Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 167-170 recibido 31-03-2017 revisado 16-05-2017 aceptado 29-05-2017

VYTEDU: Un corpus de vídeos y sus transcripciones para investigación en el ámbito educativo

VYTEDU: A corpus of videos and transcriptions for research in the education domain

Jenny Alexandra Ortiz Zambrano
Universidad de Guayaquil
090514 Guayaquil, Ecuador
jenny.ortizz@ug.edu.ec

Arturo Montejo-Ráez
Universidad de Jaén
23071 Jaén, España
amontejo@ujaen.es

Resumen: El presente trabajo introduce un nuevo corpus de vídeos con sus transcripciones desarrollado en la Universidad Estatal de Guayaquil para el estudio de sistemas de simplificación de textos en el ámbito educativo. Para ello, se han producido hasta ahora 55 vídeos con sus transcripciones a texto, que se pone a disposición de la comunidad científica para su uso como herramienta de investigación. La orientación de demostración de este trabajo supone un intento para la difusión del material que posibilite el aprovechamiento temprano del mismo.
Palabras clave: Corpus multimodal, vídeo, transcripciones de vídeos, simplificación de textos, recurso

Abstract: This work introduces a new corpus of videos and their transcriptions developed in the Guayaquil National University for research in automatic text simplification in the education domain. To this end, 55 videos have been recorded, along with their literal transcriptions to text, offered freely to the scientific community for research purposes. This paper is oriented as a demonstration of the corpus, as a first attempt to disseminate its existence, enabling an early use of the corpus by other researchers.
Keywords: Multimodal corpus, video transcriptions, video, text simplification, resource

1 Introducción

La Universidad Estatal de Guayaquil (Ecuador) tiene interés en el desarrollo de tecnologías que faciliten la integración de los estudiantes en el proceso formativo académico. Para ello, anima al desarrollo de trabajos de investigación en este ámbito. Actualmente hay en marcha un trabajo de doctorado orientado a la simplificación de textos docentes obtenidos de las transcripciones de vídeos con contenido docente.

El corpus de Vídeos y Transcripciones en Educación (VYTEDU) realizado supone una fuente de datos fundamental para el desarrollo de la investigación, porque, si bien existen numerosas colecciones de vídeos y transcripciones para investigación, como el corpus AMI sobre vídeo-conferencias (Carletta, 2016), usado en anotación de roles semánticos (Sapru y Bourlard, 2015), escasean los recursos para español en general y educación en particular. En concreto, para español, destaca el corpus generado para análisis de sentimientos de (Rosas et al., 2013), conformado por 105 vídeos extraídos del popular servicio YouTube.

En este trabajo se introduce el proceso de generación del corpus tras una justificación del mismo como necesidad identificada en la propia universidad. Después se presenta una descripción más detallada del corpus para, finalmente, comentar la orientación práctica del mismo en tareas de simplificación de textos.

2 Justificación del corpus

Como se ha comentado en la introducción, trabajar con el español en el tema objeto de nuestra investigación, la subtitulación con textos simplificados de vídeos educativos, supone el reto de elaborar una colección de

ISSN 1135-5948

datos controlada para dicho fin. La simplificación automática de textos (SAT) es una tecnología usada para adaptar el contenido de un texto a las necesidades específicas de los individuos o de un colectivo determinado con el objeto de hacer dichos textos más legibles y comprensibles por ellos (Saggion et al., 2015). La simplificación automática de textos ha sido objeto de estudio desde hace más de veinte años (Chandrasekar et al., 1996) y puede servir para mejorar la accesibilidad a los contenidos (Saggion et al., 2011) y se ha estudiado con anterioridad para el español (Bott et al., 2012). En todo caso, no tenemos conocimiento del uso de simplificación de texto de transcripciones de vídeos para facilitar su compresión mediante la inclusión de subtítulos.

Adicionalmente, la necesidad de desarrollar un sistema de simplificación de textos para la Universidad Estatal de Guayaquil fue detectada tras un proceso de diagnóstico, en el que se elaboró una encuesta tomando en consideración la población estudiantil matriculada en el periodo 2015-2016 en dicha universidad.

Se consideró tomar la muestra de los estudiantes por categoría universitaria, esta categoría corresponde a una clasificación que presenta la Universidad de Guayaquil donde agrupa las 18 facultades y donde cada facultad posee uno o más programas académicos de pre grado[1]. Para las categorías: Ingeniería Industrial y Construcción, Salud y Bienestar, Ciencias Naturales Agricultura y Veterinaria, Ciencia Sociales Periodismo e Información, se tomó una muestra de 600 estudiantes; y una muestra de 100 estudiantes para las categorías: Administración de Empresas, y, Educación, Artes y Humanidades.

Los resultados reflejan que la comunidad de estudiantes valora enormemente la disponibilidad de vídeos docentes, así como herramientas que faciliten su seguimiento y comprensión.

3 Creación del corpus

Es proceso tomó un mes de trabajo, en el que se enviaron solicitudes a los diferentes decanatos para pedir autorización para la realización de un vídeo dentro del aula y así grabar la clase magistral del docente. En esta etapa colaboraron 10 estudiantes de primer semestre de la carrera de Ingeniería de Sistemas Computacionales de la Facultad de Ciencias Matemáticas y Físicas.

La grabación de los vídeos fue realizada en las diferentes carreras de las distintas facultades de la Universidad Estatal de Guayaquil.

4 Descripción del corpus

Los vídeos contienen en su grabación diferentes temáticas que corresponden a las diferentes asignaturas de programas académicos, tales como: Biblioteca Virtual (Sistemas de Información), Principios biomecánicos de las preparaciones dentarias (Odontología), Botánica (Ingeniería Agronómica), El problema de la deuda como problema de desarrollo (Economía), Economía de Mercado (Contaduría Pública Autorizada), Los sistemas de Información (Ingeniería en TeleInformática), Psicología Educativa (Psicología), La Hidráulica (Ingeniería Civil), Administración Estratégica (Ingeniería Comercial), Redes LAN (Networking), La Reiteración (Ingeniería Ambiental), Procesos para dirigir y gestionar la ejecución del proceso (Ingeniería Industrial), Procesos Constructivos Ingresados y Re-ingresados (Arquitectura), y algunas más.

Algunos ejemplos de los textos transcritos son los que se presentan a continuación:

"Para que un sistema muestre un comportamiento oscilante es necesario que tenga al menos dos niveles que son elementos del sistema en los que se producen acumulaciones. En ocasiones se observa un comportamiento oscilante como algo natural en todos los procesos ejemplos: al verano le sigue el invierno, al calor el frío, la noche el día y siempre vuelve al estado inicial, entonces tengo un sistema oscilante por ejemplo cuando sabemos que llegó el verano, a continuación llega el invierno y así sucesivamente, en conclusión si el estado actual del sistema no nos gusta o no es el correcto, no es necesario hacer nada ya que todo parece ser cíclico y volverá a la normalidad por sí solo". (Video-51) Ingeniería en TeleInformática, tema "Clasificación de los Sistemas".

"¿Qué pasa si las personas incumplen?

Lo primero es que en el caso de que tengan ya sus dos no-conformidades, tienen que pagar multas impuestas que vamos a ver que van desde veinte hasta doscientos sueldos mínimos, entonces tenemos: pago de multas impuestas, ejecución inmediata de correctivos de la no conformidad (eso lo tenemos como en

[1] http://www.ug.edu.ec/unidades-academicas/

el caso de la gente de Daule inmediatamente tuvimos que contratar empresa mediadoras para que san guiarán el área de hecho fue muy interesante porque utilizaron incluso de a un profesor de ingeniería química de la universidad de Guayaquil ellos generaron un polímero y ese polímero lo dispersaron en la zona y le pegaron palores y se hizo como un plástico y en ese plástico se pegaron todos los hidrocarburos fue bien interesante porque contrataron a varias compañías y una de las compañías lo que hizo fue llevar plumas de aves y ciertos compuestos para que se adhirieran los aceites pero la más interesante fue esa de ingeniería química porque realmente se vio cómo es un muy buen absorbente de las grasas", (Video-40) Ingeniería Ambiental, tema "La Reiteración".

Figura 1. En las aulas de clases de la carrera de Derecho Facultad Jurisprudencia

Figura 2. En las aulas de clases de la carrera de Arquitectura – Facultad Arquitectura

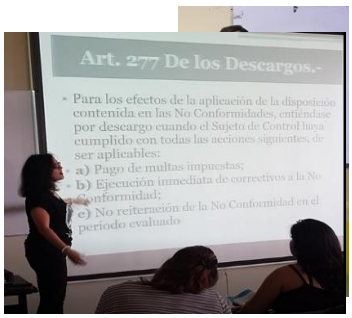

Figura 3. En las aulas de clases de la carrera de Ingeniería Ambiental – Facultad Ciencias Naturales

"Windows server está orientado a administrar grandes volúmenes de datos optimizar el uso de la agenda de red, seguro este sistema operativo como vimos la clase pasada que nos quedamos hasta la parte de las versiones de Windows server había un Windows server que estaba orientado a administrar Hardware de alto rendimiento por ejemplo pueden tener ustedes un servidor de 64 procesadores, una memoria de 400 gigas. ¿Para qué sirve el sistema operativo? Administrar el hardware, administrar el software de la máquina los periféricos de entrada y salida". (Video-46), carrera Networking, tema "Windows Server".

Las estadísticas del corpus, donde podemos observar la variabilidad de los vídeos quedan reflejadas en la Tabla 1.

	Mín	Máx	Media	Total
Vídeos				55
Duración	0:05:01	0:21:08	0:10:18	9:26:32
Tamaño Mb	4,9	2.645	804,5	44.248,1
Nº palabras	465	2.646	1.244	68.414
Nº párrafos	6	29	12.24	673

Tabla 1. Estadísticas del corpus

5 Midiendo la complejidad

El análisis de la complejidad del texto es una parte fundamental en nuestro trabajo. Esto permite tanto el estudio de los textos objeto de nuestro trabajo, como la evaluación de un futuro sistema. Para ello, el sistema mide algunas de los indicadores seleccionados por (Saggion et al., 2015). Constituyen un conjunto de métricas que permiten analizar la complejidad del texto a varios niveles: el *índice de complejidad léxica* (Fórmula 1) y el *índice de complejidad de oración* (Fórmula 5), propuestos por (Anula, 2008), y la *legibilidad del español del Spaulding* (Spaulding, 1956), detallada en la Fórmula 4.

Estos valores se calculan según las siguientes fórmulas:

$$LC = (LDI + ILFW) / 2 \quad (1)$$

$$LDI = N(dcw) / N(s) \quad (2)$$

$$ILFW = N(lfw) / N(cw) * 100 \quad (3)$$

$$SSR = 1.609\ N(w) / N(s) + 331.8\ N(rw) / N(w) + 22.0 \quad (4)$$

$$SCI = (ASL + CS) / 2 \quad (5)$$

$$ASL = N(w)/N(s) \quad (6)$$

$$CS = N(cs)/N(s) \quad (7)$$

Donde:

- LDI: *lexical distribution index* (índice de complejidad léxica)
- ILFW: index of low frequency words (índice de palabras poco frecuentes)
- N(dcw): número de palabras de contenido diferentes (sustantivos, adjetivos y verbos), generalmente lematizados
- N(cw): número de palabras de contenido totales (sustantivos, adjetivos y verbos), generalmente lematizados
- N(s): número de oraciones
- N(lfw): número de palabras de baja frecuencia (aparecen una o dos veces)
- N(w): número de palabras en el texto
- N(cs): número de oraciones complejas. Son aquellas que tienen más de un "clúster" de verbos, siendo un clúster de verbos aquellos verbos adyacentes sin la intervención de otras categorías de palabras, por ejemplo: *ha comido* o *quiere comer*.

6 Conclusiones y trabajo futuro

Nuestro objetivo es llegar al centenar de vídeos, si no más, en breve. En cualquier caso, este material ya está disponible y puede ser obtenido contactando con los autores.

También estamos trabajando en mejorar la compresión de los vídeos con un formato más compacto que reduzca el tamaño de los archivos, pues estos no han sido procesados después de su grabación directa.

Asimismo, algunos aspectos relativos al corpus como identificar el vocabulario utilizado y caracterizar las transcripciones con herramientas de lingüística computacional es también una tarea a realizar. A partir de ese momento el objetivo es iniciar la investigación de los aspectos siguientes:

- Calidad de los sistemas de transcripción automáticos.
- Análisis de las herramientas de simplificación de textos actuales.
- Estudio y diseño de nuevos algoritmos para la generación de subtítulos simplificados.

Consideramos que este corpus puede ser una aportación valiosa a la comunidad científica para seguir avanzando en el estudio de técnicas de PLN.

7 Bibliografía

Anula, A. 2008. Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad. En *XVIII Congreso Internacional de la Asociación para la Enseñanza del Español como lengua Extranjera (ASELE,)* Alicante, páginas 162-170.

Bott, S., H. Saggion y S. Mille. 2012. Text Simplification Tools for Spanish. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, páginas 1665-1671. Estambul (Turquía).

Carletta, J. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181-190.

Chandrasekar, R., C. Doran y B. Srinivas. 1996. Motivations and methods for text simplification. En *Proceedings of the 16th Conference on Computational Linguistic*s. Volumen 2, páginas 1041-1044. Association for Computational Linguistics. California (EEUU).

Rosas, V. P., R. Mihalcea y L. P. Morency. 2013. Multimodal sentiment analysis of Spanish online videos. *IEEE Intelligent Systems*, 28(3):38-45.

Saggion, H., E. G. Martínez, E. Etayo, A. Anula y L. Bourg, L. 2011. Text simplification in Simplext. Making text more accessible. *Procesamiento del Lenguaje Natural (SEPLN)*, 47:341-342.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, L. y B. Drndarevic, B. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14

Sapru, A., y H. Bourlard. 2015. Automatic recognition of emergent social roles in small group interactions. *IEEE Transactions on Multimedia*, 17(5):746-760.

Spaulding, S. 1956. A Spanish readability formula. *The Modern Language Journal*, 40(8):433-441.