X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number:	761758
Project Acronym:	X5GON
Project title:	Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date:	2017-09-01 to 2020-08-31
Project Duration:	36 months
Deliverable Title:	D3.5 – Final support for cross-lingual OER
Lead beneficiary:	UPV
Type:	Report
Dissemination level:	Public
Due Date (in months):	30 (February 2020)
Date:	
Status (Draft/Final):	Draft
Contact persons:	J. Jorge and A. Juan

Revision

Date	Lead author(s)	Commments
21-Feb-2020	J. Jorge, A. Pérez, G. Garcés, P. Baquero, J. Iranzo, J. A. Silvestre, A. Sanchis, J. Civera and A. Juan	first draft
25-Feb-2020	Erik Novak	Some comments added

Contents

List of Tables

Abstract

The main objective of WP3 is the construction of the analytics engine that provides the relevant knowledge required to drive the operation of the OER and social network. This includes cross-lingual issues in Task 3.3 (M12–M30), whose main goal is to extend the analytics engine with capabilities to deal with multi-lingual collections of OER. D3.5 is to report the work done in Task 3.3 from M25 (September 2019) to M30 (February 2020). During this period, the main goal of Task 3.3 is to provide the final support for cross-lingual OER.

1 Introduction

The main objective of WP3 is the construction of the analytics engine that provides the relevant knowledge required to drive the operation of the OER and social network. This includes analysis of learning and testing, cross-lingual aspects, links with educational theories, affective computing, etc. In addition, there are two important aspects that are studied over the duration of the project:

1. Fine-grained indexation of educational videos by transcription tools; and
2. Investigation of multicultural, pedagogical and juridical issues, with particular care on privacy.

The first of these two aspects, and cross-lingual issues in general, are covered in Task 3.3 from M12 (August 2018) to M30 (February 2020). In this regard, the work done until M24 was described in deliverable D3.4 – Early support for cross-lingual OER [1]. This deliverable, D3.5 – Final support for cross-lingual OER, is to report the work done in Task 3.3 from M25 (September 2019) to M30 (February 2020); i.e. during the first half of Year 3 (Y3).

Briefly speaking, we have improved our ASR (Automatic Speech Recognition) and MT (Machine Translation) systems for automatic transcription and translation of OER in the dominant languages of the official pilots. This includes better automatic transcription of Slovenian and Spanish OER, and also better automatic translation of OER for the language pairs {Spanish, French, Italian} $\leftrightarrow$ English. Section 2 covers the progress made in OER transcription. In it, we first provide a brief summary of the work done until M24; then, the work done from M25 to M30 is described; and lastly, as a summary on the final ASR support for cross-lingual OER, we conclude the section with a global summary of results and comparison to Google. This same structure is followed in Section 3 for OER translation. Finally, important concluding remarks and open opportunities are discussed in Section 4.

2 OER transcription

2.1 Summary of work done until M24

As said in the introduction, the work done until M24 on ASR for OER was reported in deliverable D3.4 [1]. It started during the last months of Y1, while developing X5gon-TTP, into which we integrated the ASR systems from the UPV background for English, Spanish and Slovenian. At that time, we also invested some effort to improve the Slovenian ASR system, as well as to build a new German ASR system using state-of-the-art technology and having virtUOS, the official pilot from UOS, in mind. Then, in Y2, our work was centered on improving our ASR systems for English and Slovenian, that is, the dominant languages in the main official pilot, VideoLectures.NET (VL).

For convenience, Table 1 shows a brief summary of results achieved until M24, in terms of WER (Word Error Rate) for English, Spanish, Slovenian and German (lower WER score means better results). Most of them correspond to in-domain tasks from the official pilots: VL (video lectures in English and Slovenian, from JSI), pM (poliMedia recordings in Spanish, from UPV); and scast and class (German screencast and classroom recordings, respectively, from UOS). Also, Table 1 includes

results from two additional, out-domain tasks that are widely used for comparison purposes by the ASR research community: TED (in Slovenian) and IWSLT (in English).

Table 1: WER scores provided by X5gon ASR systems until M24 for different in-domain and out-domain tasks in English, Spanish, Slovenian and German.

	VL	IWSLT	pM	VL	TED	scast	class
Month	En	En	Es	Sl	Sl	De	De
M0	19.6	13.2	11.0	32.5	27.6	-	-
M12	-	-	-	31.4	26.3	38.1	28.7
M18	-	-	-	30.1	24.3	-	-
M24	18.8	7.2	-	25.3	20.0	-	-

As can be observed in Table 1, until M24 we made good progress in supporting ASR for OER in our official pilots. Based on our prior research experience in ASR and MT for OER, WER scores that are around 20% or below are in general of publishable quality and, not least, suitable enough to try MT from them. Then, if we look at the best figures in Table 1 for each language, we may say that English, Spanish and Slovenian were already well supported by the end of Y2.

Although the ASR results for German are not as good as those for the other languages under study, two important remarks are in order. On the one hand, as discussed in deliverable D5.1 [2, p13], the scast test set consists of only two single-speaker videos of poor audio quality, and thus we tend to think that no significant conclusions can be drawn from it. On the other hand, the WER score of 28.7% achieved in the class task is in line with the WER figures we obtained in recent, on-going experiments on similar (live) class recordings (in Spanish) at UPV. Therefore, our impression is that the class task is comparatively more difficult than the tasks considered in languages other than German. Notwithstanding, while a WER score of 28.7% might be a little high for automatic multilingual subtitling, it is for sure good enough for less ambitious purposes, such as building cross-lingual recommendations on the basis of automatic transcriptions and translations. This is precisely the main interest of UOS and, indeed, the main topic UOS has been exploring within the piloting work package (see deliverables D5.1 [2] ,and D5.2 [3] for more details).

2.2 Work done from M25 to M30

In the period from M25 to M30, we decided to work on improving our ASR systems for Spanish and Slovenian. On the one hand, the reason behind the decision to choose Spanish was clear: although we already had an impressive WER of 11% on the pM task, the Spanish ASR system we were using was still the one we brought into production at the very beginning of X5gon, from UPV background. Thus, as we did for the two official pilots other than UPV, we tried to further improve transcription accuracy for the Spanish video lectures in the UPV's poliMedia repository.

On the other hand, the Slovenian case is different. Being a minority yet official EU language, we think that X5gon can really help in opening up Slovenian OER by means of advanced cross-lingual tools. In this regard, and from the results in Table 1 for VL-Sl, it is our impression that we did a good job for Slovenian ASR, though still not good enough to reach the "safe" area of WER scores (below 20%). With this aim in mind, we also decided trying to further improve our results for Slovenian.

Both the Spanish and Slovenian systems were updated to the latest ASR developments in a very similar way. The technical details on what we did precisely are reported in annexes A and B. In brief, the major differences from our previous Slovenian and Spanish ASR systems are:

- Speech data were processed into 85-D filter-bank vectors (instead of 48-D MFCCs).
- DNNs for acoustic modeling were replaced by Bidirectional LSTMs.

- New LSTM-based language models were built and interpolated with conventional count-based (4-gram) language models.
- Our prior multiple-pass recognition strategy was substituted by our brand new one-pass decoding approach. In it, on-the-fly LM interpolation allows us to deliver real-time transcription with (almost) no degradation to accuracy.
- As usual when updating ASR systems, more speech and text data were used for acoustic and language modeling. Often, this requires complex re-processing procedures to be applied to previous and new, noisy data.

Table 2 shows WER scores provided by the X5gon ASR systems updated from M25 to M30. It also shows the best WER scores we got until M24 and their relative improvement ($\Delta\%$). Apart from pM in Spanish, and VL and TED in Slovenian, we have included the RTVE task in Spanish, from the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge [4]. In it, which is about recognition of different TV shows broadcasted by the main Spanish public TV station between 2015 and 2018, we got the best WER score (20.0%) using only training data provided by the organization.

Table 2: WER scores provided by the ASR systems updated from M25 to M30. Also shown are the best WER scores achieved until M24, and their relative improvement ($\Delta\%$), on two Spanish and two Slovenian tasks.

	pM	RTVE	VL	TED
Month	Es	Es	Sl	Sl
M24	11.0	20.0	25.3	20.0
M30	9.1	13.0	22.0	18.3
$\Delta\%$	-17.2	-35.0	-13.0	-8.5

From the results in Table 2, we may conclude that the work done on ASR from M25 to M30 has led to large improvements on OER transcription accuracy for Spanish and Slovenian. In the case of Spanish, the WER of 9% achieved in pM is really impressive and, as confirmed by the also excellent WER of 13% on a very different kind of speech data (RTVE), it is *not* due to over-training on pM. In the case of Slovenian, now we are more comfortable with the WER scores achieved; i.e., we managed to cross into the "safe" area ($\leq 20\%$) for TED-Sl, and also more than halved the gap to it for VL-Sl. And, in both cases, we are now ready to transcribe large volumes of X5gon OER content much more accurately and efficiently than in the past.

2.3 Summary of results and comparison to Google Cloud Speech-To-Text

Table 3 shows a summary of ASR results in terms of WER scores for automatic OER transcription on the in-domain and out-domain tasks we have considered from M12 to M30. In it, the best scores from X5gon ASR systems are compared to those delivered by the Google Cloud Speech-To-Text web service [5]. The only exception is the IWSLT-En task, for which we do not have the corresponding score from the Google service, and thus it is not taken into account for the computation of the average scores given in the rightmost column.

From the results in Table 3, we can conclude that X5gon support for OER transcription is excellent. In general, with an average WER score of 21.1%, our ASR systems provide a relative error reduction of 40% with respect to the performance offered by the renowned Google Cloud Speech-To-Text web service. Moreover, this figure is even higher in the in-domain tasks for Spanish (54% for pM-Es) and Slovenian (56% for VL-Sl).

Table 3: WER scores provided by X5gon ASR systems and Google Cloud Speech-To-Text, on different in-domain and out-domain tasks in English, Spanish, Slovenian and German.

ASR	VL En	IWSLT En	pM Es	RTVE Es	VL Sl	TED Sl	scast De	class De	Avg.
Google	28.6	-	19.9	49.3	50.0	38.1	38.9	22.7	35.4
X5gon	18.8	7.2	9.1	13.0	22.0	18.3	38.1	28.7	21.1
Δ%	-34.3	-	-54.3	-73.6	-56.0	-52.0	-2.1	26.4	-40.4

3 OER translation

3.1 Summary of work done until M24

The work done in Task 3.3 on MT for OER started in M12 and, as indicated in the introduction, until M24 was covered in deliverable D3.4 [1]. A main goal was to end Y2 with full MT support for any pair of languages relevant to X5gon, maybe using English as a pivot language. And this goal was accomplished by developing state-of-the-art, Transformer-based [6] Neural MT (NMT) systems for the language pairs: {German, Spanish, French, Italian, Slovenian}↔English, German↔French and Portuguese↔Spanish. All of these NMT systems were assessed in terms of BLEU scores (higher score means better results) on the in-domain tasks poliMedia (pM) and VideoLectures.NET (VL), and also on the well-known (out-domain) tasks WMT and IWSLT (WMT, for short, in what follows) [7, 8]. The reader is referred to Figure 1(b) of deliverable D3.4 for a detailed, easy to read summary of results. In brief, many of the systems developed showed BLEU scores clearly above 35, or just below 35, which is a common reference for experts to consider them good enough for practical use. Moreover, the X5gon MT systems in M24 were found to be truly competitive with the very accurate NMT systems Google Translate [9] is using recently, with the exception of Italian↔English, for which Google Translate was clearly ahead of X5gon.

3.2 Work done from M25 to M30

The work done in the first half of Y3 has focused on improving X5gon MT systems for {Italian, Spanish, French}↔English translations. To this end, our training pipeline was improved by updating it with the Paracrawl dataset (to its latest release), and also the data filtering procedure we were using until M24 [10]. Following [11], the new data filtering procedure performs language identification after filtering in length the whole corpus. Also different to what we were doing until M24 is that model training is now done with a batch size of 16000 tokens (2000 tokens per GPU), using gradient accumulation when convenient.

Table 4 shows BLEU scores provided by X5gon MT systems in M24 and M30 on pM, VL and WMT tasks, for each language pair considered from M25 to M30. At this point, it is worth to mention that we have updated the BLEU score on pM Es-En (30.0) with respect to that in deliverable D5.2 [3], where a better figure was reported due to a tokenization error in the input pipeline.

From the results in Table 4, it is clear that all M30 systems have outperformed their M24 counterparts, especially in the case of Italian→English and English→Italian, with relative improvements of 36% and 28%, respectively. Generally speaking, most BLEU scores are in the mid-30s to mid-40s, and thus they are certainly of practical use in X5gon.

3.3 Summary of results and comparison to Google Translate

To summarize MT results, and for comparative reference, Table 5 shows BLEU scores provided by Google Translate and X5gon MT systems, on the VL and WMT tasks, for each language pair con-

Table 4: BLEU scores provided by X5gon MT systems in M24 and M30 on pM, VL and WMT tasks, for each language pair considered from M25 to M30. Relative improvements from M24 to M30 (Δ%) are also given for each language pair and task, and on average for each language pair (Avg. Δ%).

Lang.	pM			VL			WMT			Avg.
pair	M24	M30	Δ%	M24	M30	Δ%	M24	M30	Δ%	Δ%
Es-En	30.0	34.1	13.7	36.4	40.3	10.7	32.3	35.9	11.1	11.8
En-Es	-	-	-	39.4	44.4	12.7	32.2	34.6	7.4	10.1
Fr-En	-	-	-	29.0	30.1	3.8	36.8	39.7	7.9	5.9
En-Fr	-	-	-	26.2	29.2	11.5	37.9	41.1	8.4	10.0
It-En	-	-	-	-	-	-	25.9	35.2	35.9	35.9
En-It	-	-	-	-	-	-	23.3	29.8	27.9	27.9

sidered. Note that rows are sorted by average relative improvement for the X5gon MT systems over Google Translate (Avg. Δ%). Table 5 can be considered an update of Table 35 in deliverable D3.4, also including the latest MT results achieved from M25 to M30.

Table 5: BLEU scores provided by Google Translate and X5gon MT systems, on the VL and WMT tasks, for each language pair considered. Rows are sorted by average relative improvement for the X5gon MT systems over Google Translate (Avg. Δ%).

Lang.	VL			WMT			Avg.
pair	Google	X5gon	Δ%	Google	X5gon	Δ%	Δ%
Es-Pt	-	-	-	43.4	**70.7**	62.9	62.9
Pt-Es	-	-	-	47.6	**72.4**	52.1	52.1
Sl-En	15.0	**26.4**	76.0	29.2	**34.3**	17.4	46.7
En-Sl	16.5	**22.9**	38.8	23.6	**29.4**	24.6	31.7
De-En	25.7	**27.0**	5.1	43.9	**48.0**	9.4	7.3
Es-En	37.8	**40.3**	6.6	34.4	**35.9**	-2.0	2.3
En-Es	41.3	**44.4**	7.5	**35.3**	34.6	-2.0	2.8
Fr-En	**30.3**	30.1	-0.7	38.6	**39.7**	2.8	1.1
De-Fr	**19.6**	18.6	-5.1	32.2	**34.4**	6.8	0.9
En-Fr	**29.4**	29.2	-0.7	40.4	**41.1**	1.7	0.5
It-En	-	-	-	**35.7**	35.2	-1.4	-1.4
Fr-De	**18.6**	17.2	-7.5	26.6	**26.9**	1.1	-3.2
En-It	-	-	-	**32.1**	29.8	-7.2	-7.2
En-De	**24.7**	21.5	-13.0	**47.0**	45.7	-2.8	-7.9

As shown in Table 5, X5gon MT systems for Spanish↔Portuguese and Slovenian↔English provide far better results than those provided by Google Translate. On the other hand, the results for all other language pairs are more or less on par with those by Google Translate.

Clearly, the progress made in Y2, now completed during the first half of Y3, has allowed us to catch up with the high quality Google Translate is delivering in recent years. Moreover, it is also clear to us that it is certainly possible to go far beyond Google Translate's quality by *adapting* MT systems to the X5gon domain, especially for language pairs with comparatively less support from Google Translate. This is the case of the Slovenian↔English pairs, which are very relevant to X5gon.

In line with the comparison to Google Translate, the MT systems developed in the framework of X5gon were also thoroughly assessed in international competitions such as WMT [7, 12]. In this regard, these MT systems were ranked on average among the top five systems including those from strong

academia (UCambridge, RWTH Aachen, DFKI, JHU, UEDIN and LIUM) and industry (Microsoft and Facebook) representatives.

4 Conclusions and open opportunities

This deliverable closes the work done in Task 3.3 from M12 (August 2018) to M30 (February 2020), to provide support for cross-lingual OER in X5gon. On the one hand, good progress was made in ASR for automatic OER transcription in the dominant languages of the official pilots (English, Spanish, Slovenian and German). Starting from background ASR systems developed by UPV prior to X5gon, they were largely improved by applying the latest scientific developments in the ASR field. Proof of this is their average relative error reduction of 40% with respect to the performance offered by the renowned Google Cloud Speech-To-Text web service, which was even higher in the in-domain tasks for Spanish (54% for pM-Es) and Slovenian (56% for VL-Sl). To us, the performance of our final ASR systems is in general good enough to support OER transcriptions of publishable quality, also amenable for machine translation and other cross-lingual uses of great interest to X5gon such as searching and recommendation.

On the other hand, as with OER transcription, the progress made in MT for automatic OER translation is also remarkable. Our work in this area paralleled that in ASR, by updating background UPV systems, and developing new ones, in line with the state-of-the-art, Transformer-based technology now prevailing in (Neural) MT. We did not limit ourselves to MT pairs of languages from those dominant in the official pilots. Instead, we also covered other, relevant languages in the EU (i.e. French, Italian and Portuguese) with the aim of providing full MT support to (potential) new X5gon network sites, maybe using English as a pivot language. Again as in ASR, the quality of the MT systems delivered in M30 was compared to that provided by Google, in this case Google Translate, and we found that our final MT systems are far better than Google Translate for Spanish↔Portuguese and Slovenian↔English, though more or less on par for all other language pairs. Given the high quality this Google service is delivering in recent years, it is good to report that we managed not only to catch up with it, but also to achieve much better results for language pairs such as Slovenian↔English, possibly much more relevant to X5gon than to Google.

Although we made good progress in Task 3.3 to provide support for cross-lingual OER in X5gon, we are totally convinced that there are still open research opportunities worth of exploration. On the one hand, our brand new one-pass decoding approach for ASR system building is allowing us to deliver real-time transcription with (almost) no degradation to accuracy. A direct benefit from this approach is a largely increased efficiency of ASR systems, which enables us to process larger volumes of OER content from X5gon. Also, we think that it would be good to extend this live transcription technology to *live speech translation* in general, and then explore how to best use it in *live cross-production and cross-consumption of OER content* across the X5gon OER network. In connection to this, on the other hand, we would like to "close the loop" of (live) OER translation by also including dubbing from *Text-To-Speech (TTS)* at the end of the cross-lingual processing pipeline. Indeed, we have been already moving along this direction in the framework of Task 5.2, in which the work plan for the second half of Y3 includes piloting of advanced cross-lingual and cross-modal features.

A Work done from M25 to M30 on Slovenian OER transcription

During this period, we worked on improving the Slovenian ASR system reported in deliverable D3.4 [1]. Briefly speaking, this was done by gathering more data and improving models, particularly acoustic models. As a result, we got a reduction of $\sim$ 13% WER in the VideoLectures.NET task. In what follows, this is described in detail.

Acoustic modelling

For the new Slovenian ASR system, DNNs were replaced by Bidirectional LSTMs for acoustic modelling (BLSTMs) [13]. Another important difference is that 85-dimensional Filter Bank input vectors were used instead of the original 16 MFCCs with derivatives (48-dimensional vectors).

Regarding the training procedure, the transLectures-UPV toolkit (TLK) was used to train a DNN-HMM model which was then used to bootstrap BLSTMs [14]. In particular, BLSTM training consisted in a cross-entropy training procedure with a limited back propagation through time window of 50 frames, in a way similar to that described in [13]. BLSTM training was carried out using TensorFlow [15]. The main features of the resulting BLSTM model are:

- Output layer size (HMM topology): 14497 tiedphoneme 3-state HMM.
- 4 hidden layer BLSTM with 512 output cells per direction (1024 per layer).
- Input layer size: 85 Filter Bank features.

The BLSTM acoustic model was trained using about 177 hours of Slovenian speech data from open resources using the transLectures UPV toolkit (TLK) [14].

Language modelling

The language model (LM) was an interpolation of two different LMs, a 4-gram LM and LSTM LM, with a vocabulary size of 500K words. Both models were trained from scratch using the datasets summarized in Table 6 and other open linguistic resources. The 4-gram model was built as a linear

Table 6: Slovenian linguistic resources for language modelling (Sentences =S, Running words = RW, Vocabulary = V).

Corpus	S(K)	RW(K)	V(K)
Europarl-v7 [16]	623	12559	135
ccGigafida [17]	7448	110098	1212
Newscrawl2011 [18]	1000	18599	433
TED [19]	54	339	39
VideoLectures.NET [20]	10	229	27
Wikipedia [21]	1804	22330	776
Wit3 [22]	15	200	29

interpolation of 4-gram LMs trained for each one of the domain corpus. The interpolation weights were tuned to optimize the perplexity of the VideoLectures.NET development set.

In the case of the LSTM LM, all linguistic resources were used for *Noise Contrastive Estimation (NCE)* training by means of the CUED-RNNLM toolkit [23]. The full vocabulary (500K words) was modeled in the output layer. Thus, the final topology for the LSTM LM consisted of an embedding

layer, a hidden LSTM layer of dimension 1024, and an output softmax layer of dimension $500K$. Finally, the LSTM and the 4-gram LMs were interpolated to optimize the perplexity of the VideoLectures.NET development set. The final weight for the LSTM LM in the interpolation was 0.67. Perplexities for each LM over the VideoLectures.NET development and test sets are shown in Table 7. Out-of-Vocabulary (OOV) rates of 0.7% and 1.7% for the development and test sets, respectively, were achieved with the new lexicon size of 500k words.

Table 7: Perplexities of the Slovenian LMs on VideoLectures.NET.

Model	dev	test
4-gram	473	666
LSTM	319	401
Interpolation	284	363

The decoding was carried out using a new developed one-pass-decoder which is able to directly interpolate on-the-fly the LSTM LM and the 4-gram LM during the regular decoding [24].

Evaluation

Apart from the VideoLectures.NET task, the Slovenian ASR system was also assessed on the publicly available SI-TEDx-UM corpus derived from TEDx Talks [19]. Table 8 shows basic statistics for the development and evaluation sets used in these tasks.

Table 8: Statistics for the Slovenian dev and test sets (Running Words = RW).

Set	Corpus	Dur.(h)	RW(K)	Voc.(K)
Dev	VideoLectures.NET [20]	2.9	27.4	5.2
Test	VideoLectures.NET [20]	3.2	23.2	5.9
	SI-TEDx-UM [19]	3.2	26.9	7.0

Table 9 shows the WER figures achieved on the evaluation sets for the Slovenian ASR systems developed throughout the project. It shows new Slovenian ASR system's relative improvements of 32.3% and 33.7% on the VideoLectures.NET (VL) and SI-TEDx-UM (TED) evaluation tasks, respectively.

Table 9: WERs figures for Slovenian ASR achieved throughout the project.

Period	System	VL	TED
Y1	Initial ASR System	32.5	27.6
	+ LSTM-LM NBest Rescoring	31.4	26.3
Y2	+ more speech data	30.1	24.3
	+New (N-gram and LSTM) LMs	28.1	22.6
	+ more speech data & one-pass decoding	25.3	20.0
M25-M30	+ BLSTM & more speech data	22.0	18.3

B Work done from M25 to M30 on Spanish OER transcription

During this period we worked on improving the Spanish ASR system we were using since the very beginning of X5gon. To this end, we resorted to the latest advances in the ASR field, which resulted in a reduction of $\sim 20\%$ WER in the PoliMedia task. This is described below in detail.

Acoustic modelling

As with the Slovenian ASR system, our former DNN-based acoustic models were replaced by BLSTMs trained from 85-dimensional Filter Banks vectors as input. The training procedure was also similar to that of the Slovenian ASR system, using our in-house toolkit TLK [14]. The model topology is:

- Output layer size (HMM topology): 10041 tiedphoneme 3-state HMM.
- 8 hidden layer BLSTM with 512 output cells per direction (1024 per layer).
- Input layer size: 85 Filter Bank features.

BLSTMs were trained from Spanish speech data resources described in Table 10. Apart from the open resources available from different sources and the PoliMedia audio tracks, we included a dataset provided during the IberSpeech2018 conference for the Speech-To-Text challenge "RTVE Albayzin" [25]. This dataset is a very valuable resource as it is a clean database of speech data from different TV shows (i.e: news broadcast, debates, quiz shows, etc.).

Table 10: Statistics of annotated speech resources for Spanish acoustic modelling.

Corpus	Duration(h)
Internal data	3624
PoliMedia (Es)	261
RTVE Albayzin (Es)	186

Language modelling

As before, we followed a one-pass decoding approach in which a 4-gram and a LSTM LMs were interpolated on the basis of a vocabulary of 200K words drawn from the linguistic resources in Table 11.

Table 11: Spanish resources for language modelling (S=Sentences, R=Running words, V=Vocabulary).

Corpus	S(M)	R(M)	V(M)
Common Crawl [26]	1719	41792	1337
eldiario.es [27]	1665	47542	892
El Periódico [28]	2677	46637	435
Europarl [16]	2112	55301	202
News Commentary [18]	207	5448	99
News Crawl [18]	7532	198545	958
Opensubtitles [29]	212635	1146861	5751
PoliMedia [30]	76	1372	60
UFAL [31]	92873	910728	2179
UN [32]	11196	343594	381
VideoLectures MT [33]	2	48	5
Wikipedia [21]	32686	586068	8357

The quality of the resulting Spanish LMs was measured in terms of perplexity on the development and the test sets for the PoliMedia task, using both models separately, and in combination by means of a linear interpolation. Table 12 shows the results.

Table 12: Perplexities of the Spanish LMs on PoliMedia.

Model	dev	test
4-gram	211.3	256.4
LSTM	117.8	135.0
Interpolation	112.4	129.4

As indicated above, we have now followed a one-pass decoding approach to produce fast yet accurate systems in which LM interpolation is carried out on the fly [24]. This allows us to build systems whose computational performance is below real time (Real Time Factor -or RTF- below one), and thus they are suitable for use under streaming requirements.

Evaluation

Table 13 shows basic statistics for the development and evaluation sets of the PoliMedia task and the international benchmark "RTVE Albayzin". It is important to note that, while we used the official *dev2* subset as a test set, we used as the development set only a subset of what was provided with the name *dev1* during the challenge.

Table 13: Statistics for the Spanish dev and test sets (Running Words = RW).

Corpus	Set	Dur.(h)	RW(K)	Voc.(K)
PoliMedia	Dev	3.9	35101	4708
	Test	3.4	30134	4252
RTVE albayzin	Dev	7	157821	13379
	Test	6	149265	12342

Table 14 shows the WER figures achieved on the PoliMedia and RTVE Albayzin tasks by the new Spanish ASR system built during the latest reporting period. Considering the complexity of the tasks (different speakers, different acoustic conditions, etc), specially for the RTVE set, these figures clearly show that the new ASR system for Spanish delivers results of high quality, and thus can be safely used in X5gon for Spanish OER transcription.

Table 14: WERs figures for Spanish ASR on Polimedia and RTVE Albayzin sets.

Task	Dev	Test
PoliMedia	8.3	9.1
RTVE Albayzin	15.2	13.0

References

[1] D3.4: Early support for cross-lingual OER. Technical report, X5gon project, M24, 2019.

[2] D5.1: First report on piloting. Technical report, X5gon project, M12, 2018.

[3] D5.1: Second report on piloting. Technical report, X5gon project, M24, 2019.

[4] Javier Jorge, Adrià Martínez-Villaronga, Pavel Golik, Adrià Giménez, et al. MLLP-UPV and RWTH Aachen Spanish ASR Systems for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge. In *IberSpeech2018*, pages 257–261, Barcelona (Spain), 2018.

[5] Google Cloud API - Speech-to-Text. `https://cloud.google.com/speech-to-text/`.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *NIPS2017*, pages 6000–6010, 2017.

[7] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, et al. Findings of the 2019 conference on machine translation. In *WMT2019*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics.

[8] J. Niehues, R. Cattoni, S. Stüker, et al. The IWSLT 2019 Evaluation Campaign. In *IWSLT2019*. Zenodo, November 2019.

[9] Google Translate. `https://translate.google.com/`.

[10] Marcin Junczys-Dowmunt. Dual conditional cross-entropy filtering of noisy parallel corpora. In *WMT2018*, pages 888–895, 2018.

[11] Nathan Ng, Kyra Yee, Alexei Baevski, et al. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation*, pages 314–319, Florence, Italy, August 2019. Association for Computational Linguistics.

[12] Javier Iranzo-Sánchez, Gonçal V. Garcés Díaz-Munío, Jorge Civera, and Alfons Juan. The MLLP-UPV Supervised Machine Translation Systems for WMT19 News Translation Task. In *WMT19*, pages 218–224, Florence (Italy), 2019.

[13] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, et al. A Comprehensive Study of Deep Bidirectional LSTM RNNs for Acoustic Modeling in Speech Recognition. In *ICASSP2017*, pages 2462–2466, New Orleans, LA, USA, March 2017.

[14] M.A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, et al. The Translectures-UPV Toolkit. In *Advances in Speech and Language Technologies for Iberian Languages*, volume 8854 of *Lecture Notes in Computer Science*, pages 269–278. Springer International Publishing, 2014.

[15] TensorFlow. `https://www.tensorflow.org/`.

[16] Europarl Corpus: European Parliament Proceedings Parallel Corpus v7. `http://www.statmt.org/europarl/`.

[17] Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, and Peter Holozan. Written corpus ccGigafida 1.0, 2013. Slovenian language resource repository CLARIN.SI.

[18] News Crawl corpus (WMT workshop) 2015. `http://www.statmt.org/wmt15/translation-task.html`.

[19] Andrej Zgank, Mirjam Sepesy Maucec, and Darinka Verdonik. The SI TEDx-UM speech database: a new Slovenian Spoken Language Resource. In *LREC2016*, Paris, France, may 2016. European Language Resources Association (ELRA).

[20] UPVLC, XEROX, JSI-K4A, RWTH, and EML. D3.1.3: Final report on massive adaptation. Technical report, transLectures, 2014.

[21] Wikipedia. `https://www.wikipedia.org/`.

[22] Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit^3: Web inventory of transcribed and translated talks. In *EAMT2012*, pages 261–268, Trento, Italy, May 2012.

[23] Xi Chen, Xin Liu, Y. Qian, Mark J. F. Gales, et al. Cued-rnnlm — an open-source toolkit for efficient training and evaluation of recurrent neural network language models. In *ICASSP2016*, pages 6000–6004, 2016.

[24] Javier Jorge, Adrià Giménez, Javier Iranzo-Sánchez, et al. Real-time One-pass Decoder for Speech Recognition Using LSTM Language Models. In *Interspeech 2019*, Graz (Austria), 2019. in press.

[25] Eduardo Lleida, Alfonso Ortega, Antonio Miguel, et al. Albayzin 2018 Evaluation: The IberSpeech-RTVE Challenge on Speech Technologies for Spanish Broadcast Media. *Applied Sciences*, 9(24):5412, 2019.

[26] commoncrawl 2014. `http://commoncrawl.org/`.

[27] Eldiario.es. `https://www.eldiario.es/`.

[28] ElPeriodico.com. `https://www.elperiodico.com/`.

[29] OpenSubtitles. `http://www.opensubtitles.org/`.

[30] poliMedia. `https://media.upv.es/#/catalog`.

[31] UFAL Medical Corpus. `http://ufal.mff.cuni.cz/ufal_medical_corpus`.

[32] Chris Callison-Burch, Philipp Koehn, Christof Monz, et al. Findings of the 2012 workshop on statistical machine translation. In *WMT2012*, pages 10–51, Montréal, Canada, June 2012. Association for Computational Linguistics.

[33] VideoLectures.NET. `http://videolectures.net/`.