Different Contributions to Cost-Effective Transcription and Translation of Video Lectures

Joan Albert Silvestre-Cerdà, Alfons Juan, and Jorge Civera

Machine Learning and Language Processing (MLLP) Research Group
Departament de Sistemes Informàtics i Computació (DSIC)
Universitat Politècnica de València (UPV)
{jsilvestre,ajuan,jcivera}@dsic.upv.es
http://www.mllp.upv.es

Abstract. In recent years, on-line multimedia repositories have experiencied a strong growth that have made them consolidated as essential knowledge assets, especially in the area of education, where large repositories of video lectures have been built in order to complement or even replace traditional teaching methods. However, most of these video lectures are neither transcribed nor translated due to a lack of cost-effective solutions to do so in a way that gives accurate enough results. Solutions of this kind are clearly necessary in order to make these lectures accessible to speakers of different languages and to people with hearing disabilities, among many other benefits and applications.

For this reason, the main aim of this thesis is to develop a cost-effective solution capable of transcribing and translating video lectures to a reasonable degree of accuracy. More specifically, we address the integration of state-of-the-art techniques in Automatic Speech Recognition and Machine Translation into large video lecture repositories to generate high-quality multilingual video subtitles without human intervention and at a reduced computational cost. Also, we explore the potential benefits of the exploitation of the information that we know a priori about these repositories, that is, lecture-specific knowledge such as speaker, topic or slides, to create specialised, in-domain transcription and translation systems by means of massive adaptation techniques.

The proposed solutions have been tested in real-life scenarios by carrying out several objective and subjective evaluations, obtaining very positive results. The main outcome derived from this multidisciplinary thesis, *The transLectures-UPV Platform*, has been publicly released as an open-source software, and, at the time of writing, it is serving automatic transcriptions and translations for several thousands of video lectures in many Spanish and European universities and institutions.

Keywords: Audio Segmentation, Automatic Speech Recognition, Machine Translation, Language Modelling, Massive Adaptation, Intelligent Interaction, Multilingualism, Acessibility, Education, Technology Enhanced Learning, Video Lectures, Recommender Systems.

1 Introduction and Motivation

In recent years, the growth of the world wide web has offered a great opportunity for academic institutions to enhance the learning process of their students with digital media contents that complement and even replace conventional teaching methods such as face-to-face lectures [4]. Indeed, these digital resources are being incorporated into existing university curricula around the world with enthusiastic response from students [5].

In this sense, on-line multimedia repositories have become established as fundamental knowledge assets, specially in those specialised on serving on-line video lectures. These repositories are being built on the back of on increasingly available and standardised infrastructure [2, 1]. A well-known example of this is VideoLectures.NET [7], a free and open access web portal that has already published more than 20.000 educational videos and conference recordings given by relevant world-wide researchers and professors.

However, the utility of these audiovisual assets could be further extended by adding subtitles that can be exploited to incorporate added-value functionalities such as searchability, accessibility, and discovery of content-related videos, among others. In fact, most of the video lectures available in large university repositories are neither transcribed nor translated, despite the clear need to make their content accessible to speakers of different languages and people with disabilities [8]. Also, the subtitles can be used to develop advanced educational functionalities like content summarisation to assist student note-taking [3].

For this reason, this thesis[1] aims to developing a cost-effective solution that can do so to a reasonable degree of accuracy. More specifically, we propose the integration of state-of-the-art techniques in ASR and MT into large video lecture repositories to generate high-quality multilingual video subtitles without human intervention and at a reduced computational cost. Of course, although it would be the most desirable scenario, we do not expect to produce error-free transcriptions and translations, and, for this reason, we also aim to create efficient and ergonomic tools to allow the review of transcription and translations under a collaborative-editing scenario.

2 Thesis Overview

In this section we give a brief summary of this work, with references to the corresponding chapters of the document. We want to highlight that this is a multidisciplinary thesis, since it provides scientific contributions to many different research and technological areas: Statistical Machine Translation, Automatic Speech Recognition, Audio Segmentation and Recommender Systems.

The generation of multilingual subtitles for video lectures involves the consecutive application of both technologies: on a first step, ASR to generate speech

[1] Thesis document can be found on-line here:
http://hdl.handle.net/10251/62194

transcripts from the lecturer, and on a second step, MT to translate these transcripts into other languages. Assuming that recognition errors are likely to arise on the first step, and that these errors are propagated to the second step, we need to ensure that our MT technology yields good quality translations regardless the input source language text. In this line, the Chapter 3 of this thesis (*Explicit Length Modelling for Statistical Machine Translation*) discusses how length information is modelled in state-of-the-art Statistical MT (SMT) systems, proposing a novel approach in which length variability of word sequences among source and target languages is explicitly taken into account when translating sentences from one language to another.

It is important to note that ASR systems are the bottleneck of the generation of multilingual subtitles: MT systems can be parallelized in order to reduce the overall computation time, however, they cannot start generating translations until the speech transcript is available. Consequently, ASR systems must be boosted as much as possible without compromising significantly the quality of their outputs. Since the temporal cost of generating an automatic transcription strongly depends on the length of the input audio signal, a simple way to speed up the whole process is to apply a previous step in which the input audio signal is split into homogeneous acoustical regions to detect speech segments, and delivering these isolated speech segments to the ASR system. Furthermore, transcription quality may be improved due the fact that the ASR system does not have to deal with non-speech segments, which are usually but erroneously transcribed by their closest phonetic transcripts. This process of segmenting the input audio signal to detect speech regions is addressed by Audio Segmentation (AS) systems. Since their application is motivated to hasten the overall process of transcribing a video lecture, these systems must be as fast as possible. In Chapter 4 (*Efficient Audio Segmentation for Speech Detection*), we present a simple yet powerful approach for Audio Segmentation that meets our purposes.

Despite state-of-the-art ASR and MT systems have been proved to yield accurate speech transcriptions in most cases, their outputs can be greatly improved through the application of massive adaptation techniques. Massive adaptation refers to process of exploiting the wealth of knowledge available in video lecture repositories, that is, lecture-specific knowledge, such as speaker, topic and slides, to create a specialised, in-domain transcription or translation system. A system adapted using this knowledge is therefore likely to produce a far better ASR and MT output than a general-purpose system. These techniques are reviewed and tested in Chapters 5 and 8. In addition, a novel approach to topic adaptation for ASR systems using lecture-related text documents downloaded from the internet is proposed and evaluated in Chapter 7 (*Language Model Adaptation Using External Resources for Speech Recognition*).

As for the integration of ASR and MT technologies into large video lecture repositories, it is needed to design and develop a system architecture capable of blending the existing workflows in remote repositories with transcription and translation processes, as well as to engage users and authors into subtitle review processes. This architecture should also facilitate the incorporation of techno-

logical upgrades into ASR and MT systems to allow a progressive refinement of the overall transcription and translation quality of the repository. Indeed, Chapter 5 (*The transLectures-UPV Platform*) introduces a novel system architecture that satisfies these requirements. The implementation of this architecture, called *The transLectures Platform* (TLP), was tested under a real-life environment, as it was deployed over the poliMedia [6] official video lecture repository of the Universitat Politècnica de València (UPV). The proposed system architecture is refined and extended in Chapter 8 (*Transcription and Translation Platform*).

Users that visit multimedia repositories are often overwhelmed by the vast amount of choices that these sites offer. They may not have the time or knowledge to find the most suitable videos for their needs. However, having all video lectures transcribed with our proposed solutions, we can generate accurate semantic representations of every lecture that can be used to recommend lectures to users based on their interests. Hence, Chapter 6 (*Recommender Systems for Online Learning Platforms*) describes a novel Recommender System (RS) that exploits lecture transcriptions plus other related text resources to provide better recommendations to users. This RS was developed, deployed and tested in the VideoLectures.NET web site.

3 Scientific and Technological Goals

The main scientific and technological goals pursued in this work are the following:

- Propose an approach to explicit length modelling for SMT.
- Develop an efficient Audio Segmentation system to speed up ASR systems.
- Study how massive adaptation techniques can lead to better results in transcription and translation of video lecture repositories.
- Propose alternative topic adaptation techniques for ASR.
- Develop a system architecture capable of integrating ASR and MT technologies into video lecture repositories.
- Develop appropriate solutions to enable users to edit transcriptions and translations with ease and relatively small effort under a collaborative scenario.
- Design a Recommender System capable of exploiting speech transcriptions to provide accurate recommendations to users in video lecture on-line repositories.
- Evaluate these contributions in real-life scenarios.
- Make public releases of the software tools developed in this thesis.

4 Global Conclusions

In this section we draw some conclusions of this thesis in the light of the experimental results obtained in each area.

Firstly, in Chapter 3 are proposed two novel explicit conditional phrase length models for SMT. These phrase-length models were integrated in a state-of-the-art log-linear SMT system as additional feature functions, providing in most

cases a systematic and statistically significant boost of translation quality on unrelated language pairs.

Secondly, in Chapter 4 is described an efficient AS system clearly inspired in GMM-HMM-based ASR that exhibits excellent performance detecting speech segments at near real-time speeds. This system was submitted to the Audio Segmentation competition of the *Albayzin 2012 Evaluations* within the *IberSpeech 2012* conference, achieving the 2nd place in the global standings, very close to the winner system.

Thirdly, Chapter 5 presents a system architecture that allows the integration of ASR and MT technologies into video lecture repositories. Its implementation, *The transLectures-UPV Platform*, was integrated into the UPV's poliMedia [6] repository on production. Preliminary results on automatic and human evaluations suggested that the delivered transcriptions and translations were of an acceptable quality though had to be improved, and that the provided tools to edit subtitles were comfortable, productive, and very easy to use.

Then, Chapter 6 describes a lecture Recommender System that exploits automatic speech transcriptions of video lectures to zoom in on user interests at a semantic level. This RS was implemented and deployed over the VideoLectures.NET production website. Preliminary, quantitative-based metrics computed in comparison with the previously existing RS were not encouraging, suggesting that qualitative-based metrics must be explored in order to fairly compare both systems.

Next, Chapter 7 proposed an effective method to retrieve documents from the web and use them to build topic-adapted language models for video lecture transcription. The application of this technique under a solid experimental setting reported systematic and significant WER improvements of above 10%.

Finally, Chapter 8 presented the latest version of the *transLectures-UPV Platform* as an evolution of the first version presented in Chapter 5. This software was publicly released as open-source software[2]. Similarly, the preliminary automatic and user evaluations in the poliMedia repository presented in Chapter 5 were extended, showing how the overall transcription and translation quality of a media repository can be enhanced over time by means of introducing technological upgrades into the ASR and MT systems integrated into TLP. Also, we have proven that massive adaptation techniques provide systematic and significant improvements in transcription and translation quality. Furthermore, user evaluations reflected that using automatic transcriptions or translations as a start point to generate perfect subtitles saves about two thirds of the total time that would be needed to do that from scratch.

[2] The latest version of TLP can be downloaded here:
http://www.mllp.upv.es/tlp

5 Achievements and contributions

The main contributions of this thesis are the following:

- An explicit conditional phrase length modelling approach for SMT that provide systematic and significant improvements over strong baselines for different language pairs.
- A simple yet powerful and efficient approach for AS to detect speech segments in audio signals.
- A free and open-source solution to integrate ASR and MT technologies into large video lecture repositories, capable of generating cost-effective high-quality multilingual subtitles.
- An extensive evaluation of several ASR and MT systems in different languages to gauge the positive effect of massive adaptation techniques in video lecture repositories.
- A new approach to video lecture recommendation for content-based RS using automatic speech transcripts.
- A new language model adaptation technique for ASR that yields significant WER improvements over solid baselines.

The scientific impact of this thesis can be gauged through the 9 publications that were derived from this work. More precisely, this thesis yielded 4 articles in national conferences (*IberSpeech 2012*, *IberSpeech 2014*), 3 articles in international conferences (*IbPRIA 2011*, *IEEESMC 2013*, *EC-TEL 2015*), and 2 articles in JCR journals (*Pattern Recognition*, *Speech Communication*).

In addition, we want to highlight that the *transLectures-UPV Platform* (TLP) software, at the time of writing, is running in production for the UPV's Media portal[3] (formerly poliMedia), generating and serving automatic multilingual subtitles for more than 20.000 video lectures to a potential audience of approximately 36.000 students and 2.800 university lecturers and researchers. TLP is also behind the MLLP's Transcription and Translation Platform[4], a cloud service created and hosted by the Machine Learning and Language Processing (MLLP) research group that is offering automatic subtitling services to several worldwide institutions.

Acknowledgments. Work supported by the Spanish Goverment under the FPU scolarship (AP2010-4349), and under the iTrans2 (TIN2009-14511), erudito.com (TSI-020110-2009-439) and Active2Trans (TIN2012-31723) research projects. Also supported by the EC (FEDER/FSE) under the transLectures (FP7-ICT-2011-7-287755) and EMMA (ICT-PSP/2007-2013-621030) projects, and by the Spanish (MINECO/FEDER) research project MORE (TIN2015-68326-R).

[3] http://media.upv.es
[4] https://ttp.mllp.upv.es

References

1. Coursera: Take the World's Best Courses, Online, For Free. http://www.coursera.org
2. edX: Access to free education for everyone. http://www.edx.org
3. Glass, J., et al.: Recent progress in the MIT spoken lecture processing project. In: Proc. of Interspeech 2007. vol. 3, pp. 2553–2556 (2007)
4. Ross, T., Bell, P.: "No significant difference" only on the surface. International Journal of Instructional Technology and Distance Learning 4(7), 3–13 (2007)
5. Soong, S.K.A., Chan, L.K., Cheers, C., Hu, C.: Impact of video recorded lectures among students. Who's learning pp. 789–793 (2006)
6. Universidad Politècnica de València: The polimedia repository. http://media.upv.es
7. VideoLectures.NET: Exchange ideas and share knowledge. http://www.videolectures.net
8. Wald, M.: Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. Interactive Technology and Smart Education 3(2), 131–141 (2006)