Language technologies for Education: recent results by the MLLP group

Alfons Juan
2nd Internet of Education Conference 2015
18 September 2015, Sarajevo
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The MLLP research group</td>
<td>2</td>
</tr>
<tr>
<td>transLectures</td>
<td>3</td>
</tr>
<tr>
<td>EMMA</td>
<td>11</td>
</tr>
</tbody>
</table>
• Research group at the Univ. Politècnica de València (Spain)

• Research areas:
  – Machine Learning and Applications
  – Natural Language Processing
  – Educational Technologies and Big Data

• Recent research projects:
  – transLectures: Transcription and translation of video lectures
  – EMMA: European Multiple MOOC Aggregator
  – Active Interaction for Speech Transcription and Translation
transLectures (Nov 2011 – Oct 2014)

Main goal: to develop innovative, cost-effective (quasi-automatic) solutions to produce accurate subtitles for educational videos

Two pilots:

- **VideoLectures.NET:** En, Sl → {De, Sl, Fr, Es} Sl → En
- **poliMedia:** Es, Ca ↔ En Ca ↔ {Es, En}

Three scientific and technological objectives:

- **Massive adaptation** to improve subtitling quality
- **Intelligent interaction** to improve subtitling quality
- **Integration into Opencast** to enable real-life evaluation
transLectures: VideoLectures.NET

>20000 videos (45 min. on avg.): 85% English, 13% Slovenian, ...
transLectures: poliMedia

27000 videos (2-10 min.): 88% Spanish, 3% Catalan, ...
Key advances: neural networks and “extra” resources (slides)
ASR: all error rates below 30% (Sl: 26.9%, others below 20%)
SMT: all BLEU scores above 20 (En→Sl: 20.1, mostly above 25)
transLectures: Intelligent interaction

The 10% of words recognized with minimum confidence include 30% of the actual word recognition errors
transLectures: Integration into Opencast

Opencast was designed to be a flexible and customizable video capture and distribution option for modern institutions. Opencast is built by a growing community of developers in collaboration with leading universities and organizations worldwide.

Schedule
Schedule events to automatically record based on a pre-defined timetable and, capture both video of the presenter and the PC screen.

Edit
Bulk edit and trim video recordings. The editor provides graphical visualization of elements such as audio can significantly reduce editing time.

Process
A scalable infrastructure to encode video, generate metadata and preview images, create captioning and bulk edit to support video processing at scale.

>300 different institutions worldwide
transLectures: try our tools

Try our tools at  ttp.mllp.upv.es

190 users and 1197 videos (264 hours) since May 2014
transLectures: try our tools (cont.)

<table>
<thead>
<tr>
<th>Institution</th>
<th>Country</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univ. Politècnica de València</td>
<td>Spain</td>
<td>ES, CA, EN</td>
</tr>
<tr>
<td>Univ. Carlos III de Madrid</td>
<td>Spain</td>
<td>ES, EN</td>
</tr>
<tr>
<td>Universidade Aberta</td>
<td>Portugal</td>
<td>PT, EN</td>
</tr>
<tr>
<td>University of Naples Federico II</td>
<td>Italy</td>
<td>IT, EN</td>
</tr>
<tr>
<td>Open Universiteit Nederland</td>
<td>Netherlands</td>
<td>NL, EN</td>
</tr>
<tr>
<td>University of Leicester</td>
<td>UK</td>
<td>EN, ES</td>
</tr>
</tbody>
</table>

**Examples:**

- Lecture recorded at poliMedia studios. [link]
- Text-to-Speech demo [link]
**EMMA** (Feb 2014 – Jul 2016)

**European Multiple MOOC Aggregator (EMMA)**

**Main goal:** to provide multilingual access to European MOOCs

**Motivation:**

- Most MOOCs are offered in few languages (En, Es, Fr)
- Language barrier is keeping many learners from taking MOOCs
- MOOC components: texts, images, videos, forums
- **EMMA** uses transLectures tools to translate videos and texts
  - Few hours of video in 7 En, Es, It, Nl, Et, Pt and Fr
  - Source language is the national language of the MOOC provider
  - Target languages: En, Es and It
EMMA: cost of manually translating MOOCs

Texts:
- Manual translation rate is approximately 2500 words per day
- A 6-week course with 75000 words takes 1.5 PM

Videos:
- Before translating, videos are manually transcribed (10 RTF)
- Then, transcriptions are translated (30 RTF)
- A course including 2 hours of video takes 0.5 PM

Solutions to lower costs:
- Crowdsourcing (e.g. TED talks)
- ASR and MT: user effort is reduced to 30% (2 → 0.6 PM)
EMMA: automatic video subtitling

1. Generation of **automatic transcriptions** from video

2. **Manual review** of automatic transcriptions

3. Generation of **automatic translations** from transcriptions

4. **Manual review** of automatic translations
EMMA: automatic document translation

• Course text is ingested into the translation system
• Source and target texts are reviewed in parallel
• Preview of source and target texts also available
• Translated text is imported back into the EMMA platform
## Video subtitling

<table>
<thead>
<tr>
<th>Language pairs</th>
<th>Transcription RTF (10)</th>
<th>Translation RTF (30)</th>
<th>Total RTF (40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish → English</td>
<td>3</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>English → Spanish</td>
<td>6</td>
<td>17</td>
<td>23</td>
</tr>
</tbody>
</table>

## Document translation

<table>
<thead>
<tr>
<th>Language pairs</th>
<th>Translation RTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish → English</td>
<td>7</td>
</tr>
<tr>
<td>English → Spanish</td>
<td>17</td>
</tr>
</tbody>
</table>
EMMA: conclusions

• Multilingual access to your course boosts visibility

• The cost of manually translating your course is high (2 PM)

• Automatic translation can reduce the temporal cost to 30%

• Accuracy of automatic translation depends on several factors:
  – Languages involved
  – Availability of annotated data resources related to your course
  – Specificity of the course

• Designing a multilingual MOOC should also take into account: slides, images, application interfaces (demos), bibliography