Jornadas In–Red 2014
Universitat Politècnica de València
Doi: **********

Evaluación del proceso de revisión de transcripciones automáticas para vídeos poliMedia.*

Juan Daniel Valor Miró[1], Rachel Nadine Spencer[1], Alejandro Pérez González de Martos[1], Gonçal Garcés Díaz-Munío[1], Carlos Turró[1], Jorge Civera[1] y Alfons Juan[1]

[1]Universitat Politècnica de València

Abstract

Video lectures are a tool of proven value and wide acceptance in universities that are leading to platforms like poliMedia. transLectures is a European project that generates automatic high-quality transcriptions and translations for the poliMedia platform, and improve them by using massive adaptation and intelligent interaction techniques. In this paper we present the evaluation with lecturers carried out under the Docència en Xarxa 2012-2013 call, with the aim to study the process of supervise transcriptions, compared with to transcribe from scratch.

***Keywords**: ASR, transcriptions, evaluations, online teaching, poliMedia.*

Resumen

Los vídeos docentes son una herramienta de demostrada utilidad y gran aceptación en el mundo universitario que están dando lugar a plataformas como poliMedia. transLectures es un proyecto europeo que genera transcripciones y traducciones automáticas de alta calidad para la plataforma poliMedia, mediante técnicas de adaptación masiva e interacción inteligente. En este artículo presentamos la evaluación con profesores que se realizó en el marco de Docència en Xarxa 2012-2013, con el objetivo de estudiar el proceso de supervisión de transcripciones, comparándolo con la obtención de la transcripción sin disponer de una transcripción automática previa.

***Keywords**: ASR, transcripciones, evaluaciones, docencia en red, poliMedia.*

*The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755.

1 Introducción

Los vídeos docentes son una herramienta de demostrada utilidad y gran aceptación en el mundo universitario que están dando lugar a plataformas como poliMedia (poliMedia 2007), plataforma de la Universitat Politècnica de València (UPV) que que permite la creación, publicación y difusión de este tipo de contenidos. En 2011 la UPV inició el proyecto europeo transLectures (Silvestre y col. 2012), con el objetivo de generar transcripciones y traducciones de forma automatizada en español, catalán e inglés para todo el repositorio de poliMedia, mediante el uso de técnicas estadísticas de aprendizaje automático (Valor Miró y col. 2012).

Sin embargo, a pesar de que la calidad de la transcripción automática se encuentra en un 80 % de acierto en promedio, estas transcripciones automáticas poseen errores inherentes al proceso de obtención utilizado, por lo que es necesaria una revisión manual posterior para garantizar una calidad adecuada para el alumnado. El objetivo de este trabajo es evaluar este proceso de revisión manual a partir de transcripciones automáticas, para poder compararlo en términos de coste temporal con un proceso de obtención de transcripciones totalmente manual. Con este objetivo en mente, y en el marco de las ayudas de la UPV de Docencia en Red 2012-2013, pusimos en marcha un proceso de revisión de transcripciones automáticas por parte de los autores de los correspondientes vídeos poliMedia. Dichas evaluaciones se organizaron en 3 fases, con el objetivo de evaluar diversas versiones de nuestro sistema que fue refinado tras cada fase.

2 transLectures

transLectures es el acrónimo del proyecto europeo titulado "Transcription and Translation of Video Lectures" (FP7-ICT-2011-7), en el cual se aplican técnicas de reconocimiento automático del habla y de traducción automática con el objetivo de proveer transcripciones y traducciones de forma automatizada a grandes repositorios de vídeos docentes como poliMedia (Silvestre y col. 2012) o VideoLectures.NET. En este proyecto coordinado por la UPV colaboran dos socios académicos como son el *Jozef Stefan Institute* en Eslovenia y el *Rheinisch-Westfaelische Technische Hochschule* en Alemania, y tres socios industriales *Deluxe Media Europe* de Grecia, *European Media Laboratory GmbH* en Alemania y *XEROX S.A.S.* en Francia.

El objetivo científico principal de este proyecto es generar transcripciones y traducciones automáticas de alta calidad mediante técnicas de adaptación masiva e interacción inteligente. La adaptación masiva consiste en adaptar los modelos empleados en transcripción y traducción a las variables específicas de los vídeos docentes, como pueden ser el locutor y el tema. Por otro lado, la interacción inteligente consiste en intentar detectar los errores producidos en las transcripciones y traducciones, con el objetivo de presentárselos al usuario para su corrección.

3 poliMedia

poliMedia es un sistema diseñado por la UPV para la creación de contenidos multimedia como apoyo a la docencia presencial, que abarca desde la preparación del material docente hasta la distribución a través de distintos medios a los destinatarios de los mismos (poliMedia 2007). El autor es el propietario intelectual de la obra, y es la UPV la que ofrece instrumentos, materiales y técnicos para la grabación de los vídeos docentes por parte del profesorado.

Los vídeos docentes que existen en la plataforma poliMedia, siguen un formato estándar en todos los vídeos, característico de la plataforma. Se trata de una vista conjunta de profesor y pantalla (diapositivas, programas, etc) en un mismo plano de cámara que no se mueve en todo el vídeo, como se puede observar en la Figura 1.

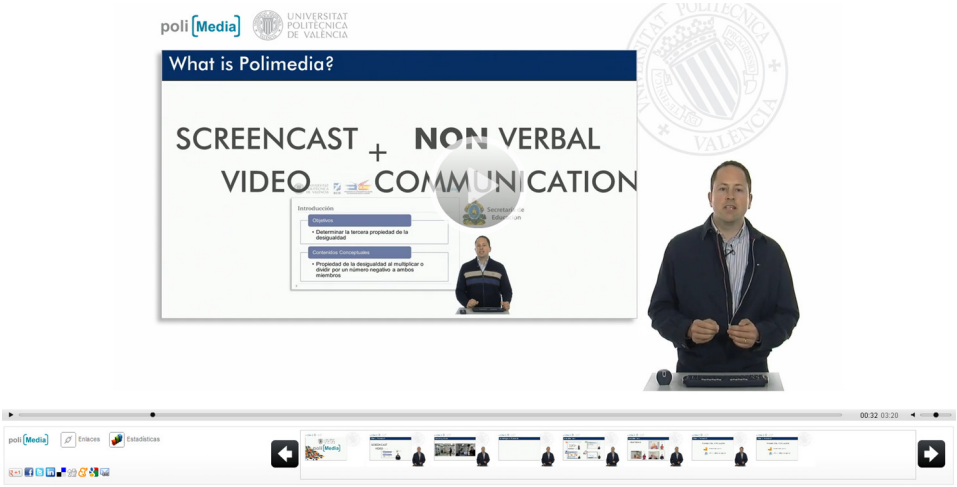

Fig. 1: Resultado final de un vídeo docente de poliMedia.

El proceso de publicación de vídeos docentes es muy simple: se graba en un estudio al profesor sobre un fondo blanco, y simultáneamente la pantalla del ordenador, que previamente habrá cargado algún tipo de recurso multimedia. Después se extrae la grabación del profesor y se mezcla con la de la pantalla según se observa en la Figura 1. Finalmente, este vídeo ya editado se distribuye a través del portal oficial de poliMedia.

Destacar que en las evaluaciones que se presentan en este artículo han participado profesores que previamente habían grabado material para el portal de poliMedia, y que se han llevado a cabo en el marco del programa *Docència en Xarxa*, que pretende incentivar entre el profesorado el uso de esta plataforma.

4 Evaluaciones

Como comentábamos, las evaluaciones se han llevado a cabo en el marco del programa *Docència en Xarxa* 2012-2013, en el que se ha invitado a los profesores a revisar las transcripciones de algunos de sus propios vídeos docentes de poliMedia. En total 27 profesores participaron en este estudio con un total de 86 vídeos docentes revisados, que fueron transcritos automáticamente utilizando el toolkit TLK (The TransLectures-UPV team 2013) desarrollado en el proyecto transLectures. Además, con el objetivo de organizar las evaluaciones, y probar nuevos modelos de interacción con el usuario, las evaluaciones fueron divididas en tres fases.

La primera fase consistió en una revisión de transcripciones generadas automáticamente basada en una interfaz web diseñada específicamente para este fin. Esta interfaz mostraba de forma sincronizada el vídeo y la transcripción para facilitar al usuario la detección de errores de transcripción. El usuario simplemente necesitaba hacer clic o presionar la tecla *Intro* sobre el segmento de transcripción errónea para poder modificarlo. Dicha interfaz puede verse en la Figura 2.

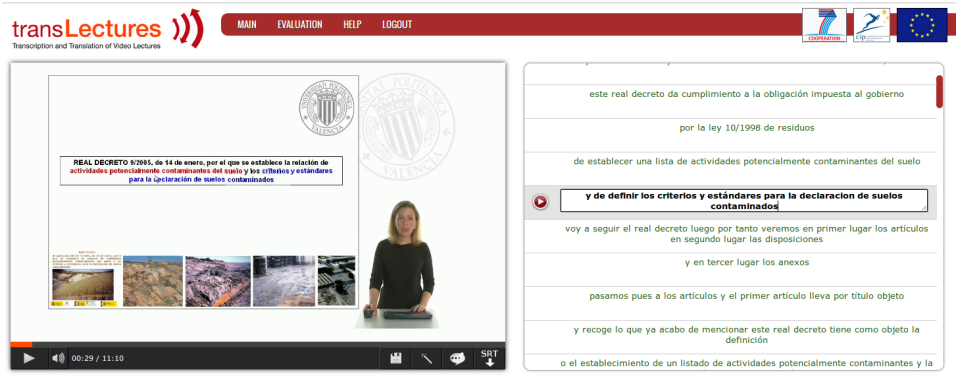

Fig. 2: Reproductor web de transLectures en modo edición manual.

En la segunda fase introdujimos un sistema basado en interacción inteligente (Serrano y col. 2013), que basándose en medidas de confianza (Sanchis, Juan y Vidal 2012) a nivel de palabra detectaba automáticamente aquellas palabras de la transcripción que probablemente fueran erróneas. De esta forma pretendíamos centrar la atención del usuario en dicha palabras que probablemente habrían sido incorrectamente reconocidas por el sistema de transcripción automática. En este caso, el sistema únicamente reproduce el segmento de audio correspondiente a estas palabras con el fin de ahorrar tiempo y esfuerzo al usuario. Sin embargo, esta detección de palabras erróneas no es perfecta, por lo que aunque la calidad de transcripción en general mejora mediante la revisión del usuario, la transcripción final sigue conteniendo errores de transcripción.

La tercera y última fase fue una combinación de las dos anteriores, diseñada para obtener transcripciones perfectas, pero aprovechando la detección eficiente de errores basada en medidas de confianza. Esta fase fue dividida en dos etapas. En la prime-

Fig. 3: Reproductor web de transLectures en modo interacción inteligente. Las palabras con baja confianza (probablemente erróneas) aparecen en rojo, y las revisadas en verde. La caja de edición puede ser expandida a izquierda y derecha pulsando en << ó >>, respectivamente.

ra etapa se mejoró la transcripción inicial mediante una revisión de transcripciones automáticas utilizando un protocolo similar al de la segunda fase. Tras esta primera revisión, el sistema de transcripción automática se volvió a entrenar incorporando las revisiones del usuario (Sanchez-Cortina y col. 2012), actualizándose seguidamente las transcripciones automáticas respetando las revisiones ya realizadas por el usuario. Finalmente se realizó una segunda etapa de revisión manual idéntica a la realizada en la primera fase. La idea es que la calidad de esta segunda transcripción sea lo suficientemente alta como para que la revisión requiera un menor tiempo, y que el tiempo total empleado en ambas etapas sea menor que el de la primera fase.

5 Resultados

Los resultados de las diferentes fases fueron recopilados en términos de *Real Time Factor* (RTF), *Word Error Rate* (WER), y una encuesta de satisfacción en escala de 0 a 10 que contempla diferentes aspectos de usabilidad y preferencias del usuario. El RTF se define como el ratio entre el tiempo empleado en la revisión y la duración del vídeo, mientras que el WER mide el ratio entre el número de palabras incorrectamente reconocidas automáticamente y el número total de palabras en la transcripción correcta proporcionada por el usuario.

Destacar que comparamos nuestros resultados con el tiempo necesario para obtener una transcripción totalmente manual (sin disponer de transcripción automática inicial), que se estima estar alrededor de 10 RTF para usuarios no expertos (Munteanu, Penn y Zhu 2009). Más concretamente, este tiempo necesario para generar

transcripciones totalmente manuales se ha corroborado experimentalmente sobre poliMedia (Valor Miró y col. 2012).

La Tabla 1 muestra por columnas de izquierda a derecha, el protocolo de revisión utilizado, los resultados de WER inicial obtenidos por la transcripción automática, los resultados de WER final tras la revisión del usuario, el tiempo empleado por el usuario en el proceso de revisión en términos de RTF y la valoración subjetiva del sistema recopilada por una encuesta tras finalizar el proceso de revisión. Por filas podemos observar los resultados de los diferentes protocolos de revisión, desde la revisión que se realiza sobre la transcripción generada automáticamente, pasando por el protocolo de interacción inteligente basado en medidas de confianza, y finalmente la revisión en dos etapas.

Tabla 1: Comparativa de protocolos de revisión de transcripciones automáticas.

Protocolo de revisión	WER inicial	WER final	RTF	Encuesta
1 - Revisión manual	16.9	0.0	5.6	9.1
2 - Interacción inteligente	14.5	8.0	2.2	7.2
3 - Revisión en dos etapas	28.4	0.0	5.3	7.8

Los resultados obtenidos en las tres fases de la evaluación apuntan a que la interacción manual más sencilla utilizada en la primera fase es la preferida por nuestros usuarios, que consiguen reducir a la mitad el tiempo de obtención de una transcripción en comparación a realizarla totalmente manual. La revisión basada en medidas de confianza de la segunda fase requiere una menor dedicación por parte del usuario, pero el resultado es una transcripción con algunos errores que no son aceptable por nuestros usuarios dado el propósito docente de los vídeos. Finalmente, la tercera fase ofrece resultados de tiempo ligeramente mejores que la primera, pero debido a que el proceso en dos etapas es más complejo, los usuarios se decantaron por el protocolo de la primera fase.

6 Conclusiones

En este artículo hemos presentado una evaluación con profesores que se realizó en el marco de *Docència en Xarxa* 2012-2013, con el objetivo de obtener un sistema eficiente para realizar supervisiones de transcripciones de vídeos docentes de la plataforma poliMedia. También hemos presentado el proyecto transLectures que actualmente está ofreciendo transcripciones y traducciones automáticas para todos los vídeos docentes de la plataforma poliMedia de la UPV.

En las evaluaciones hemos comparado tres protocolos de revisión con el usuario, obteniendo en todos los casos tiempos de transcripción menores que si realizáramos la transcripción totalmente manual. Sin embargo, la técnica de interacción inteligente que limita el tiempo dedicado por el usuario a la revisión, a costa de proporcionar unas transcripciones imperfectas, no fue aceptada por los profesores. Por otro lado,

la ligera mejora de tiempo obtenida en la tercera fase no justifica para los profesores la complejidad extra que añade una revisión en dos etapas.

Así pues, como conclusión principal destacamos que la revisión de las transcripciones generadas automáticamente es el protocolo de revisión mejor aceptado por los profesores. Este modelo obtiene una reducción de aproximadamente el 50 % del tiempo necesario para generar una transcripción perfecta, al compararlo con el tiempo necesario para transcribir de forma totalmente manual.

Referencias bibliográficas

Munteanu, Cosmin, Gerald Penn y Xiaodan Zhu (2009). "Improving automatic speech recognition for lectures through transformation-based rules learned from minimal data". En: *Proc. of ACL-AFNLP*, págs. 764-772.

poliMedia (2007). *The poliMedia tool*. http://polimedia.blogs.upv.es/?lang=en.

Sanchez-Cortina, Isaias y col. (2012). "A prototype for interactive speech transcription balancing error and supervision effort". En: *Proc. of ACM IUI*, págs. 325-326.

Sanchis, Alberto, Alfons Juan y Enrique Vidal (2012). "A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition". En: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2, págs. 565-574.

Serrano, Nicolás y col. (2013). "Interactive handwriting recognition with limited user effort". En: *International Journal on Document Analysis and Recognition*, págs. 1-13.

Silvestre, Joan Albert y col. (2012). "transLectures". En: *Proceedings of IberSPEECH 2012*.

The TransLectures-UPV team (2013). *The TransLectures UPV toolkit (TLK)*. http://www.translectures.eu/tlk.

Valor Miró, Juan Daniel y col. (2012). "Integrating a State-of-the-Art ASR System into the Opencast Matterhorn Platform". En: *Advances in Speech and Language Technologies for Iberian Languages*. Vol. 328. CCIS. Springer, págs. 237-246. ISBN: 978-3-642-35291-1.