LANGUAGE MODEL ADAPTATION FOR VIDEO LECTURES TRANSCRIPTION

Adrià Martínez-Villaronga, Miguel A. del Agua, Jesús Andrés-Ferrer, Alfons Juan

PRHLT, Universitat Politècnica de València (UPV)

ABSTRACT

Videolectures are currently being digitised all over the world for its enormous value as reference resource. Many of these lectures are accompanied with slides. The slides offer a great opportunity for improving ASR systems performance. We propose a simple yet powerful extension to the linear interpolation of language models for adapting language models with slide information. Two types of slides are considered, correct slides, and slides automatic extracted from the videos with OCR. Furthermore, we compare both time aligned and unaligned slides. Results report an improvement of up to 3.8 % absolute WER points when using correct slides. Surprisingly, when using automatic slides obtained with poor OCR quality, the ASR system still improves up to 2.2 absolute WER points.

***Index Terms*—** language model adaptation, video lectures

1. INTRODUCTION

Online multimedia repositories are rapidly growing and imposing themselves as fundamental knowledge assets. This is particularly true in the area of education, where large repositories of video lectures are being built relying on increasingly available and standardised infrastructure [1, 2, 3, 4]. This repositories are making the education accessible to a wide community of potential students. As with many other repositories, most lectures are not transcribed because of the lack of efficient solutions to obtain them at a reasonable level of accuracy. However, transcription of video lectures is clearly necessary to make them more accessible. Also, they would facilitate lecture searchability and analysis, such as classification, summarisation, or plagiarism detection. In addition, communities of people with hearing disabilities would be able to follow the lectures just by reading the transcriptions.

Manual transcription of these repositories is excessively expensive and time-consuming and current state-of-the-art automatic speech recognition (ASR) has not yet demonstrated its potential to provide acceptable transcriptions on large-scale collections of audiovisual objects. However, it has such potential by simply exploiting the rich knowledge we have at hand. More precisely, in this kind of videos the speaker is accompanied by some kind of background slides during its presentation. In these cases, a strong correlation can be observed between slides and speech. Consequently, this slides provide an interesting opportunity to adapt general-purpose ASR models by massive adaptation from lecture-specific knowledge.

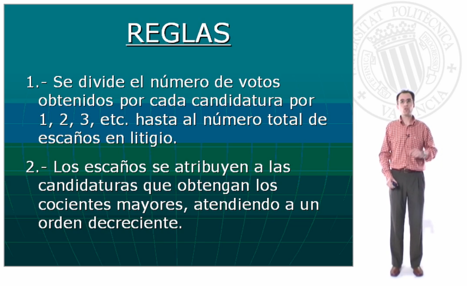
The proposed scenario is considered by some projects which aim at providing full set of transcriptions for online lecture repositories. Our work is framed in the European `trans`**Lectures** [5, 6] project, which is explained in Section 2 and whose objective is to develop innovative and cost-effective solutions to produce accurate transcriptions and translations in VideoLectures.NET and poliMedia [7] through the free and open-source platform Matterhorn [8].

Within the framework of the `trans`**Lectures** project, our intention is to improve the video lectures transcriptions of the poliMedia database by adapting language models to the content of the slides. We consider two different scenarios: (i) correct transcriptions of slides is available, (ii) only the video of the lecture is available and the text has to be extracted with automatic techniques. The first scenario represents an ideal scenario where the slide text is correctly extracted from the slide and aligned with the speech, so that for a given speech segment, the used slide is known. The second scenario is the worst case in which an electronic copy of the slides is not available and then the text has to be directly extracted from the video. Specifically, we applied the simple OCR approach described in Section 3.

The idea of using lecture slides is not new and has already been explored by some authors [9, 10]. However, these works typically assume a perfect transcription of the slides, and do not take into account slide synchronization since they are mainly focused on keyword detection for indexing. In [9] a fairly small database is used to report slight improvements in keyword detection. Surprisingly enough in [9] slide adaptation is performed without using a simple linear interpolation of models and 3 sophisticated techniques are proposed. Authors argue that preliminary results with linear interpolation did not report improvements.

In contrast to previous works, we obtain improvements of up to 3.8 absolute WER points with respect to a strong baseline, which is obtained from several external corpus such as Google n-grams [11], by simply extending linear interpolation of language models to the slide properties in section 4.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755 (transLectures). Also supported by the Spanish Government (Plan E, iTrans2 TIN2009-14511).



REGLAS

1.- Se divide el número de votos obtenidos por cada candidatura por 1, 2, 3, etc. hasta al número total de escaños en litigio.

2.- Los escaños se atribuyen a las candidaturas que obtengan los cocientes mayores, atendiendo a un orden decreciente.

Fig. 1. A poliMedia video capture

Specifically, the main contributions of this work are the following: (1) linear interpolation method [12] is extended to the peculiarities of lecture slides; (2) this extension is compared with a competitive baseline trained with several external sources (3) we compare both correct slide text and slide text automatically extracted from the video lectures by OCR; (4) the effect of using time aligned slides is assessed; (5) our approach reports experimental improvements in Section 5 of up to 3.8 % absolute WER points or 14.5 % relative, which correspond to an improvement of 59 absolute points in terms of perplexity or 34% relative.

2. *trans*Lectures

The aim of the *trans***Lectures** project is to develop innovative, cost-effective solutions to produce accurate transcriptions and translations in VideoLectures.net and poliMedia through a free and open-source platform called Matterhorn [8]. Matterhorn is a platform designed to support the creation and management of educational audio and video content.

The poliMedia database was created for production and distribution of multimedia educational content at the UPV. Lecturers are able to record lectures under controlled conditions which are distributed along with time-aligned slides. For the time being, the poliMedia catalogue includes almost 8000 Spanish videos accounting for more than 1000 hours of lectures of which only about 2000 videos can be accessed freely. The poliMedia corpus was created by manually transcribing 704 video lectures in Spanish corresponding to 100 hours so as to provide in-domain data sets for training, adaptation and internal evaluations in the *trans***Lectures** project (basic statistics are shown in table 1). Only open access video lectures were transcribed so that the corpus will be accessible by the research community beyond the scope of the *trans***Lectures** project.

A typical video capture from a poliMedia video lecture is depicted in figure 1. The lecturer is localised at the right side of the screenshot, while the slides are shown at the left side of the video.

Table 1. Basic statistics of poliMedia corpus.

		Time			Vocabulary
	Lectures	(hours)	# sentences	# words	size
train	655	96	41.5K	96.8K	28K
dev	26	3.5	1.4K	34K	4.5K
test	23	3	1.1K	28.7K	4K

3. SLIDES TEXT RETRIEVAL

In many online repositories the electronic format of the slides is typically not available together with the video. For instance, in the poliMedia case described in Section 2 uploading the slides with the video is an optional step that is many times disregarded. Consequently, there are two types of videos: those with the slides attached, and those without slides. For the former, slides text extraction only depends on tools such as *pdf2text*. Conversely, for the latter, slides must be automatically extracted from each video lecture. This automatic process is divided into 2 steps: first the slide is detected, and then a OCR tool, such as *Tesseract*, is used to extract the text from the detected slide.

Regarding the slide detection technique a very naive yet effective technique is proposed. Specifically, we count the changing pixels from frame to frame, and determine that a change in the slide has been performed if the number of changing pixels exceeds a specified threshold. Each time a new slide is detected, the corresponding frame is captured and passed to the Optical Character Recognition tool.

OCR has become an important and widely used technology for document annotation. However, when dealing with complex images the results turn out to be not as good [13] where an appropriate image preprocessing, text-line detection and text post-processing steps are fundamental. We used *Tesseract*[14] for optical character recognition (OCR).

Two different OCR approaches have been applied using Tesseract. Firstly, we carried out a slide recognition process where each slide was recognised according to different Tesseract parameter configuration in order to improve the transcriptions results. After the recognition, the output was filtered by some simple word generation rules.

Unfortunately, the previous approach provided poor performance due to irregular slide structure such as images, charts, tables and a wide variability in background and font colors. Consequently, for the second approach, we developed a preprocessing module which applies various filters such as despeckling, enhancing or pixel negation and obtains different versions of the same slide by applying several thresholds for binarisation. Each preprocessed slide version is processed by *Tesseract* and post-processed combining all the outputs. This process dramatically improved the accuracy of the obtained text.

4. LANGUAGE MODEL ADAPTATION

Our interest is to elucidate whether the slide information provides useful information with respect to a competitive LM baseline. If we compare the improvements obtained by adding slide information to a simple in-domain language model, we would obviously observe an astonishing improvement. For this reason we built a initial competitive baseline.

In order to build this competitive baseline, several n-gram models trained from different out of domain corpora were linearly mixed together with the in-domain model as follows. Let w be the current word within a sentence, and let h be the $n-1$ previous words, then the mixture is made by linear interpolation as follows:

$$p(w|h) = \sum_i \lambda_i p_i(w|h) \tag{1}$$

where λ_i is the weight of the linear interpolation corresponding to the i-th n-gram model $p_i(w|h)$. The weights $\{\lambda_1^I\}$ must add up to 1 so that the mixture is a probability. Finally, these weights are used to adapt the model by optimising them with the EM algorithm to maximise the log-likelihood or equivalently to minimise the perplexity of a given development set [12].

Corpora used for the baseline are described in Section 5.

In order to adapt the baseline, we add one or two language models to the linear interpolation discussed above. Specifically, we considered two approaches:

1. Add an extra model trained using the whole text of the slides used for the current video
2. Add two models, the previous model together with another trained only with the text in the current slide.

For the first case, given a video V, its adapted language model is defined as follows:

$$p(w|h, V) = \sum_i \lambda_i p_i(w|h) + \lambda_V p_V(w|h) \tag{2}$$

where V stands for the current video, $p_V(w|h)$ for the language model trained on the video.

In order to optimise the λ_V parameter, we had to extend the optimisation proposed in [12] to allow a changing language model, since the model $p_V(w|h)$ varies from one video to another, and the development set is supposed to be made up of several videos. In this way, we obtain a general parameter for all the slides. However, there are videos for which the slides do not contain text or do not make use of slides at all. Considering this videos as a normal videos will cause a distortion in the calculation of the interpolation weights, specially for this dynamic video-dependent n-gram model. Therefore, we add a constraint to the optimisation process such that if the slide does not contain text, then the λ_V is forced to be 0.

For the second case, we used a similar approach. However, similarly to the previous approach, there are several slides which contain little or no text at all, whereas other slides contain a lot of text. Actually, there is a higher variability than that of considering the text of the slides altogether.

Table 2. Basic statistics of corpora used to generate the LM

Corpus	# sentences	# words	Vocabulary
EPPS	132K	0.9M	27K
news-commentary	183K	4.6M	174K
TED	316K	2.3M	133K
UnitedNations	448K	10.8M	234K
Europarl-v7	2 123K	54.9M	439K
El Periódico	2 695K	45.4M	916K
news (07-11)	8 627K	217.2M	2 852K
UnDoc	9 968K	318.0M	1 854K

If we do not take this problem into consideration when linearly interpolating the language models, we will observe that the weight given to the slide dependent model will be smaller than it should for slides that contain more text, and the other way around. To amend this problem we use different weights for the slide dependent model depending on the number of words in the slide. We classify the slides in K classes depending on how many words the slide contains. For instance, we could consider 3 classes ($K = 3$): less than 10 words ($k = 0$), more than 10 but less than 100 ($k = 1$ and more than 100 words ($k = 2$).

For a slide S of a given video V that belongs to the class k, the linear interpolation of models is given by:

$$p(w|h, V, S) = \sum_i \lambda_i^{(k)} p_i(w|h) + \lambda_V^{(k)} p_V(w|h) + \lambda_S^{(k)} p_S(w|h) \tag{3}$$

where $p_S(w|h)$ stands for the model built from the current slide. For the special case in which the current slide is void, the first proposed model in Equation 2 is used as a back-off.

5. EXPERIMENTS

The proposed techniques for language model adaptation are measured in terms of both perplexity and WER obtained with state of the art ASR system [15]. The acoustic model has been trained using the poliMedia corpus (Table 1), employing triphonemes inferred using the conventional CART with almost 3900 leaves. Each triphoneme was trained for up to 64 mixture components per Gaussian, 4 iterations per mixture and 3 states per phoneme with the typical left-to-right topology without skips. Additionally, speaker adaptation was performed applying CMLLR feature normalisation (full transformation matrices). The results obtained with this model were competitive in the last trans**Lectures** evaluation.

As for the language model, we computed the baseline model as discussed in Section 4 by interpolating several individual language models trained in several corpora. Table 2 summarises the main statistics of the used corpora. For each out-of-domain corpora we trained a 4-gram language model with SRILM [16] toolkit . The individual 4-gram models were smoothed with modified Kneser-Ney absolute interpolation method [17]. In addition, a language model was trained from the Google counts corpus [11]. Finally, the training set of poliMedia (Table 1) was also used as the in-domain corpus. The final interpolated model were also pruned such that the perplexity increased by less than 2^{-10}. Perplexities obtained for each of this individual models are reported in Table 3. As for the vocabulary, we used the top 50K most frequent words over all the corpora plus the in-domain vocabulary.

We carried out experiments to analyse the improvement obtained from using both proposed techniques in section 4 with both automatic and correct slides.

The automatic slides were obtained as described in Section 3. The first OCR approach obtained a text WER of 70% when compared with the correct slides whereas the improved OCR method dropped it down to 43%. Both automatic approaches have high WER, and as an additional filtering step the n-gram counts extracted from the automatic slides were filtered erasing all the n-gram containing characters out of the Spanish alphabet.

Results are summarised in Table 4. First row show the baseline result of 24.8 points of WER (a). When adding a a 3-gram language model trained with the text of the correct slides for the whole video, as discussed in Equation (2), the WER drops down to 21.2 (c). This is an improvement of 3.6 absolute WER points. Since including the slides also extends the vocabulary, we also computed the results obtained including this vocabulary in the baseline (b). It is then observed that the vocabulary accounts for 1.4 of improvement out of the 3.6 WER points reported by the video-dependent LM.

In Table 4, we also assess the improvement obtained when introducing synchronized fined-grain slide models besides the video-dependent model (d). Due to the small amount of text in the development slides we decided to use 2-grams instead of 3-grams. The models were smoothed with modified Kneser Ney, and whenever the smoothing method failed, we reduced the order of the 2-grams down to 1. As explained in Section 4 we need to consider different situations depending on the amount of text in the slide. Specifically, we used 5 length classes depending on the number of words in the slide: between 1 and 10, between 11 and 40, more than 40 or 0 words, differentiating in that last case if only the current slide is empty or the whole video presentation is. Unfortunately, the results (d) report only small gains of 0.2 % absolute points in terms of WER. Hence, the synchronisation information does not seem relevant for our approach.

Finally, we ran experiments with the video-dependent language models but using the OCR extracted slides. Results are shown in Table 4. Surprisingly, the first OCR approach (e) (70% of WER with respect to the correct slide text) incurred in an improvement of 2 % absolute points of WER. This results is quite impressive taking into account that most of the text extracted from the slide is wrong. Finally, the improved OCR approach (f) furher improved this result up to 2.2 % absolute points of WER.

Table 3. Perplexities and OOV words on the development and test sets for all corpora

	Perplexity		OOVs (%)	
Corpus	dev	test	dev	test
EPPS	543.7	710.8	8.21	12
news-commentary	636	747.7	6.73	9.4
TED	615.6	521.2	6.59	7.94
UnitedNations	754	802.9	7.77	10.94
Europarl-v7	460.6	605.7	5.75	8.59
El Periódico	450.2	545.9	5.95	8.61
news (07-11)	358.9	747.6	5.64	7.99
UnDoc	544.9	802.8	6.10	9.21
Google	1370.3	1954.8	4.71	6.95
poliMedia train	317.9	332.5	4.61	5.23

Table 4. WER (%) and PPLs on the poliMedia corpus for several adapted language models

	Development		Test	
	PPL	WER	PPL	WER
(a)Baseline	140.8	22.1	172.1	24.8
(b)Baseline + Slides vocab	150.8	21.6	195.7	23.4
(c) Correct Slides(3-gram)	96.6	20.5	113.2	21.2
(d) C. Slides+SYNC(±0)	96.6	20.6	113.3	21
(e) OCR Slides(3-gram)	126.8	21.7	147.3	22.8
(f) OCR Improved(3-gram)	111.3	21.4	131.8	22.6

6. CONCLUSIONS

A simple yet effective method for adapting language models for video lectures using the information from slides is proposed. Our proposal involves two levels of adaptation: add the full presentation language model and adding both full presentation and current slide models. The proposed models report improvements up to 15% relative and 3.8% absolute in terms of WER.

Time aligned slides seem not to provided valuable information over the text slides, since currend-slide models do not significantly improved video-based models.

Two different scenarios were considered: one in which the correct slides text was available and the other where only the slide image was available and the text had to be extracted using OCR tools. Surprisingly, automatic slides proved to be valuable even when they contain a large number of errors as long as they are automatically filtered.

It has been observed that many times, lecturers refer to future slides when showing current slides. In the future, we intend to use a window of ±1 or ±2 slides for the single slide model to get more context for better using time-aligned slides. Additionally, we intend to compare our approach with cache language models.

7. REFERENCES

[1] "coursera.org: Take the World's Best Courses, Online, For Free," http://www.coursera.org/.

[2] "Videolectures.NET: Exchange ideas and share knowledge," http://www.videolectures.net/.

[3] "poliMedia: Videolectures from the "Universitat Politècnica de Valencià," http://polimedia.upv.es/catalogo/.

[4] "SuperLectures: We take full care of your event video recordings.," http://www.superlectures.com.

[5] UPVLC, XEROX, JSI-K4A, RWTH, EML, and DDS, "Transcription and Translation of Video Lectures," in *Proc. of EAMT*, 2012.

[6] UPVLC and XEROX and JSI-K4A and RWTH and EML and DDS, "*trans***Lectures** ," https://translectures.eu/, 2012.

[7] Universitat Politècnica de València, "poli[media]," https://polimedia.upv.es/, 2012.

[8] Markus Ketterl, Olaf A. Schulte, and Adam Hochman, "Opencast matterhorn: A community-driven open source solution for creation, management and distribution of audio and video in academia," in *Proc. of the 11th IEEE International Symposium on Multimedia (ISM 2009)*, San Diego (USA), dec 2009, pp. 687–692.

[9] Hiroki Yamazaki, Koji Iwano, Koichi Shinoda, Sadaoki Furui, and Haruo Yokota, "Dynamic language model adaptation using presentation slides for lecture speech recognition," in *In Proc. INTERSPEECH*, 2007, pp. 2349–2352.

[10] Tatsuya Kawahara, Yusuke Nemoto, and Yuya Akita, "Automatic lecture transcription by exploiting presentation slide information for language model adaptation," in *In Proc. ICASSP*, 2008, pp. 4929–4932.

[11] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Holberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden, "Quantitative analysis of culture using millions of digitized books," *Science*, 2010.

[12] Frederick Jelinek and Robert L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *In Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland, May 1980, pp. 381–397.

[13] Datong Chen, Jean-Marc Odobez, and Herv Bourlard, "Text detection and recognition in images and video frames," *Pattern Recognition*, vol. 37, no. 3, pp. 595 – 608, 2004.

[14] R. Smith, "An overview of the tesseract ocr engine," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, Washington, DC, USA, 2007, ICDAR '07, pp. 629–633, IEEE Computer Society.

[15] "Another Kit (AK) for building and use Hidden Markov Models," http://sourceforge.net/projects/aktoolkit/.

[16] Andreas Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. of ICSLP*, 2002.

[17] R Kneser and Hermann Ney, "Improved backing-off for M-gram language modeling," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, 1995, pp. 181–184.