Language Model Adaptation for Lecture Transcription by Document Retrieval

A. Martínez-Villaronga, M. A. del-Agua, J. A. Silvestre-Cerdà, J. Andrés-Ferrer, and A. Juan

MLLP, DSIC, Universitat Politècnica de València,
Camí de Vera s/n, 46022, València, Spain
{amartinez1,mdelagua,jsilvestre,jandres,ajuan}@dsic.upv.es

Abstract. With the spread of MOOCs and video lecture repositories it is more important than ever to have accurate methods for automatically transcribing video lectures. In this work, we propose a simple yet effective language model adaptation technique based on document retrieval from the web. This technique is combined with slide adaptation, and compared against a strong baseline language model and a stronger slide-adapted baseline. These adaptation techniques are compared within two different acoustic models: a standard HMM model and the CD-DNN-HMM model. The proposed method obtains improvements on WER of up to 14% relative with respect to a competitive baseline as well as outperforming slide adaptation.

Keywords: language model adaptation, video lectures, document retrieval

1 Introduction

As part of the continuous development and advances in information technologies, new channels and communication possibilities have been established. In the field of education, universities have made a great effort in knowledge dissemination, which has resulted in the creation of large multimedia repositories of lecture recordings [7, 2, 3] or MOOCs (Massive Open Online Courses) [1, 6]. The transcription of these repositories is an increasing necessity so as to achieve their massive dissemination for several reasons: for instance, transcriptions help to improve searchability, classification and analysis within huge repositories; they also help to reach wider audience of students by overcoming linguistic, as well as acoustic, barriers.

Each of these repositories is made up of hundreds or even thousands of videos, rendering manual transcription unfeasible in terms of both time and cost. Despite current state-of-the-art automatic speech recognition (ASR) systems are achieving continuous improvements over time, these repositories can be greatly improved through the use of specifically retrieved in-domain data. For instance, in [15] the video-dependent in-domain data is extracted/retrieved from the slides used in each video. Specifically, a general-purpose ASR system was adapted

through language model interpolation from different resources (out-of-domain and in-domain,) including the text of video-dependent slides. The conclusion is that slide-dependent language models could significantly improve the transcription quality.

Unfortunately, it is not always possible to obtain slides for a given video lecture. For instance, it is usual that either the author does not give access to the slide document, or the repository does not keep track of such files. When slides are not available in electronic format, they can be extracted from the video recording using OCR techniques [15]. However, due to the video quality, even this is not possible in many cases.

In addition to slide adaptation, some works have explored language model adaptation by building language models using documents retrieved from the web. Document retrieval techniques are fundamental in gathering relevant in-domain data, a large part of which are based on building search queries to locate documents through common search engines. Previous works have focused on document retrieval for broad and general ASR systems. Some authors have tried to build these queries from a first pass recognition [9, 14, 18] or from keyword detection [17]. Other works have tried to use the training set itself [23] or the text of the slides [16] to build the query. In summary, previous works have a strong focus on studying how to build the queries in order to have competitive recall and precision trade-off. In contrast, we propose to use the title of each video lecture as the search query, since they tend to be quite descriptive in video lecture repositories.

In this work, we focus on language model adaptation by document retrieval for the automatic transcription of video lectures. We compare our approach with a strong baseline computed from a large collection of out-of-domain and in-domain documents comprising 46 billion words. Furthermore, we compare our results with those obtained by slide adaptation [15], using as slides the text extracted from the video using OCR. We also combine both approaches to further improve adaptation which yields significant improvements with respect to both the baseline model and the slide-adapted model. In order to assess language model adaptation with increasingly better acoustic models, this comparison is made with two different acoustic models: the standard Hidden-Markov-Model (HMM) and the Context-Dependent Deep-Neural-Network Hidden-Markov-Model (CD-DNN-HMM) approach [19]. All these techniques fall within the scope of the European transLectures [20, 4] project whose objective is to develop innovative and cost-effective solutions to produce accurate transcriptions and translations in VideoLectures.NET and poli[Media] [2] through the free and open-source platform Matterhorn [12].

2 Document Retrieval

In this work, we focus on document retrieval from the web by building queries using the title of the video lecture. This is in contrast to other works where more complex techniques are proposed, such as rendering the lines of each slide

as queries [16], or extracting keywords from a first pass recognition to build queries [14]. Our method is built on the hypothesis that the title is very informative and tends to contain the most important keywords and they appear in the proper order. The proposed method has several advantages among other works, such as it can be used without the need of slides or a first pass recognition on the video. Furthermore, it is a general and simple technique for this kind of repositories. We should remark that sometimes the paper on which the lecture is based is downloaded. This is very useful for the adaptation, although finding this exact document is not the primary goal of the search.

To ensure that the documents retrieved are of a minimum-required quality, we constrain the search to pdf documents only, and not webpages. Note that typically pdf files are of a higher standard since they are usually papers, books or notes related to the lecture topic. Unfortunately, some of the retrieved documents might be in languages different from that of the video and must be filtered out.

We propose two search methods for retrieving N documents per video:

- **Exact search:** we download documents that exactly match the title of the video lecture, i.e. the title is contained within the text of the document. Sometimes the search produces less than N results. For instance, the lecture "*Especies de interés I. Aromáticas. Lamiaceae. Thymus*" ("*Species of interest I. Aromatic. Lamiaceae. Thymus*") produced 0 results.
- **Extended search:** we perform an exact search and the search is extended with documents that partially match the title if less than N documents are found. The extended search will retrieve all the documents from the exact search plus other documents that contain some of the words of the lecture title, up to N documents.

3 Language Model Adaptation Technique

The language model adaptation technique for video lectures was introduced in [15]. It combines out-of-domain language models, in-domain models and video-specific models by means of a linear interpolation:

$$p(w|h) = \sum_i \lambda_i p_i(w|h)$$

where weights are optimised in a development set according to [11].

This technique is extended to consider language models built from documents retrieved from the web as follows:

$$p(w|h, V) = \sum_i \lambda_i p_i(w|h) + \lambda_D p_D(w|h)$$

where V stands for the current video and $p_D(w|h)$ for the language model trained on the documents retrieved for V.

In this work, we further consider the scenario where the lecture slides can be extracted from the video using OCR and they are available to adapt the

models [15], or a mixed scenario that combines both the text in the slides and the retrieved documents as follows

$$p(w|h, V) = \sum_i \lambda_i p_i(w|h) + \lambda_D p_D(w|h) + \lambda_S p_S(w|h)$$

Note that, in the case where no document is retrieved for a given video, the corresponding λ_D is constrained to 0, as done in [15] for slides.

4 Experiments and Results

In this section we compare our approach with a strong baseline language model computed with out-of-domain corpora (Table 1) and in-domain corpus (Table 2). The baseline is tested against three systems: a system adapted with documents, a system adapted with slides, and a mixture thereof. In all cases, two acoustic models are used: a Deep Neural Networks model [19] with 4 hidden layer and a classical HMM with fCMLLR.

4.1 Corpora

Several corpora were used to build the baseline. Regarding the out-of-domain corpora, Table 1 summarises their main statistics. As for the in-domain corpus, we used the poli[Media] corpus, created by manually transcribing a number of video lectures from the Spanish poli[Media] repository for training, adaptation and internal evaluation as part of the **transLectures** project. Details of this corpus are given in Table 2.

Table 1. Main statistics of the out-of-domain corpora used by the baseline 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
Google Ngram	–	45 360M	292K

In order to conduct experiments with mixed models (slides and documents), it is necessary to have the text from the slides. However, a correct transcription of the slides (pdf) is not usually available and the text must be extracted automatically using OCR. These OCR slides have been recognised using *Tesseract* OCR tool [21] and applying various preprocessing and postprocessing steps. The

Table 2. Main statistics of the poli[Media] corpus.

	Videos	Time (h)	# sentences	# words	Vocabulary
Train	655	96	41.5K	968K	28K
Dev	26	3.5	1.4K	34K	4.5K
Test	23	3	1.1K	28.7K	4K

Table 3. poli[Media] slides details.

	# slides	# words	Vocabulary
Dev	107	17.4K	3.9K
Test	363	16.4K	3.1K

final WER of the OCR slides is 40%. Table 3 contains some statistics about the slides.

Regarding the downloaded documents, we performed experiments with models trained with up to 5, 10 and 20 documents per video. As we will see in Section 4.4, the extended search reports significantly better results than the exact search for 5 documents. So for 10 and 20 documents we only considered the extended search. Details of the retrieved documents are depicted in Table 4. As we mentioned, it is possible that the paper the lecture is based on is among the documents downloaded. However, this is not likely to happen in the repository used in this work.

Table 4. Statistics of downloaded documents for poli[Media] development and test sets.

			# documents	# words	Vocabulary
5 docs	Exact search	dev	96	1.3M	40K
		test	102	1.2M	41K
	Extended search	dev	130	2.0M	55K
		test	115	1.4M	42K
10 docs	Extended search	dev	260	5.2M	91K
		test	230	2.7M	65K
20 docs	Extended search	dev	515	9.0M	128K
		test	459	6.4M	104K

4.2 Acoustic Models

The language model adaptation techniques are tested with different acoustic models (AMs): standard HMM and deep neural network (DNN). In both cases, the software used to train the systems is the **transLectures** -UPV toolkit [5, 8], and the data used to train the systems is the poli[Media] corpus training set described in Table 2.

The HMMs are based on triphonemes, modelled with a 3-state left-to-right topology. A decision tree based state-tying is applied, resulting in a total of 5039 triphone states. Each triphoneme was trained for up to 128 mixture components per Gaussian and 4 iterations per mixture. Moreover, in order to reduce the speaker variability, fCMLLR was applied.

Regarding the DNN, we used the hybrid approach proposed by [19]; the so-called CD-DNN-HMM. In this case, we need to train a classical HMM system in order to provide an accurate forced alignment of the input features at triphone state (senone) level. Subsequently, a Deep Neural Network is trained in two steps: pre-training and fine-tuning. With respect to its topology, the output layer was set as large as the number of phonetic targets derived during the previous forced alignment, in this case 5039 classes. After trying several networks configurations, 4 hidden layers of 3000 units each was found to provide the best performance. Finally, the Gaussian mixtures of the acoustic model are replaced with the neural network state posterior probabilities.

4.3 Language Models

As regards the language model, we computed the baseline model as discussed in Section 3, interpolating several individual language models trained on the corpora described in Section 4.1. For each out-of-domain corpora we trained a 4-gram language model with the SRILM [22] toolkit . The individual 4-gram models were smoothed with the modified Kneser-Ney absolute interpolation method [13, 10]. Finally, the training set of poli[Media] was used as the in-domain corpus. For the vocabulary, we obtained a base vocabulary using 200K words over all the out-of-domain corpora plus the in-domain vocabulary, resulting in a 205K words vocabulary.

The vocabulary of the adapted models is built extending the base vocabulary with the words in the slides and/or the documents. In the case of the documents, the vocabulary extension will result in much larger vocabularies. From here on, we consider the standard full version of the vocabulary, and a restricted version in which only those words that occur more than three times in the documents are added.

4.4 Experiments

First we run experiments to assess whether the exact or the extended search is better for querying documents. For these experiments the number of documents per video is set to 5. Table 5 depicts these results, in which it is observed that document adaptation significantly improves the baseline results independently of the AM used. The extended search obtains better results than the exact search where the smaller amount of documents retrieved leads to slightly higher WER values. Constraining the vocabulary also results in higher error rates.

After setting the retrieval technique to extended search, we assessed the impact of the number of documents retrieved when using up to 10 and 20 documents, instead of 5. In Table 6 significant improvement is observed for all AMs

Table 5. WER (%) on the poli[Media] corpus for the adapted models with 5 documents retrieved per video.

Language Model	Acoustic Model			
	HMM		CD-DNN-HMM	
	Dev	Test	Dev	Test
Baseline (BL)	20.4	21.8	14.3	15.7
BL + Exact search	20.3	20.7	14.2	14.6
+ Restricted voc	20.2	20.8	14.1	14.8
BL + Extended search	19.8	20.6	14.0	14.4
+ Restricted voc	19.7	20.6	13.9	14.4

when using 20 documents. Note that the improvements of up to 10.8% relative WER, depicted in Table 6, can be used to effectively adapt language models for video lectures in those scenarios where no other resources but the lecture title is available.

Table 6. WER (%) on the poli[Media] corpus dev and test sets for the adapted models with 10 and 20 documents retrieved by extended search.

Language Model	Acoustic Model			
	HMM		CD-DNN-HMM	
	Dev	Test	Dev	Test
Baseline (BL)	20.4	21.8	14.3	15.7
BL + 10 Documents	19.6	20.6	13.8	14.4
+ Restricted voc	19.5	20.5	13.8	14.4
BL + 20 Documents	19.6	20.0	13.8	14.2
+ Restricted voc	19.5	19.9	13.8	14.0

In cases where slides are available, not only it is possible to perform adaptation by using either the documents or the slides [15], but also a combination of these two resources. These combined results are summarised in Table 7. It is observed that the inclusion of documents significantly improves the results of all the previous systems (adapted or not) where documents were not used. It is also interesting to note that the combination of slides and documents outperforms both the system without slides and the system without documents.

5 Conclusions

We have proposed a new simple yet effective method to retrieve documents from the web and use them to build adapted language models for video lecture transcription. These documents have proven to be a very valuable resource for adapting language models, obtaining a WER improvement of up to 1.9 absolute WER points (8.7 % relative) using HMM, and 1.7 points (10.8 % relative) when using DNN, with respect to a strong baseline.

Table 7. WER (%) for the adapted models with documents and slides.

Language Model	Acoustic Model			
	HMM		CD-DNN-HMM	
	Dev	Test	Dev	Test
Baseline (BL)	20.4	21.8	14.3	15.7
BL + Slides	19.8	19.4	13.8	13.8
+ Documents	18.7	18.9	13.4	13.5
+ Restricted voc	18.7	19.0	13.4	13.5

Furthermore, if we combine the document adaptation with slide adaptation the system yields improvements of 2.9 and 2.2 absolute WER points (13.3 % and 14.0 % relative) with respect to a strong baseline, depending on the acoustic model used. If instead we compare these results with the models adapted with slides only, it is observed that documents can still provide improvements of up to 1 absolute WER points for HMMs, and 0.5 for DNNs.

It is worth noting that, in general, the improvements are consistent for all proposed acoustic models, which makes us think that this kind of adaptation will provide significant improvements as the acoustic models get even better.

The documents obtained have led to significant improvements, proving that this method is a good way of retrieving documents for the purpose of adapting language models. However, in the future, we plan to compare this document retrieval method with the alternative methods proposed by ohter authors.

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 (transLectures) and ICT Policy Support Programme (ICT PSP/2007-2013) as part of the Competitiveness and Innovation Framework Programme (CIP) under grant agreement no 621030 (EMMA), the Spanish MINECO Active2Trans (TIN2012-31723) research project and the Spanish Government with the FPU scholarships FPU13/06241 and AP2010-4349.

References

1. coursera.org: Take the World's Best Courses, Online, For Free, `http://www.coursera.org/`
2. poliMedia: Videolectures from the "Universitat Politècnica de València, `http://polimedia.upv.es/catalogo/`
3. SuperLectures: We take full care of your event video recordings. `http://www.superlectures.com`
4. trans**Lectures** , `https://translectures.eu/`
5. transLectures-UPV Toolkit (TLK) for Automatic Speech Recognition, `http://translectures.eu/tlk`
6. Udacity: Learn, Think, Do., `http://www.udacity.com/`
7. Videolectures.NET: Exchange Ideas and Share Knowledge, `http://www.videolectures.net/`

8. del Agua, M.A., Giménez, A., Serrano, N., Andrés-Ferrer, J., Civera, J., Sanchis, A., Juan, A.: The transLectures UPV Toolkit. In: Proc. of IberSPEECH. p. Submitted (2014)
9. Chang, P.C., shan Lee, L.: Improved language model adaptation using existing and derived external resources. In: Proc. of ASRU. pp. 531–536 (2003)
10. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech & Language 13(4), 359–393 (1999)
11. Jelinek, F., Mercer, R.L.: Interpolated Estimation of Markov Source Parameters from Sparse Data. In: Proc. of the Workshop on Pattern Recognition in Practice. pp. 381–397 (1980)
12. Ketterl, M., Schulte, O.A., Hochman, A.: Opencast matterhorn: A community-driven open source solution for creation, management and distribution of audio and video in academia. In: Proc. of ISM. pp. 687–692 (2009)
13. Kneser, R., Ney, H.: Improved Backing-off for M-gram Language Modeling. In: Proc. of ICASSP. pp. 181–184 (1995)
14. Lecorv, G., Gravier, G., Sbillot, P.: An unsupervised web-based topic language model adaptation method. In: Proc. of ICASSP 2008. pp. 5081–5084 (2008)
15. Martínez-Villaronga, A., del Agua, M.A., Andrés-Ferrer, J., Juan, A.: Language model adaptation for video lectures transcription. In: Proc. of ICASSP. pp. 8450–8454 (2013)
16. Munteanu, C., Penn, G., Baecker, R.: Web-based language modelling for automatic lecture transcription. In: Proc. of INTERSPEECH. pp. 2353–2356 (2007)
17. Rogina, I., Schaaf, T.: Lecture and presentation tracking in an intelligent meeting room. In: Proc of. ICMI. pp. 47–52 (2002)
18. Schlippe, T., Gren, L., Vu, N.T., Schultz, T.: Unsupervised language model adaptation for automatic speech recognition of broadcast news using web 2.0 pp. 2698–2702 (2013)
19. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. of ASRU. pp. 24–29 (2011)
20. Silvestre, J.A., et al.: translectures. In: Proc. of IberSPEECH 2012. pp. 345–351 (2012)
21. Smith, R.: An overview of the tesseract ocr engine. In: Proc. of ICDAR. pp. 629–633. ICDAR '07 (2007)
22. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. of ICSLP. pp. 901–904 (2002)
23. Tsiartas, A., Georgiou, P., Narayanan, S.: Language model adaptation using www documents obtained by utterance-based queries. In: Proc. of ICASSP. pp. 5406–5409 (2010)