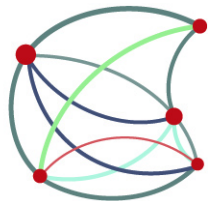LISTEN Workshop / Summer School

Bonn, Germany, July 17-19, 2018

Multilingual videos for education

Alfons Juan

MLLP | Machine Learning and Language Processing

Universitat Politècnica de València, València, Spain

Contents

MLLP

1   Introduction

- Multilingual subtitles are needed for MOOCs and OER

- Automatic (ASR&MT) subtitles are often good enough

- If not, post-edition of automatic subtitles is far more efficient than generating them ex novo

- Real-life results are reported for a MOOC platform (EMMA FP7 project, 2014–2016) and an OER repository (transLectures FP7 project, 2011–2014), that have been just published in:

J. D. Valor, P. Baquero, J. Civera, C. Turro and A. Juan
Multilingual videos for MOOCs and OER
Journal of Educational Technology & Society, 21 (2), 2018.

MLLP

2 Case studies

- **_The EMMA platform_**

  – Developed in the FP7 project EMMA (2014–2016)

  – 12 partners delivering more than 30 multilingual MOOCs

- **_The UPV media repository_**

  – UPV's repository of educational video lectures

  – Studied in the FP7 project transLectures (2011–2014)

  – Basic statistics:

Language	Videos	Hours	Lecturers
Spanish	15013	2709	1572
English	1221	173	203
Catalan	434	52	80

MLLP

3 Systems, tools and integration components

- transLectures-UPV toolkit (TLK) for ASR (similar to Kaldi/RASR)

- Moses toolkit for MT (still state-of-the-art just 3 years ago)

- Language model adaptation to in-domain data

- transLectures-UPV Platform (TLP) for subtitling post-edition:

MLLP

4 Transcription and translation quality

Transcription quality

Language	WER
Spanish	18.4
Italian	25.7
English	21.9
Dutch	29.4
French	23.2
Average:	*23.7*

Translation quality

Tl. pair	TER
Es → En	33.2
En → Es	27.0
It → En	37.5
En → It	33.8
Nl → En	30.7
Fr → En	58.9
Average:	*36.9*

Language	WER EMMA	WER Mains.
Spanish	14.8	22.5
Italian	17.1	31.6
English	39.2	65.9
Dutch	24.5	41.1
French	20.6	32.0
Average:	*23.2*	*38.6*

Tl. pair	TER EMMA	TER Mains.
Es→En	33.9	44.3
En→Es	35.8	42.4
It→En	33.4	39.2
En→It	39.7	43.3
Nl→En	42.5	45.0
Fr→En	52.8	52.6
Average:	*39.7*	*44.5*

5   Reviewing time

Transcription: RTF vs WER
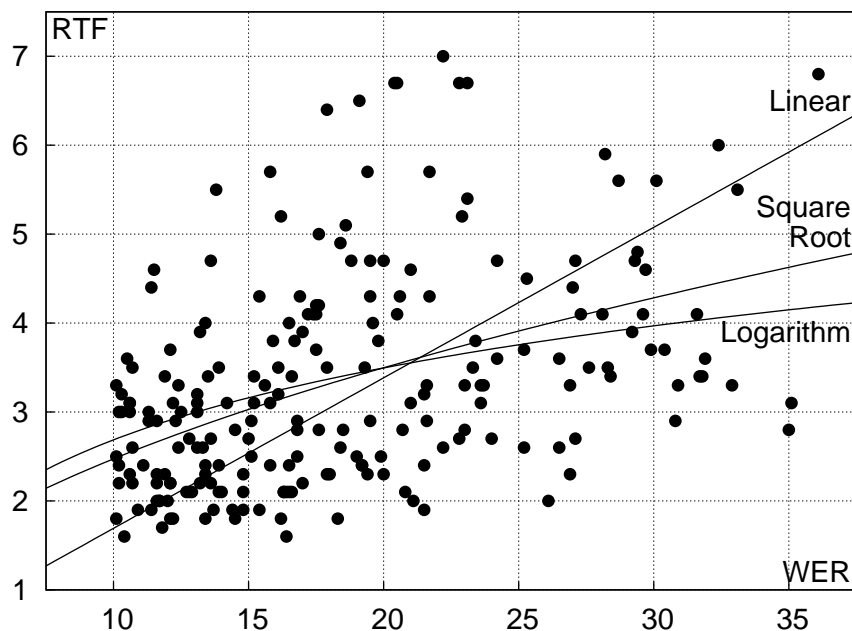
Language	WER	RTF
Spanish	18.4	3.3
Italian	25.7	3.9
English	21.9	5.3
Dutch	29.4	5.8
French	23.2	6.7
Ex novo		10.0

Translation: RTF vs TER
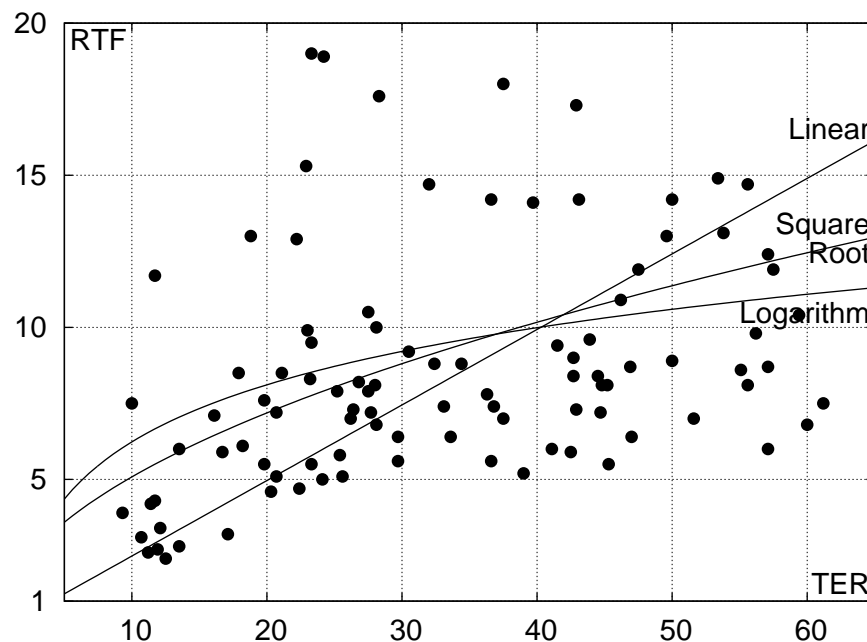
Tl. pair	TER	RTF
Es→En	33.2	9.1
En→Es	27.0	7.8
It→En	37.5	11.3
En→It	33.8	9.6
Nl→En	30.7	9.5
Fr→En	58.9	23.2
Ex novo		30.0

MLLP

Reviewing time across languages

Transcription: RTF vs WER

Translation: RTF vs TER
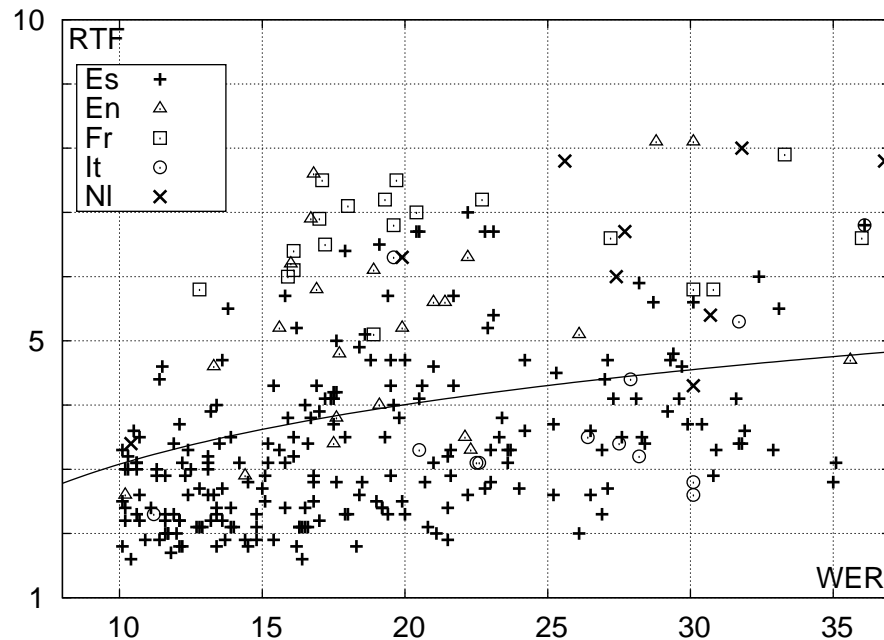
Log fit stat. significant

For 1h at 10% WER: 3h

Log fit stat. significant

For 1h at 30% TER: 9h

Time saving: 70% independently of the lang. (pair)

MLLP

6   Impact on the case studies

The EMMA platform

(Non-)native enrollments:

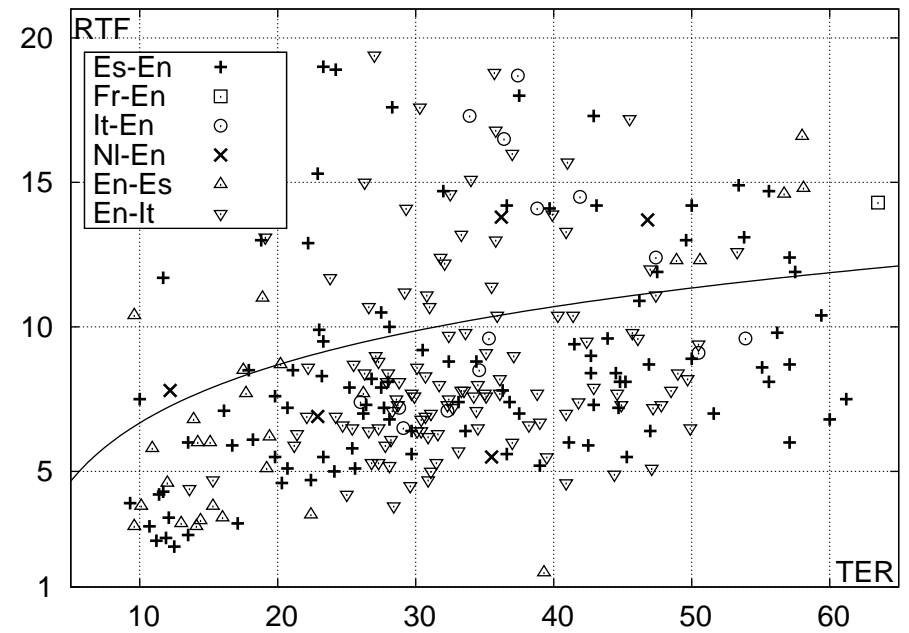| Lang.	Nat.	Non	△%
Spanish	161	547	340
French	983	879	89
Italian	609	259	43
Dutch	501	104	21
English	351	27	8
Total	2605	1816	70

The UPV media repository

Video/subtitle Kviews (15/16):

Video lang.	Video views	Sub. views Es	En
Spanish	629	6.9	1.1
English	63	1.3	0.5
Total	692	8.2	1.6

Enrollment increase: 70%

$$\frac{1816}{2605}$$

Sub. views increase: 24%

$$\frac{1.1+1.3}{6.9+1.1+1.3+0.5}$$

MLLP

7   Conclusions and future work

- Reported experience (2011–2016) on using ASR and MT to produce (automatic) multilingual subtitles for educational videos.

- Automatic subtitles are often good enough to be directly used.

- If not, post-edition of automatic subtitles is far more efficient than generating them ex novo: 70% of time saving on average.

- Impact on case studies was also important: 70% of enrollment increase in the EMMA platform and 24% of subtitle views increase in an OER repository.

- Future work: multilingual, on-line (distant) ASR (+MT+TTS).

MLLP

Future work: 2017 update

Language	Data (hours) 2016	Data (hours) 2017	Word Error Rate 2016	Word Error Rate 2017	△	RTF 2016	RTF 2017
Spanish	200	800	15.0	11.7	− 22%	1.5	0.8
Catalan	50	2430	17.8	16.6	− 7%	1.2	0.7
English	430	2550	26.7	22.2	− 17%	2.0	1.5
German	300	880	27.0	22.5	− 17%	1.2	1.0
French	80	670	21.3	16.5	− 23%	2.1	1.0
Italian	130	740	27.8	25.4	− 9%	0.7	0.7
Dutch	300	620	27.5	26.7	− 3%	1.5	0.9
Portuguese	120	530	34.4	23.3	− 32%	1.0	1.0
Slovene	90	90	38.6	33.9	− 12%	1.9	1.1
Average	190	1030	26.2	22.1	− 16%	1.4	0.9

Follow-up: H2020 project X5gon from 2017 to 2020

MLLP

Acknowledgements

- ***transLectures:*** FP7 ICT grant 287755, 2011–2014

- ***EMMA:*** FP7 ICT grant 621030, 2014–2016

- ***MORE:*** MINECO/FEDER TIN2015-68326-R, 2016–2018

- ***X5gon:*** H2020 ICT grant 761758, 2017–2020

Thank you very much for your attention!