transLectures

J. A. Silvestre-Cerdà, M. A. del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan

Universitat Politècnica de València (UPV), Camí de Vera s/n, 46022 València
{jsilvestre,mdelagua,ggarces,ggasco,agimenez,aperez,isanchez,nserrano rnadine,jvalor,jandres,jcivera,josanna,ajuan}@dsic.upv.es
http://translectures.eu

Abstract. transLectures (Transcription and Translation of Video Lectures) is an EU STREP project in which advanced automatic speech recognition and machine translation techniques are being tested on large video lecture repositories. The project began in November 2011 and will run for three years. This paper will outline the project's main motivation and objectives, and give a brief description of the two main repositories being considered: VideoLectures.NET and poliMedia. The first results obtained by the UPV group for the poliMedia repository will also be provided.

Keywords: language technologies, machine translation, automatic speech recognition, massive adaptation, intelligent interaction, education, video lectures, multilingualism, accessibility, opencast matterhorn

1 Introduction

transLectures is the acronym adopted for the EU (FP7-ICT-2011-7) STREP project entitled "Transcription and Translation of Video Lectures". It began on 1 November 2011 and will run until 31 October 2014. The transLectures consortium includes video lecture providers (users), experts in automatic speech recognition (ASR) and machine translation (MT), and professional transcription and translation providers:

UPV Universitat Politècnica de València, València, Spain.
XEROX Xerox S.A.S., Grenoble, France.
JSI Jozef Stefan Institute, Ljubljana, Slovenia.
RWTH Rheinisch-Westfaelische Technische Hochschule, Aachen, Germany.
EML European Media Laboratory GmbH, Heidelberg, Germany.
DDS Deluxe Digital Studios Ltd, London, UK.

The aim of this paper is to outline the motivation behind the transLectures project (sec. 2) and the main objectives established for it. Brief descriptions will be provided of the poliMedia and VideoLectures.NET repositories (sec. 3), as well as of the first results obtained by UPV for poliMedia (sec. 4).

2 Motivation and project objectives

Online multimedia repositories are growing rapidly and becoming evermore consolidated as key knowledge assets. This is particularly true in the area of education, where large repositories of video lectures are being established on the back of increasingly available and standardized infrastructures. A well-known example of this is VideoLectures.NET, a free and open access web portal that has already published more than 10,000 educational videos. VideoLectures.NET is a major player in the diffusion of the open-source Matterhorn platform currently being adopted by many education institutions and organizations within what is known as the Opencast community[1].

As with many other repositories, most of the video lectures available on VideoLectures.NET are neither transcribed nor translated, despite the clear need to make their content accessible to speakers of different languages and people with disabilities. Transcription and translation would also enable the incorporation of search and analysis functions, such as lecture classification, summarisation or plagiarism detection. However, a cost-effective solution that can do so to a reasonable degree of accuracy has yet to be developed.

The aim of the transLectures project is to do just this: to develop innovative, cost-effective solutions for producing accurate transcriptions and translations of the lectures published on VideoLectures.NET, with their compatibility and usability across other Matterhorn-related repositories informing all design decisions. Our starting hypothesis is that the gap that current automatic speech recognition (ASR) and machine translation (MT) technologies must bridge in order to achieve acceptable results for the kind of audiovisual collections being considered in the project is relatively small, and can be closed by pursuing the following three scientific and technological objectives:

1. *Improvement of transcription and translation quality by massive adaptation.* ASR has yet to reveal its full potential in the generation of acceptable transcriptions for large collections of audiovisual material. However, that potential is within reach and relatively little further research into ASR technology is required; rather we must learn to better exploit the wealth of knowledge we have at hand. More precisely, through this project we hope to demonstrate that transcriptions of a reasonable quality can be obtained through the massive adaptation of general-purpose models on the basis of lecture-specific variables, such as speaker, topic and, more importantly, time-aligned slides. Only once we have achieved acceptable transcriptions can we hope to address the adaptation of translation models with any degree of success.
2. *Improvement of transcription and translation quality by intelligent interaction.*
   Massive adaptation can and will deliver substantial contributions to the improvement of overall quality, but it is our belief that sufficiently accurate translations are unlikely to be achieved through fully-automated approaches alone: in order to reach the desired levels of accuracy, we must consider user

[1] See http://www.opencastproject.org/project/matterhorn

interaction. Current user models for the transcription and translation of audiovisual material tend to be batch-oriented. Under this model, an initial transcription/translation is generated by the system offline and then sent to the user to be post-edited manually without system participation. In our view, these models only give satisfactory results when highly collaborative users are working on near-perfect system output. Otherwise, a more intelligent interaction model is required that saves on user supervision and allows the system to learn from user supervision actions dynamically. In transLectures, our aim is to develop innovative, truly-interactive models in which the system learns from and reacts to each user supervision action as and when received. We are also exploring various modes of interaction, to reflect the multiple roles users can have.

3. *Integration into Matterhorn to enable real-life evaluation.*
   In contrast to many past research efforts in which system prototypes are evaluated in the lab alone and are largely inapplicable to real-life settings, we will be developing tools and models for use with Matterhorn. We will therefore be able to evaluate their usefulness using real-life data in a real-life context.

We are confident that the innovative solutions we develop in these three areas will be deployed rapidly across many educational repositories in Europe and worldwide, allowing these portals to overcome language barriers and reach wider audiences while supporting linguistic diversity. In transLectures, we will be testing our ideas on VideoLectures.NET and on a smaller repository of Spanish video lectures, poliMedia, which is also part of the Matterhorn community. For transcription, we are considering English and Slovenian through VideoLectures.NET and Spanish through poliMedia, with the former accounting for more than 90% of all lectures. Meanwhile, for translation we are considering the language pairs {Spanish, Slovenian} into English, and English into {French, German, Slovenian, Spanish}.

3 VideoLectures.NET and poliMedia

The best proof of the effectiveness of the transLectures approach would be the cost-effective production of accurate transcriptions and translations within the context of one or more video lecture repositories. As discussed above, we are testing our ideas on two case studies: VideoLectures.NET and poliMedia.

VideoLectures.NET was founded in 2001 as an internally-funded project and is now run by the dedicated Centre for Knowledge Transfer in Information Technologies at the Josef Stefan Institute (JSI), Ljubljana, Slovenia. It is a free and open access repository of video lectures mostly filmed by people from JSI[2] at major conferences, summer schools, workshops and promotional events from many scientific fields. Indeed, VideoLectures.NET is being used as an educational platform for several EU-funded research projects[3]; and by various open education

[2] See promo video at http://videolectures.net/april2010_videolectures_promo
[3] See http://videolectures.net/eu_supported

resource organizations such as the OpenCourseWare Consortium, MIT OpenCourseWare and Open Yale Courses; as well as other scientific institutions like CERN. In this way, VideoLectures.NET brings together high-quality educational content that has been recorded to similarly high quality standards. All lectures, accompanying documents, information and links are systematically selected and classified through the editorial process, and user comments are also taken into account. Video editing is carried out in-house and is completely uncensored, in the sense that lectures are never edited in a way that would allow content or viewer manipulation. Most lectures are also accompanied by time-aligned presentation slides[4].

poliMedia is a recent, innovative service for the creation and distribution of multimedia educational content at the UPV. It is designed primarily to allow UPV professors to record their courses in video blocks lasting up to 10 minutes. As in VideoLectures.NET, the videos are accompanied by time-aligned slides. However, in contrast to VideoLectures.NET, the video recordings are filmed at specialized studios under controlled conditions to ensure maximum recording quality and homogeneity. Professors are filmed against a constant-colour background in order to be able to post-produce presentations in which the professor appears, properly scaled, against a backdrop of the corresponding slides. Please visit http://polimedia.blogs.upv.es/?lang=en for more details and examples[5].

transLectures is focusing on English and Slovenian video lectures from VideoLectures.NET and on Spanish video lectures taken from poliMedia. For an idea as to the complexity of the task, some basic statistics have been provided in Table 1.

	English	Slovenian	Spanish	Total
Authors	6900	1347	734	8981
Lectures	9720	1103	5056	15879
Avg. lecture duration (min)	45	45	9	34
Transcribed lectures	111	0	7	118
Lectures with slides	7013		734	

Table 1. Basic statistics on the video lectures considered in transLectures.

4 First UPV results on poliMedia

The UPV obtained its first results for poliMedia working from a set of 115 hours of video lectures manually transcribed using the Transcriber tool [1]. From this set, a standard partition was defined with three speaker-independent sets: training, development and test. This will allow ongoing scientific evaluation through-

[4] e.g. http://videolectures.net/translingeu2010_uszkoreit_rossi_welcome

[5] See http://polimedia.upv.es/visor/?id=39f62a9a-4cf5-bd4e-92f3-cb34e4792a85 for a brief presentation in Spanish, with subtitles available in English.

out the project. Table 2 shows the basic statistics for this standard partition.

	Training	Development	Test
Videos	559	26	23
Speakers	71	5	5
Hours	99h	3.8h	3.4h
Sentences	37K	1.3K	1.1K
Vocabulary	28K	4.7K	4.3K
Running words	931K	35K	31K
OOV words	-	4.6%	5.6%
Perplexity	-	222	235

Table 2. Statistics for the standard partition for scientific evaluations on poliMedia.

The baseline UPV ASR system is based on the RWTH ASR system [2, 4] for acoustic modelling and the SRILM toolkit [6] for language modelling. The RWTH ASR system includes state-of-the-art speech recognition technology for acoustic model training and recognition. It also includes speaker adaptation, speaker adaptive training, feature extraction for audio files, unsupervised training, a finite state automata library and an efficient tree search recognizer. For its part, the SRILM toolkit is a well-known language modelling toolkit used across different natural language applications.

The audio data was extracted and preprocessed from the videos, in order to then extract the mel-frequency cepstral coefficients (MFCCs) [5]. Then, monophoneme and triphoneme acoustic models were trained by adjusting different parameters such as the number of states, Gaussian components, leaves of the triphoneme CART, etc. in the development set. The lexicon model was obtained by applying phonetic transliteration to all of the training vocabulary words. An n-gram language model was trained on the transcribed text after filtering out functional symbols such as punctuation marks, silence annotations, etc. Meanwhile, external resources were used to enrich the in-domain language model. Specifically, we considered the linear combination of our in-domain language model with a large, external out-of-domain language model computed from the Google n-gram corpus [3]. A single parameter λ governs the linear combination of the poliMedia language model and the Google n-gram model, which is optimized in terms of perplexity on the development set.

The entire system, including the acoustic, lexicon and language models, was trained on the poliMedia training set. The ASR system parameters were optimized in terms of word error rate (WER) on the development set. A significant improvement of more than 5 WER points was observed when moving from monophoneme to triphoneme acoustic models. Triphoneme models were inferred using the conventional CART model using 800 leaves. In addition, other parameters obtained to train the best acoustic model included 2^9 components per Gaussian mixture, 4 iterations per mixture, and 5 states per phoneme. The in-domain

language model was an interpolated trigram model with Kneser-Ney discount. Higher and lower order n-gram models were also assessed, but no better performance was observed.

From the triphoneme ASR system, several refinements to the language model were evaluated. The in-domain language model trained on the poliMedia corpus was interpolated with the out-of-domain Google n-gram corpus [3]. These two language models were interpolated in order to minimize perplexity in the development set, using an approximate λ value of 0.65 for the in-domain language model and of 0.35 for the out-of-domain language model. Two interpolations were performed using different vocabulary sets, the first containing only vocabulary matching poliMedia ("LM Interpolation") and a second made up of the poliMedia vocabulary plus the 50,000 most frequent words in the Google n-gram corpus ("Google 50K"). The final experimental results in terms of WER in the test set are shown in Table 3.

System	WER	OOV
Monophoneme Model	44.6	5.6%
Triphoneme Model	39.4	5.6%
+LM Interpolation	34.6	5.6%
+Google 50K Vocab	33.7	3.5%

Table 3. Test-set WER for several ASR system refinements.

As shown in Table 3, there is a significant improvement by 5.7 WER points over the triphoneme system when the language model interpolated with the "Google 50K Vocab" vocabulary set is applied. As expected, the decrease in WER is directly correlated with the number of out-of-vocabulary words (OOVs) in the test set, since the Google n-gram corpus provides a better vocabulary coverage. A similar trend is observed when comparing perplexity figures for the triphoneme system with those observed for the "LM Interpolation" system. Specifically, perplexity drops significantly from 235 to 176 simply by interpolating our in-domain language model with the Google n-gram language model containing poliMedia vocabulary alone.

5 Concluding remarks

In this paper, we have outlined the transLectures project's main motivation and objectives, and given a brief description of the two video lecture repositories being considered: VideoLectures.NET and poliMedia. We have also reported the first ASR results obtained by UPV for poliMedia.

Our current work is aimed at the application of massive adaptation techniques on the basis of lecture-specific knowledge and, in particular, time-aligned slides. We are also running similar experiments on the larger VideoLectures.NET repository.

Acknowledgments. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755. Funding was also provided by the Spanish Government (iTrans2 project, TIN2009-14511; FPI scholarship BES-2010-033005; FPU scholarship AP2010-4349).

References

1. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication special issue on Speech Annotation and Corpus Tools 33(1–2) (2000)
2. Lf, J., Gollan, C., Hahn, S., Heigold, G., Hoffmeister, B., Plahl, C., Rybach, D., Schlter, R., Ney, H.: The rwth 2007 tc-star evaluation system for european english and spanish. In: Proc. of Interspeech. pp. 2145–2148 (2007)
3. Michel, J.B., et al.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176–182.
4. Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lf, J., Schlter, R., Ney, H.: The rwth aachen university open source speech recognition system. In: Proc. of Interspeech. pp. 2111–2114 (2009)
5. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition 54(4), 543–565 (2012)
6. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proc. of ICSLP (2002)