Efficient generation of high-quality multilingual subtitles for video lecture repositories

Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró, and Alfons Juan

Universitat Politècnica de València,
Camino de Vera s/n, 46022 Valencia, Spain
{jvalor,jsilvestre,jcivera,ajuan}@dsic.upv.es
turro@cc.upv.es

Abstract. Video lectures are a valuable educational tool in higher education to support or replace face-to-face lectures in active learning strategies. In 2007 the Universitat Politècnica de València (UPV) implemented its video lecture capture system, resulting in a high quality educational video repository, called poliMedia, with more than 10.000 mini lectures created by 1.373 lecturers. Also, in the framework of the European project transLectures, UPV has automatically generated transcriptions and translations in Spanish, Catalan and English for all videos included in the poliMedia video repository. transLectures's objective responds to the widely-recognised need for subtitles to be provided with video lectures, as an essential service for non-native speakers and hearing impaired persons, and to allow advanced repository functionalities. Although high-quality automatic transcriptions and translations were generated in transLectures, they were not error-free. For this reason, lecturers need to manually review video subtitles to guarantee the absence of errors. The aim of this study is to evaluate the efficiency of the manual review process from automatic subtitles in comparison with the conventional generation of video subtitles from scratch. The reported results clearly indicate the convenience of providing automatic subtitles as a first step in the generation of video subtitles and the significant savings in time of up to almost 75% involved in reviewing subtitles.

Keywords: video lecture repositories, automatic speech recognition, machine translation, efficient video subtitling

1 Introduction

Video lectures are fast becoming an everyday educational resource in higher education used to supplement and complement face-to-face lectures [6], and are being incorporated into existing university curricula around the world with enthusiastic response from students [8].

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 university large repositories are neither transcribed nor translated, despite the clear need to make their content accessible to speakers of different languages and people with disabilities ([10]). Also, the subtitles can be used to develop advanced educational functionalities like content summarisation to assist student note-taking ([2]). For this reason, it is important to develop a cost-effective solution that can do so to a reasonable degree of accuracy. In this work, we propose the application of state-of-the-art techniques in Automatic Speech Recognition (ASR) and Statistical Machine Translation (SMT) to generate high-quality video subtitles under the supervision of lecturers involved in this process.

This paper is organised as follows. Next section is devoted to present poliMedia [5], a video lecture repository of the Universitat Politènica de València (UPV), and transLectures [7], the European project under which framework ASR and SMT technology was developed to be applied for video subtitling. Section 3 describes the user evaluation campaign performed at UPV to review video transcription and translations, and its results. Finally, conclusions are drawn in Section 4.

2 poliMedia and transLectures

In 2007 the UPV implemented its poliMedia lecture capture system for the cost-effective creation and publication of quality educational video content. It now has a collection of over 10,000 video objects created by 1373 lecturers, in part incentivised by the *Docència en Xarxa* (DeX) action plan. poliMedia is primarily designed to allow UPV lecturers to record pre-prepared mini lectures for use by students in supplement to the traditional live lecture. For the most part they consist of concise overviews of a given topic and have a typical duration of around ten minutes.

In addition, in 2011 the UPV embarked on the EU-subsidised project transLectures to generate automatic subtitles in Spanish, English and Catalan for all videos in the poliMedia repository. These poliMedia videos were automatically transcribed using the TLK toolkit for ASR [1], which consists of a set of tools that allows acoustic model training and speech decoding. The language model of these ASR systems was a linear mixture trained on the poliMedia manual transcriptions along with other external resources. Then, poliMedia videos were automatically translated into other languages (Spanish, English and Catalan) using the well-known SMT toolkit Moses [3] trained on large parallel external resources. Also, we apply adaptation techniques based on a selection of topic-specific sentences related to the topic of the videos being translated.

A conventional post-editing protocol was adopted for both, transcription and translation review using a web interface called transLectures player [9], an innovative web player with editing capabilities. As shown in Figure 1, the player plays a video lecture and its corresponding transcription in synchrony, allowing the user to correct the transcription while watching the video.

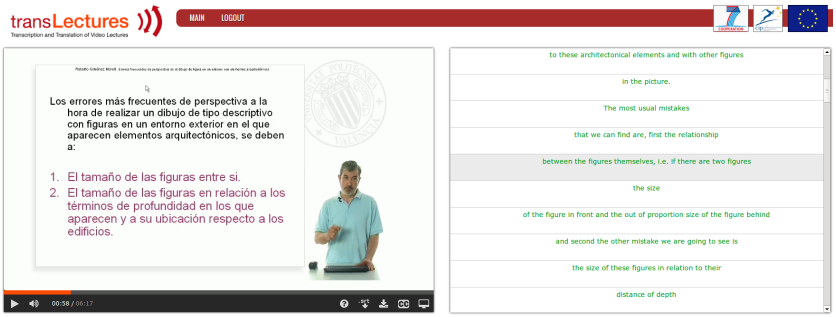

Fig. 1. transLectures web player for the review of a video transcription.

Similarly, an interface for the review of video translations was designed in order to show the original transcription, the corresponding translation and the video in synchrony (see Figure 2). It is important to note that the transLectures player is able to log precise user interaction statistics that will be useful to analyse users' behaviour or develop a user model.

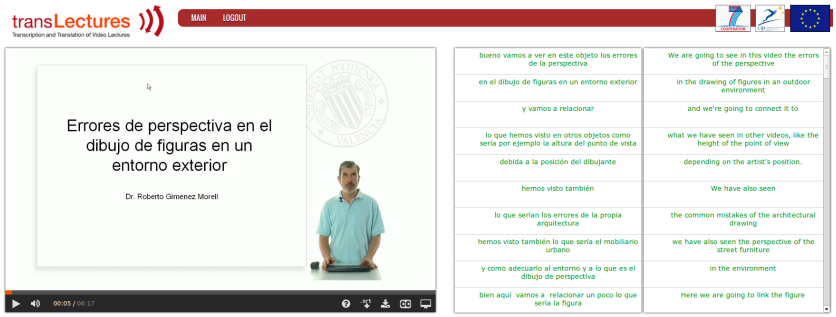

Fig. 2. transLectures web player for the review of a video translation.

3 User Evaluations

In this section we describe user evaluations carried out under UPV 2013-2014 DeX programme. User evaluations are considered from two viewpoints, the transcription or translation quality of the subtitles automatically generated and how much time takes for lecturers to review these automatic subtitles. Lecturers' background is diverse but mainly grounded on engineering studies, however a few lecturers were related to other areas, such as business management, social

science, and biology. Lecturers committed to reviewing the automatic transcriptions of five poliMedia videos or the automatic translations of three poliMedia videos. The topics of the video lectures reviewed had also different topics, but we had not found any influence in the results arising from this variety.

Transcription and translation errors are automatically gauged according to Word Error Rate (WER) and Translation Edit Rate (TER), respectively. These measures are widely accepted in the ASR and SMT fields. WER is the ratio, expressed as a percentage, of the number of basic word editing operations required to convert the automatic transcription into the correct reviewed transcription, and the total number of words in the reviewed transcription. Similarly, TER is the ratio, expressed as a percentage, of the number of edit operations (including phrasal shifts) required to convert the automatic translation into the correct reviewed translation, and the total number of words in the reviewed translation. So if, for example, a user needs to correct 20 words in a transcription (or translation) containing 100 words, the WER (or TER) is 20%.

The time required for lecturers to review automatic transcriptions or translations is computed in terms of Real Time Factor (RTF). RTF is the ratio between the time devoted to reviewing the transcription or translation of a video and the duration of said video. So if, for example, a video lasts 20 minutes and its review takes, by way of example, one hour, then the RTF for this video would be 3.

Next sections report the results obtained on the review of Spanish, English and Catalan transcriptions and Spanish into English translations.

3.1 Review of Spanish Transcriptions

Spanish automatic transcriptions were reviewed by 39 lecturers accounting for 18.3 hours (135 videos). A simple linear regression model $RTF = \beta \cdot WER$ was fitted to all collected data that resulted in a statistically significant (β=0.184, Sig< $2.2{*}10^{-16}$) dependency between RTF and WER. On average, WER was as low as 12 WER points (high transcription quality) and RTF was 2.7. In practical terms, the review of an average length poliMedia video of 10 minutes takes approximately 100 minutes (non-expert users usually needs 10 RTF [4]) to do it from scratch, but using the post-editing protocol with our high-quality ASR system, the review time would be less than 30 minutes.

3.2 Review of English Transcriptions

In English, 57 video transcriptions accounting for 7.9 hours were reviewed by 12 non-native volunteers. The process was the same as in the Spanish review.

As in Spanish, there is a statistically significant linear dependency between WER and RTF (β=0.168, Sig< $2.2{*}10^{-16}$). The average RTF was 6.2 still far below 10 RTF achieved by non-expert users. The reason behind this higher RTF was the transcription quality, since the ASR English system trained attained 36.0 WER points in this evaluation. Since this user evaluation was performed, the English ASR system has notably improved up to reaching 21.4 WER points.

In terms of more qualitative feedback, volunteers valued the simplicity and efficiency of the player interface. Volunteers agreed that the quality of the English automatic transcription must increase to reduce review time. In summary, results were largely positive and, as desired, volunteers prefered this system than transcribing videos from scratch.

3.3 Review of Catalan Transcriptions

The review of Catalan transcriptions was carried out by 5 lecturers that reviewed 19 video transcriptions accounting for 1.5 hours. The review process was the same described in the Spanish and English reviews. It is important to note that, in this case, Catalan transcriptions had a significantly lower quality than those of Spanish, since this transcription system was at the first stages of development.

The average RTF was 5.6, while the average WER of automatic transcriptions was 40.4. As in Spanish and English, there is a clearly statistically significant linear dependency between WER and RTF (β=0.138, Sig$< 2.2*10^{-16}$). Lecturers expose the idea that the transcription quality must be improved.

Fortunately, the Catalan transcription system has significantly improved since these evaluations were carried out, and nowadays Catalan WER figures are 17.4 points, decreasing the review time to attain correct transcriptions.

3.4 Review of Spanish into English Translations

Ten lecturers took part in the review of Spanish into English translations accounting for 13 video translations fully reviewed (about 2.1 hours of video). The average RTF was 12.2, while the average TER was 41.9. If we compare this RTF to that achieved by manual translations (about 30 RTF), we observe a significant relative decrease in user effort. As with WER, there is a statistically significant linear dependency between RTF and TER (β=0.255, Sig=$3.73*10^{-7}$).

As in transcription evaluations, generally speaking, lecturers were satisfied with the interface, however they requested that the player stopped at the end of each segment to have more time to review the translation. This request was not necessary while reviewing transcription, since the cognitive load is notably lower than when translating. Finally, lecturers demanded higher translation quality.

4 Conclusions

This work describes the efficient generation of high-quality multilingual subtitles for the poliMedia video repository. Our results on user evaluations with lecturers reflect significant reductions in review time to generate subtitles. As expected, review times depend on the automatic transcription or translation quality and user expertise. However, WER figures achieved by our ASR systems provided automatic transcriptions whose review was already more efficient than transcribing from scratch with significant effort reductions about 70%, 40% and 35% for

Spanish, English and Catalan, respectively. The Spanish-English translation review task also proved to save time compared to translating from scratch. No evidence exists regarding the difference in quality between subtitles (transcriptions and translations) generated after reviewing automatic subtitles and those that were obtained from scratch.

In practical terms, the review of transcription and translation for an average length poliMedia video of 10 minutes takes approximately 400 minutes (approximately 10 RTF for transcription plus 30 RTF for translation) to do it from scratch. Using transLectures technology, the user would need about 30 minutes (2.7 RTF) to review the transcription plus about 120 minutes (RTF 12.2) to correct the translation, that is, 150 minutes. This means a user time-saving of approximately two thirds with respect to do it from scratch.

The system presented in this work is currently in production and available at the poliMedia site [5]. Nowadays, all video lectures generated by UPV lecturers are automatically transcribed and translated into Spanish, English and Catalan, being available for review using the transLectures player.

Acknowledgments

The research leading to these results has received funding from the European Union FP7/2007-2013 under grant agreement no 287755 (transLectures) and ICT PSP/2007-2013 under grant agreement no 621030 (EMMA), and the Spanish MINECO Active2Trans (TIN2012-31723) research project.

References

1. del Agua, M.A., et al.: The transLectures-UPV toolkit. In: Proc. of IberSpeech (2014)
2. Glass, J., et al.: Recent progress in the MIT spoken lecture processing project. In: Proc. of Interspeech 2007. vol. 3, pp. 2553–2556 (2007)
3. Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: Proc. of ACL. pp. 177–180 (2007)
4. Munteanu, C., et al.: Improving ASR for lectures through transformation-based rules learned from minimal data. In: Proc. of ACL-AFNLP. pp. 764–772 (2009)
5. poliMedia: polimedia platform. `http://media.upv.es/` (2007)
6. Ross, T., Bell, P.: "No significant difference" only on the surface. International Journal of Instructional Technology and Distance Learning 4(7), 3–13 (2007)
7. Silvestre, J.A., et al.: translectures. In: Proc. of IberSPEECH 2012 (2012)
8. 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)
9. Valor Miró, J.D., et al.: Integrating a State-of-the-Art ASR System into the Opencast Matterhorn Platform. In: Advances in Speech and Language Technologies for Iberian Languages, CCIS, vol. 328, pp. 237–246. Springer (2012)
10. 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)