
<tr>
<th>Metric</th>
<th>En-&gt;Es</th>
<th>Es-&gt;En</th>
<th>Zh-&gt;En</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASR</td>
<td>Text</td>
<td>Verb</td>
</tr>
<tr>
<td><strong>BLEU ↔ mPER</strong></td>
<td>97.93</td>
<td>97.72</td>
<td>98.19</td>
</tr>
<tr>
<td><strong>BLEU ↔ WNM</strong></td>
<td>99.25</td>
<td>98.93</td>
<td>97.01</td>
</tr>
<tr>
<td><strong>IBM ↔ mPER</strong></td>
<td>97.56</td>
<td>97.50</td>
<td>98.06</td>
</tr>
<tr>
<td><strong>IBM ↔ WNM</strong></td>
<td>99.55</td>
<td>99.24</td>
<td>96.87</td>
</tr>
<tr>
<td><strong>mPER ↔ WNM</strong></td>
<td>96.57</td>
<td>96.56</td>
<td>94.41</td>
</tr>
</tbody>
</table>

*up: scoring correlation; down: ranking correlation*
# Correlations of automatic Metrics vs Human Evaluation (EnEs)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>ASR</th>
<th>Text</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU vs. Fluency</td>
<td>98.16</td>
<td>86.68</td>
<td>92.93</td>
</tr>
<tr>
<td>IBM vs. Fluency</td>
<td>98.47</td>
<td>86.71</td>
<td>92.51</td>
</tr>
<tr>
<td>mPER vs. Fluency</td>
<td>94.87</td>
<td>78.1</td>
<td>85.62</td>
</tr>
<tr>
<td>WNM vs. Fluency</td>
<td>98.97</td>
<td>87.85</td>
<td>94.34</td>
</tr>
<tr>
<td>BLEU vs. Adequacy</td>
<td>97.26</td>
<td>84.23</td>
<td>93.83</td>
</tr>
<tr>
<td>IBM vs. Adequacy</td>
<td>97.46</td>
<td>84.13</td>
<td>93.46</td>
</tr>
<tr>
<td>mPER vs. Adequacy</td>
<td>96.57</td>
<td>73.74</td>
<td>87.14</td>
</tr>
<tr>
<td>WNM vs. Adequacy</td>
<td>98.48</td>
<td>81.19</td>
<td>89.36</td>
</tr>
</tbody>
</table>

*up: scoring correlation; down: ranking correlation*
• What are our conclusions?
• For whom?
Improvement of SLT Performances (En→Es)

![Graph showing BLEU scores for different systems and input types.]

- Input = Text, Verbatim, Speech recognition

- Systems: IBM (FTE), IRST (FTE), RWTH (FTE), UPC (FTE), IBM (Verb), IRST (Verb), RWTH (Verb), UPC (Verb), IBM (ASR), IRST (ASR), UPC (ASR)
Improvement of SLT Performances (Es→En)

BLEU [%]

Input = Text, Verbatim, Speech recognition

BLEU - 2005 system  BLEU - 2006 system  BLEU - 2007 system
End-to-End evaluation
End-to-End

- The end-to-end evaluation is carried out for 1 translation direction: English-to-Spanish
- Evaluation of ASR (Rover) + SLT (Rover) + TTS (UPC) system
- Same segments as for SLT human evaluation
- Evaluation tasks:
  - **Adequacy**: comprehension test  ➔ *Very surprising results!*
  - **Fluency**: judgement test with several questions related to fluency and also usability of the system
End-to-End (2/2)

• Test data:
  – Input: audio data: 20 * 3 minutes of speech in English. For each segment:
    • The ASR ROVER output (English).
    • The ASR ROVER output (Spanish).
    • The synthesis (TTS) by UPC (Spanish).
  – The speech from the interpreter (ITP) is collected and evaluated as a “top-line”.
Fluency questionnaire

• [Understanding] Do you think that you have understood the message?
  1: Not at all , ..........5: Yes, absolutely

• [Fluent Speech] Is the speech in good Spanish?
  1: No, it is very bad ...... 5: Yes, it is perfect

• [Effort] Rate the listening effort
  1: Very high .......... 5: Low, as natural speech

• [Overall Quality] Rate the overall quality of this audio sample
  1: Very badm unusable ...... 5: It is very useful
End to End results
(subjective test: 1…5 )

![Bar diagram showing End to End results for ITP and TCSTAR.](chart.png)
End to End results from 2006 tests
The Key Lesson from TC-STAR
Comprehension test: 0-1)
-2 evaluated systems: ITP for the interpreter version and TC-STAR for the automatic speech-to-speech translation system
Hamon Olivier; 16.06.2006

- E2E Evaluation: the evaluation was done by the same assessors who did the subjective evaluation.
Hamon Olivier; 16.06.2006

- ITP / TTS: as it was not foreseen that results would be better for TC-STAR than for ITP, the audio files had been validated to check whether they contained the answers to the questions. The first conclusions that can be drawn from this are: it was difficult for the assessors to find the answers (questions too hard?) and as the interpreter selects and reformulates the information, missing some details, then the question becomes too specific and not appropriate.
Hamon Olivier; 16.06.2006

- TTS, SLT, ASR: in order to determine where the information was lost for the TC-STAR system, files from each component (recognized files for ASR, translated files for SLT, synthesized files for TTS) have been checked. The overall loss is 15% of the information, 5% being lost at each step.
Hamon Olivier; 16.06.2006

- Only ITP: in the end, we used the questions whose answers were included in the interpreter files. So the TC-STAR system lost 10% of the information regarding the ITP evaluation (instead of 15%).
Hamon Olivier; 16.06.2006
TC-STAR Tasks

• More results from the 2007 Campaign

http://www.tc-star.org/

• Evaluation packages available 😊

  E0002 TC-STAR Evaluation Package - ASR English
  E0003 TC-STAR Evaluation Package - ASR Spanish
  E0004 TC-STAR Evaluation Package - ASR Mandarin Chinese
  E0005 TC-STAR Evaluation Package - SLT English-to-Spanish
  E0006 TC-STAR Evaluation Package - SLT Spanish-to-English
  E0007 TC-STAR Evaluation Package - SLT Chinese-to-English
Some open issues/topics

1. "Role of the user in the evaluation process of MT (& Assisted MT); How much difficult taking the user in the evaluation process?"

2. "How to measure MT performance/success: user satisfaction or technology accuracy?" – do they correlate?

3. "How to quantify the success in each situation?

4. How much is it dependent from scenarios and context (application)?"

5. "Are the best (performance) systems the most successful commercially?" –

6. "How useful the evaluation is? Pushing a head the knowledge or killing the innovation?"

7. ..................
Thank you very much for your attention

Slides will be made available