Presencia de IRRODL

Evaluación automática del aprendizaje electrónico utilizando el análisis semántico latente: un caso de uso

Autores: Mireia Farrús y Marta R. Costa-jussà

Automatic Evaluation for E-Learning Using Latent Semantic Analysis: A Use Case

Resumen

El presente trabajo tiene como objetivo analizar y discutir algunos de los más recientes sistemas de evaluación de la educación. Para ello, se expone un caso de uso específico desarrollado para la Universidad Abierta de Cataluña, que es una universidad en línea. Se propone una herramienta de evaluación automática que permite al estudiante autoevaluarse en cualquier momento y recibir retroalimentación inmediata. Esta herramienta es una plataforma basada en la red, y se ha diseñado para asignaturas de Ingeniería (es decir, con símbolos matemáticos y fórmulas) en catalán y español. En particular, la técnica utilizada para la evaluación automática es el análisis semántico latente. Aunque el marco experimental del caso de uso es bastante desafiante, los resultados son prometedores.

Palabras clave: aprendizaje electrónico, evaluación de prueba automática, plataforma de red, análisis semántico latente

Abstract

Assessment in education allows for obtaining, organizing, and presenting information about how much and how well the student is learning. The current paper aims at analysing and discussing some of the most state-of-the-art assessment systems in education. Later, this work presents a specific use case developed for the *Universitat Oberta de Catalunya*, which is an online university. An automatic evaluation tool is proposed that allows the student to evaluate himself anytime and receive instant feedback. This tool is a web-based platform, and it has been designed for engineering subjects (i.e., with math symbols and formulas) in Catalan and Spanish. Particularly, the technique used for automatic assessment is latent semantic analysis. Although the experimental framework from the use case is quite challenging, results are promising.

Keywords: E-Learning, Automatic Test Assessment, Web Platform, Latent Semantic Analysis

Introducción

La evaluación de la educación permite obtener, organizar y presentar información acerca de cuánto y qué tan bien está aprendiendo el estudiante. La evaluación utiliza varias técnicas durante el proceso de enseñanza-aprendizaje y es especialmente útil en la evaluación de las preguntas abiertas, ya que permite a los profesores comprender mejor cómo se ha relacionado el alumno con la asignatura. En algunos casos, por ejemplo, los estudiantes con alta puntuación en pruebas de pregunta cerrada reportan errores conceptuales subyacentes al ser entrevistados por un profesor (Tyner, 1999).

Durante los últimos años se ha incrementado sustancialmente el uso de una computadora para propósitos de evaluación. Entre los objetivos que se persiguen al utilizar la evaluación computarizada están el logro y la consolidación de las ventajas de un sistema con las siguientes características (Brown *et al*., 1999.).

En primer lugar, para reducir la carga de trabajo de los profesores mediante la automatización parcial de la evaluación de los estudiantes; en segundo lugar, para proporcionar a los estudiantes información detallada sobre su periodo de aprendizaje de una manera más eficiente que con la evaluación tradicional; y, por último, para integrar la cultura de la evaluación en el trabajo cotidiano de los estudiantes en un entorno de aprendizaje electrónico. De hecho, hoy día una de las cosas más importantes en la evaluación es la retroalimentación, por lo que la evaluación del aprendizaje se destina generalmente para medir los resultados del aprendizaje y reportarlos a los estudiantes (y no sólo al sistema o al profesor).

Enseguida describiremos brevemente el uso del análisis semántico latente como un algoritmo analizador semántico de documentos relacionados entre sí y los explicaremos en el contexto de las tareas de evaluación. Más adelante, expondremos el caso de uso anteriormente señalado, que aprovecha el análisis semántico latente con el fin de obtener los resultados de la evaluación. Por último, se presentan las conclusiones.

Plataformas de evaluación del aprendizaje electrónico

Algunos estudios están orientados a la investigación de la calificación automatizada de ensayos. Los más relevantes se pueden encontrar en Miller (2003), Shermis y Burstein (2003), Hidekatsu *et al*. (2007) y Hussein (2008). Sin embargo, los estudios que cubren la calificación automatizada de ensayos en asignaturas de Ingeniería son limitados, hasta donde sabemos. En Quah *et al*. (2009), por ejemplo, los autores utilizan una máquina de apoyo vectorial para construir un prototipo capaz de evaluar las ecuaciones y respuestas cortas. El sistema extrae los datos textuales y matemáticos de los archivos de entrada en forma de palabras perceptibles para texto y ecuaciones matemáticas utilizando árboles de ecuación basados en el formato MathTree. Enseguida, el sistema aprende cómo evaluarlos, basado en las calificaciones dadas al principio, aprendiendo el esquema de evaluación y evaluando automáticamente las secuencias de comandos posteriores.

Muchos portales se encuentran en línea. Algunos que pueden darnos un panorama general son, por ejemplo, Aprendizaje en Línea y Servicios de Colaboración (OLCS, http://www.olcs.lt.vt.edu); VirginiaTech, que proporciona administración de sistemas, soporte y capacitación para los académicos, evaluaciones de los cursos en línea y otro software educativo. Por su parte, la Herramienta Educativa Colaborativa Ville (http://ville.cs.utu.fi/) es un entorno completo capaz de hacer muchos tipos

de evaluación, por lo que es ideal para las personas que desean elaborar su propio material en lugar de crear un nuevo sitio web. Además, es más fácil obtener retroalimentación sobre el material si se hace en colaboración con otros docentes.

Otro ejemplo de una plataforma de aprendizaje es la Academia Khan (http://www.khanacademy.org), que ha creado un marco genérico para ejercicios de construcción. Este marco, junto con los propios ejercicios, se puede utilizar de forma totalmente independiente de la aplicación de la Academia. El marco existe en dos componentes: un formato HTML para los ejercicios específicos y un plug-in para generar un ejercicio útil e interactivo desde el formato HTML.

Además, se pueden encontrar algunos sistemas específicamente para ejercicios de matemáticas. Por ejemplo, STACK (http://www.stack.bham.ac.uk) es un sistema de código abierto para la evaluación asistida por computadora en matemáticas y disciplinas afines, con énfasis en la evaluación formativa. Y algunos sistemas, tales como texto reestructurado (http://docutils.sourceforge.net/rst.html), proporcionan las técnicas que se pueden utilizar para desarrollar nuevos materiales.

Análisis semántico latente en el aprendizaje electrónico

La tarea de evaluar un documento en el contexto de nuestra educación implica analizar el contenido semántico de dicho documento. Para este fin, el Análisis Semántico Latente (por sus siglas en inglés, LSA), también conocido como Indexación Semántica Latente, consiste en una técnica que analiza una relación semántica entre un conjunto de documentos y los términos que contienen (Hofmann, 1999), se ha aplicado exitosamente en múltiples áreas de procesamiento del lenguaje natural tales como la recuperación de información interlingüística (Dumais *et al*., 1996), la combinación de enunciados interligüísticos (Banchs y Costa-Jussà, 2010) y la traducción automática estadística (Banchs y Costa-Jussà, 2011).

El objetivo del LSA es analizar documentos con el fin de encontrar su significado o conceptos subyacentes. La técnica surge a partir del problema de cómo comparar palabras para encontrar documentos relevantes, ya que lo que realmente queremos hacer es comparar conceptos y significados que están detrás de las palabras, en lugar de las palabras en sí mismas. En el LSA, tanto las palabras como los documentos se esquematizan en un espacio conceptual. Es en este espacio donde se realiza la comparación. Este espacio se crea por medio de la conocida técnica de Descomposición en Valores Singulares (por sus siglas en inglés, SVD), que es una factorización de una matriz real o compleja (Greenacre, 2011).

En el ámbito específico de la evaluación de ensayos, el LSA ha mostrado resultados prometedores en el análisis de contenido de ensayos (Landauer *et al*., 1997), donde las medidas basadas en esta técnica estaban estrechamente relacionadas con los juicios humanos para predecir hasta qué punto el alumno aprenderá a partir del texto (Wolfe *et al*., 1998; Rehder, *et al*., 1998) y en calificar las respuestas del ensayo (Kakkakonen *et al*., 2005). Otras aplicaciones educativas son sistemas tutoriales inteligentes que proporcionan ayuda para estudiantes (Wiemer-Hastings *et al*., 1999; Foltz *et al*., 1999b) y la evaluación de resúmenes (Steinhart, 2000). En este contexto, como el LSA es independiente del idioma, se ha aplicado en ensayos escritos en inglés (Wiemer-Hastings y Graesser, 2000), en francés (Lemaire y Dessus, 2001) y en finlandés (Kakkakonen *et al*., 2005). Todos estos estudios muestran que, a pesar de que toma en cuenta el orden de las palabras, el LSA es capaz de captar una parte

significativa del significado no sólo de las palabras individuales, sino también de pasajes enteros, tales como oraciones, párrafos y ensayos cortos. Es por eso que hemos elegido el LSA para comparar la similitud semántica de los documentos en el espacio conceptual (Pérez *et al*., 2006).

En particular, en este trabajo y a diferencia de la literatura previa, se investiga si el LSA se puede aplicar para la evaluación electrónica de ensayos matemáticos. Además, los experimentos se realizan tanto en catalán como en español. El LSA está integrado de la siguiente manera.

Los documentos que contienen las respuestas de los estudiantes se comparan con uno o más documentos de referencia que contienen las respuestas correctas creadas por los profesores. Entonces esta comparación semántica entre los documentos de los estudiantes y los documentos de referencia permitirá a los profesores generar una evaluación aproximada de losalumnos. Para la comparación o recuperación de documentos, éstos se transforman normalmente en una representación adecuada, por lo general un modelo espacio vectorial (Salton, 1989). Un documento se representa como un vector en el que cada dimensión corresponde a un término independiente. Si se produce un término en el documento, su valor en el vector no es cero. Se han desarrollado varias formas de calcular estos valores conocidas como ponderaciones (temporales). Uno de los esquemas más conocidos es el de ponderación tf-idf, por sus siglas en inglés (frecuencia de término-frecuencia inversa de documento).

La ponderación tf-idf define estadísticamente la importancia que tiene una palabra para un documento en una colección. Tal representación es conocida por ser confusa y dispersa. Es por ello que con el fin de obtener representaciones espacio vectoriales más eficientes, se aplican técnicas de reducción de espacio (Deerwester *et al*., 1990; Hofmann, 1999; Sebastiani, 2002) de modo que el nuevo espacio reducido supuestamente capta las relaciones semánticas entre los documentos de la colección. La *figura 1* muestra una representación esquemática del uso del LSA para la calificación automática de ensayos.

Finalmente se calcula una medida de similitud de distancia del coseno entre cada examen y su solución en el espacio reducido, obteniendo una calificación que muestra cómo un conjunto particular de exámenes es similar en la semántica con su solución correspondiente.

El caso de uso de la Universidad Abierta de Cataluña

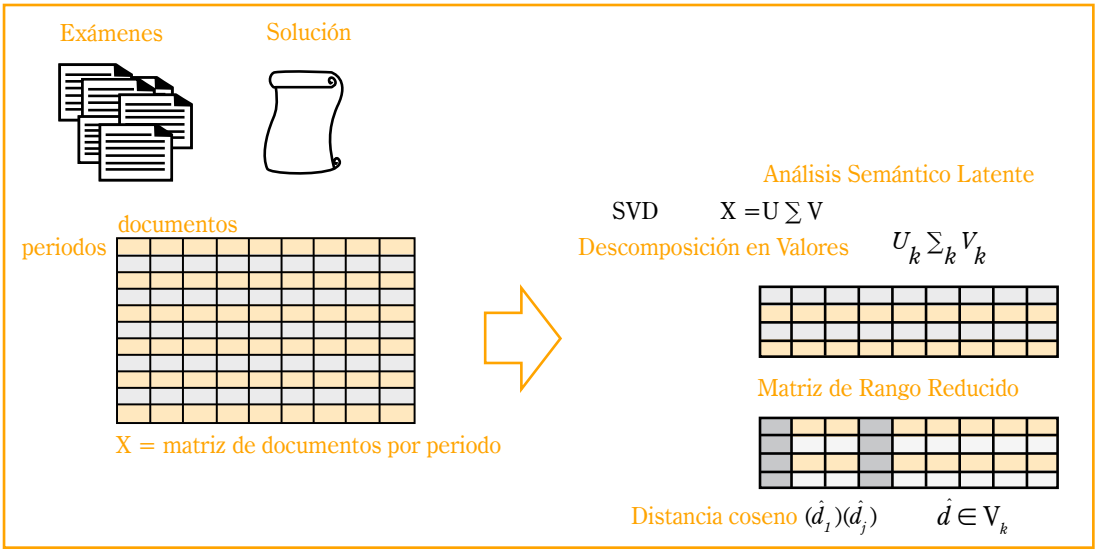
En esta sección abordaremos la creación de una herramienta de evaluación de texto libre a través de Internet, lo que permite la evaluación automática de los alumnos de la Universidad Abierta de Cataluña (Universitat Oberta de Catalunya, UOC). Las principales características del sistema de evaluación de la universidad y la herramienta desarrollada se describen enseguida.

La Universidad Abierta de Cataluña

La UOC es una universidad en línea con sede en Barcelona, con más de 54 mil estudiantes y más de 2000 mil tutores, con apoyo administrativo conformado por 500 personas aproximadamente, que prestan servicios a todos estos estudiantes.

Los alumnos siguen un sistema de evaluación continua, que se lleva a cabo en línea a lo largo del semestre. Aunque este sistema se utiliza con éxito para completar sus estudios, uno de los principales problemas es el creciente número de matrícula estudiantil cada año, lo que hace que la tarea de calificar sus continuas evaluaciones sea tediosa y tardada. Asimismo, se necesitan más tutores externos para realizar esta tarea, lo que hace que sea difícil acordar criterios.

Figura 1. La representación esquemática del uso del LSA para la calificación automática de ensayos es la matriz del documento por periodo y es la descomposición en valores singulares de la matriz, lo que permite calcular una matriz de reducción de rango sobre la que se calcula la distancia del coseno entre los documentos.



La herramienta de evaluación

La herramienta desarrollada en la UOC tiene como objetivo proporcionar una evaluación automática de las tareas para las asignaturas de Ingeniería mediante el uso de la técnica de Análisis Semántico Latente, tomando en cuenta el trabajo realizado por Miller (2003), en el que se examina la aplicación del LSA para calificación automática de ensayos y se compara con métodos estadísticos previos para evaluar la calidad del ensayo. La implementación del LSA se realiza utilizando JAVA.

La herramienta de evaluación de texto libre basada en la red, permite a los profesores diseñar tantas pruebas de evaluación como deseen, con tantas preguntas como consideren necesarias para la misma. Por un lado, para cada pregunta, el profesor asocia varios modelos de respuesta correcta con el fin de generar suficientes respuestas de referencia para garantizar que el sistema de evaluación automática funcione correctamente. Por otro lado, la plataforma basada en la red permite a los estudiantes hacer tantas pruebas de evaluación como deseen, generando, después de cada prueba realizada, un informe que contenga los resultados de la evaluación de cada pregunta individual, así como los resultados generales. Además, la herramienta proporciona a los estudiantes la posibilidad de comparar las respuestas de referencia generadas por el profesor con sus propias respuestas con el fin de dar retroalimentación detallada y mejorar su proceso de aprendizaje. La plataforma también incluye un editor de texto que propicia la inserción de fórmulas, tanto en las declaraciones como en las respuestas con el JavaScript plug-in MathML (Su *et al.*, 2006).

Experimentos de evaluación

Es aquí donde precisamente describiremos el marco experimental en nuestro estudio de caso. Incluimos subsecciones que describen particularmente el marco de trabajo, la interfaz de la red, los experimentos de evaluación y los resultados que obtuvimos.

Marco de trabajo

El objetivo principal de la herramienta es ayudar a los profesores en sus tareas de evaluación de un gran número de estudiantes. Estos primeros experimentos implican un grupo controlado y relativamente pequeño de alumnos con el fin de establecer las bases para experimentos posteriores y más amplios. El marco de aplicación abarca a los estudiantes en dos semestres consecutivos (con 54 y 70 alumnos matriculados, respectivamente) de una sola asignatura llamada "Teoría de Circuitos", una asignatura básica del primer año del Grado de Ingeniería de Telecomunicaciones de la UOC .

Además de la evaluación individual que se realiza al final del semestre, el modelo de evaluación de la asignatura consta de cuatro tareas de evaluación continua (por sus siglas en inglés, CAA) distribuidas a lo largo del semestre, así como un trabajo práctico que incluye ejercicios de simulación por computadora, estructurados de la siguiente manera. Las tres primeras CAA se componen de una sección de preguntas cortas y una sección de ejercicios. La cuarta y última CAA sólo contiene una sección de ejercicios. Específicamente, las secciones de preguntas cortas consisten en una serie de cinco a seis preguntas sobre temas muy concretos. Para cada una de estas preguntas se proporcionan cuatro posibles respuestas, y sólo una de ellas es correcta; de este modo, los estudiantes tienen que especificar la respuesta correcta y las razones de sus elecciones. Debido a la naturaleza técnica de la asignatura, las ecuaciones matemáticas suelen aparecer en el texto de ambas preguntas y respuestas, así como en las justificaciones correspondientes de los estudiantes.

En este contexto, la sección de preguntas cortas de las tres primeras CAA se han elegido como un marco de aplicación específica para realizar los experimentos de evaluación automática, debido a la pertinencia de la estructura y la longitud de las preguntas y las respuestas, así como a la naturaleza (breve texto más algunas ecuaciones matemáticas) de las justificaciones que los estudiantes tienen que ofrecer.

Interfaz de red

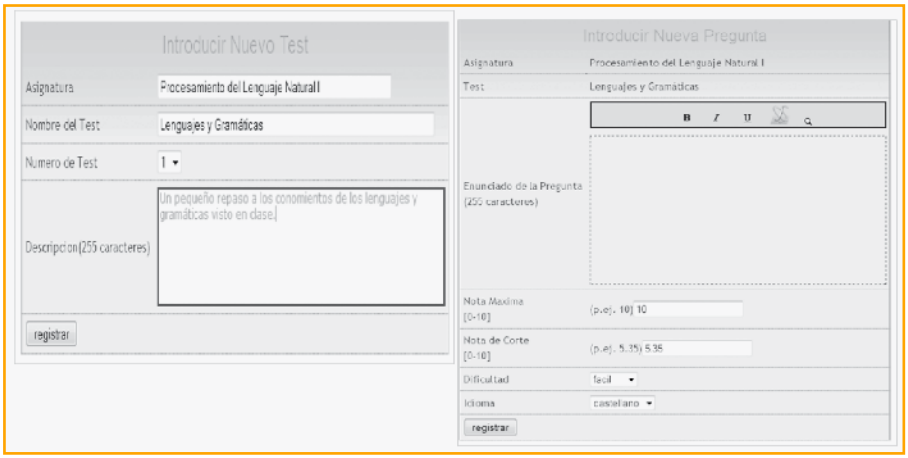
El sistema automático de evaluación de pruebas se presenta como una plataforma de red, donde el acceso se puede llevar a cabo desde dos perfiles diferentes: el profesor y el estudiante. La principal tarea del profesor es proporcionar preguntas y respuestas correctas de referencia. Así, un profesor puede realizar dos acciones distintas para cada asignatura: crear una nueva prueba y modificar una existente.

Para crear una nueva prueba, el profesor debe definir primero el nombre de la prueba, la asignatura a la que pertenece, la posición dentro del conjunto de pruebas de la asignatura, así como una breve descripción (*figura 2a*). Una vez introducidas estas características, el profesor puede registrar la prueba vacía en la base de datos y después insertar tantas preguntas como lo desee en la prueba. Por cada pregunta nueva, deben cubrirse las siguientes características:

a. una declaración
b. puntuación máxima posible
c. puntuación mínima para aprobar la pregunta
d. dificultad de la pregunta
e. el idioma de la declaración (*figura 2b*)

Además, un conjunto de respuestas de referencia está asociado con cada pregunta. Asimismo,

Figura 2. Página de creación para una nueva prueba (a) y formato de creación para una nueva pregunta (b).



el profesor puede consultar tanto los resultados obtenidos, como las respuestas proporcionadas por los estudiantes.

Una vez autorizado o legalizado, los estudiantes pueden ejecutar las siguientes acciones:

1. Autoevaluarse mediante la realización de una prueba.
2. Revisar el historial de las pruebas realizadas.
3. Consultar las calificaciones obtenidas, así como el máximo y el mínimo
4. definido por el profesor.

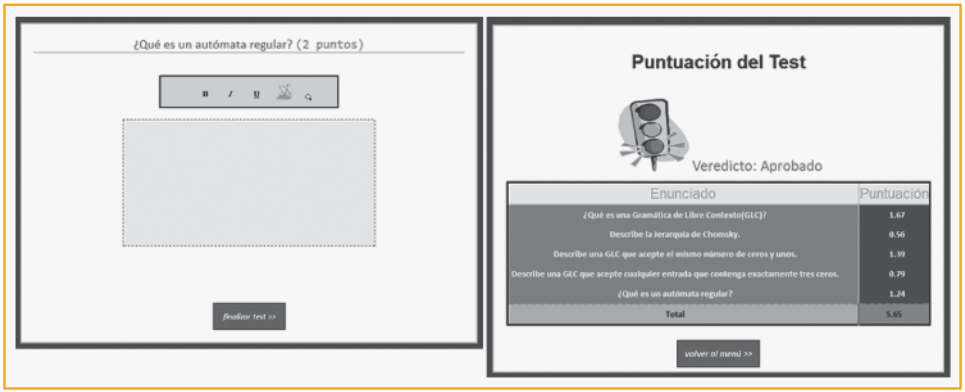
Para autoevaluarse, se proporciona a los estudiantes una lista de temas ordenados alfabéticamente en la que puedan llevar a cabo la evaluación eligiendo un tema y seleccionando la prueba que desean comenzar, así como su nivel de dificultad. Además, se presenta a los alumnos la declaración de cada pregunta junto con su correspondiente puntuación. Los estudiantes responderán en un editor de texto en el que pueden insertar fórmulas gracias al plug-in JavaScript llamado MathEdit (Su *et al*., 2006), como se muestra en la *figura 3a*. Una vez que la respuesta ha sido tecleada y la prueba está terminada, el sistema proporciona una puntuación al estudiante, junto con las calificaciones obtenidas en cada una de las preguntas (*figura 3b*). Del mismo modo, los alumnos pueden revisar, para cada pregunta, las respuestas que escribieron, así como las preguntas de referencia escritas por el profesor.

Además de la realización de las pruebas, los estudiantes tienen la posibilidad de ingresar a la plataforma con el fin de evaluar su progreso. Por tanto, cada estudiante tiene acceso a un historial en el que puede ver una lista de las pruebas realizadas. Una vez que se elige una de estas pruebas, las preguntas se pueden ver con detalle, incluyendo la respuesta dada por el estudiante, la calificación obtenida, la puntuación máxima y la mínima definidas por el profesor, y las respuestas de referencia utilizadas por el sistema de evaluación automático con el fin de hacer las correcciones.

Experimentos de evaluación

En esta sección se describe la evaluación automática elaborada a través de las tareas de evaluación continua (CAA) de los estudiantes.

Figura 3. Editor de preguntas y texto con MathEdit (a) y puntuación de la prueba una vez terminada (b).



Los experimentos llevados a cabo utilizaron las CAA de dos semestres consecutivos, S1 y S2, en la que 54 y 70 estudiantes fueron registrados, respectivamente. Cada semestre incluyó un conjunto de tres diferentes CAA (CAA1, CAA2 y CAA3). Los datos se convirtieron en componentes léxicos, en minúsculas. Las 20 palabras más frecuentes fueron descartadas. Enseguida se describe el procedimiento para el tratamiento de un conjunto de soluciones con el LSA:

1. Calcular soluciones de N en términos de tf-idf:
   a. Extraer vocabulario.
   b. Cada solución es un vector de dimensiones M.
2. Solución de la matriz N*M.
3. Calcular SVD.
4. Seleccionar valores singulares L.

Luego, para cada respuesta del estudiante el procedimiento es el siguiente:

1. Vectorizar la respuesta en términos de tf-idf, usar el vocabulario de la serie de soluciones. Tenemos un vector de dimensión M.
2. Proyectar el vector en el espacio reducido.
3. Calcular la similitud de este vector de espacio reducido con cada solución. Mantenemos la distancia máxima.

El material utilizado en el análisis presentó tres problemas principales.

1. Archivos de formato. Las CAA de los estudiantes se entregan en varios formatos distintos, aunque se encuentran principalmente en PDF, Word y Open Office Writer. Algunas incluso son documentos escaneados y pegados como archivos de imagen en documentos de Word o Writer. Por tanto, no todas las CAA se pueden transformar fácilmente en formato TXT para que se puedan manejar adecuadamente. En consecuencia, los documentos PDF y todos aquellos documentos que contienen archivos de imagen se han retirado del conjunto original de archivos. La *tabla 1* muestra, para cada semestre, el número de alumnos matriculados, el número de documentos originales, y el número de documentos utilizados después de retirar los documentos con imágenes pegadas y documentos PDF. La tabla también muestra el vo-

Tabla 1. Estudiantes registrados, número de CAA originales (#orig.), número de CAA utilizados (# utilizado), y tamaño de vocabulario empleado (vocab.) para cada semestre.

Semestre	Estudiantes	CAA1			CAA2			CAA3		
		# orig.	# utilizado	vocab.	# orig.	# utilizado	vocab.	# orig.	# utilizado	vocab.
S1	54	20	14	857	19	13	730	15	10	712
S2	70	28	20	1027	25	9	699	20	16	1291

cabulario para cada CAA. Como puede verse, el tamaño del vocabulario no se correlaciona con el número de CAA, por lo que el contenido de vocabulario de las CAA varía en gran medida entre cada conjunto.

2. Formulación matemática. Teniendo en cuenta que estamos utilizando un enfoque de bolsa-de-palabras, la formulación que se obtiene de los documentos de Open Office se codificó en MathML (Lenguaje de Marcado Matemático), pero no así con la formulación extraída de documentos de Word, lo que hizo una gran diferencia entre las CAA respecto al vocabulario final.
3. Idioma. Los estudiantes presentaron las CAA, tanto en catalán como en español. En este caso, asumimos que el método presentado en el documento actual es capaz de aprovechar el vocabulario que es independiente del idioma, tal como las variables matemáticas.

Resultados

Para llevar a cabo los experimentos preliminares de la evaluación, se utilizaron las CAA1 y CAA2 del semestre S1 como material de desarrollo, lo que permitió llegar a la conclusión de que la mejor reducción de rango en el análisis semántico latente era de cinco.

Los resultados se muestran en términos de la correlación obtenida entre las evaluaciones automáticas y humanas. Definimos evaluación humana como la evaluación realizada por el profesor de manera tradicional, mientras que la evaluación automática se define como una evaluación computarizada determinada por la metodología propuesta en el presente trabajo (es decir, las cuantificaciones obtenidas de forma automática mediante el análisis semántico latente y la distancia de coseno).

Por tanto, utilizando el análisis semántico latente, se obtuvieron evaluaciones automáticas para cada estudiante, CAA, y semestre. Después, se calcularon las correlaciones entre las evaluaciones automáticas y las humanas para cada semestre y conjunto de CAA. Los resultados obtenidos de la correlación se presentan en la *tabla 2* (columna de correlación), junto con la significación estadística de los resultados de la correlación (columna p).

Como puede verse en la tabla, en los resultados estadísticamente significativos (es decir, donde $p < 0.05$), la correlación varía de 52 % a 69 % (véase CAA1 y CAA2 del semestre S2). Aunque estos resultados son inferiores a los presentados en Miller (2003), son prometedores ya que se trata de una asignatura completamente textual, pero que contiene un número considerable de fórmulas matemáticas. El resto de los resultados (S1 y CAA3 de S2) no son estadísticamente significativos.

Tabla 2. Resultados de correlación (corr.) y significancia estadística (p) entre la evaluación automática y la humana, y el porcentaje de CAA cumpliendo con las mismas características como soluciones de referencia (mismas caract.).

Semestre	CAA1			CAA2			CAA3		
	corr.	p	Mismas caract.	corr.	p	Mismas caract.	corr.	p	Mismas caract.
S1	16%	0.60	14%	12%	0.68	15%	15%	0.68	10%
S2	52%	0.04	30%	69%	0.04	28%	29%	0.27	25%

Por un lado, hay que considerar que las respuestas de referencia fueron escritas en catalán por los profesores, mientras que los estudiantes podían elegir si querían responder las pruebas en catalán o en español, por lo que el idioma de las pruebas no era el mismo en todas las CAA de los estudiantes. Por otro lado, a diferencia de las CAA de los estudiantes, todas las soluciones de referencia estaban disponibles en formato Writer. Dado que sólo las fórmulas matemáticas de los documentos de Writer se transformaron en MathML, también hubo disparidad en las fórmulas en cada colección de CAA. Con el fin de evaluar cómo estas diferencias podrían haber afectado los resultados, se calculó el porcentaje de CAA en cada conjunto que cumplieron al mismo tiempo con los dos requisitos siguientes(es decir, cumplir las dos características por las soluciones de referencia):

1. Las fórmulas fueron codificadas en MathML.
2. Los estudiantes respondieron en el idioma catalán.

El porcentaje de CAA que cumple con ambas características se muestra en la tabla 2, en la tercera columna de cada resultado de la CAA. Se puede observar que los dos resultados estadísticamente significativos con una correlación de más de 50 % (es decir, las CAA1 y CAA2 del semestre S2) corresponden a los resultados en los que la codificación y el lenguaje utilizado es el mismo que las soluciones de referencia en más de 25 % de los casos. Por tanto, a partir de los resultados se puede afirmar que la correlación entre las evaluaciones humanas y las automáticas depende tanto de la coherencia de la codificación matemática como del lenguaje usado en las pruebas.

Por ejemplo, desde CAA1 y S1, una respuesta a una pregunta corta ser evaluada fue: “Si introduïm un senyal sinusoidal en un circuit, la resposta forçada serà una sinusoide que l’entrada amplificada per H(s)” (en español: “Si se introduce una señal sinusoidal en un circuito, la respuesta forzada es una sinusoide amplificada por la entrada de H(s)”). La respuesta fue: “La resposta del sistema és una senyal sinusoidal de la mateixa freqüència amplificada per H(s)” (en español: “La respuesta del sistema es una señal sinusoidal de la misma frecuencia amplificada por H(s)”). Sólo hay un

detalle de la mateixa freqüència (en español, "la misma frecuencia"), que no está presente en la respuesta del estudiante. La respuesta es calificada por el profesor con un ocho y por el sistema con un nueve.

Para concluir los resultados presentados, puede ser interesante analizar brevemente la función de MathML, a diferencia de las palabras en los informes escritos. En el momento de la realización de estos experimentos, las fórmulas matemáticas fueron simplemente tratadas como palabras. De hecho, uno de los inconvenientes de este estudio es que se trata del método de la bolsa de palabras, por tanto, el orden de las palabras, que sin duda es importante en el significado de las fórmulas matemáticas, no se toma en cuenta. Por ejemplo, el método no distingue entre I=V/R e I=R/V. Sin embargo, ya que la primera es totalmente correcta, la última es completamente errónea. Este es uno de los retos que hay que resolver en futuras investigaciones.

Conclusiones

En este trabajo se ha presentado un análisis y una discusión sobre los sistemas más modernos de evaluación en la educación. Además, se muestra un estudio detallado de una herramienta de corrección automática incorporada como parte de las aulas virtuales en el entorno de enseñanza-aprendizaje basado en la red de la UOC, con el fin de ayudar con la autoevaluación de los estudiantes, al proporcionarles retroalimentación instantánea. De esta manera, los alumnos adultos a distancia, que suelen carecer de tiempo, no tienen que esperar que las evaluaciones de los profesores sean calificadas. Esta herramienta, basada en una interfaz de red, está diseñada para ser utilizada en un entorno en línea, tanto por el profesor (el diseño correcto y pruebas de evaluación) como por los estudiantes (la autoevaluación). El proceso de evaluación automática se basa en técnicas de prueba usando el procesamiento del lenguaje natural y el procesamiento semántico latente.

El estudio de caso realizado en este documento ha tenido que superar algunos problemas relacionados con el material disponible, el primero de los cuales es la existencia de una gran cantidad de fórmulas matemáticas en las asignaturas de Ingeniería tratadas. Aunque muchos trabajos de investigación han abordado el tema de la calificación de ensayos automatizada, por lo que a nosotros respecta, no han tratado con el lenguaje matemático. Por otra parte, las pruebas de los estudiantes están disponibles en diferentes idiomas y formatos de archivo, lo que hace aún más difícil tratar las fórmulas matemáticas al convertirlas en un código homogéneo.

Con el fin de ser capaces de tratar el material disponible, desde el inicio se eliminaron los documentos PDF y los documentos de Word o Writer que contenían imágenes pegadas como respuestas. Sin embargo, estamos conscientes de que éste no es el mejor método para recopilar datos y ambos (PDF y archivos de imagen) serán estudiados en futuras investigaciones.

A pesar de las dificultades con el material utilizado, los experimentos preliminares han mostrado algunos resultados interesantes. Después de calcular la correlación entre las pruebas de evaluación automáticas y humanas, se ha demostrado que sólo dos de las seis pruebas de evaluación proporcionaron una correlación mayor a 50 % con resultados estadísticamente significativos. Estos dos conjuntos corresponden a los conjuntos de PAC que tienen más similitud con las soluciones de referencia PAC: las fórmulas matemáticas se codifican en MathML y las respuestas de los estudiantes eran, en su mayoría, escritas en el mismo idioma.

En evaluaciones automáticas de ensayo esperaríamos una mayor correlación. Sin embar-

go, se trata de una cuestión difícil, ya que no incluyen símbolos y fórmulas matemáticas, lo que hace que el análisis actual sea más difícil. Por tanto, aunque por el momento los resultados de la correlación no son satisfactorios, han establecido un punto de partida que nos permite trabajar con este tipo de material en asignaturas de Ingeniería. Así, el trabajo futuro se centrará en mejorar el formato de los materiales para darles coherencia (es decir, usando la misma fórmula y enfrentando el problema del idioma). Además, tenemos la intención de experimentar con la reducción de espacio no lineal como la escalabilidad multidimensional para encontrar más similitudes semánticas.

Agradecimientos

Los autores deseamos agradecer a la Universidad Abierta de Cataluña por haber proporcionado los materiales y el contexto necesario para el desarrollo de esta investigación, y por financiar parcialmente este trabajo en el marco del Proyecto de Innovación Docente número IN-PID 1043. Nos gustaría agradecer especialmente a Germán Cobo, David García, Jordi Durán, Francisco Cortés, Lluis Villarejo y Rafael E. Banchs por su apoyo para este trabajo. Esta investigación también ha sido parcialmente financiada por el Séptimo Programa Marco de la Comisión Europea a través de la Beca Internacional de Movilidad Marie Curie (IMTraP-2011-29951).

Referencias

Banchs, R. E., & Costa-jussˆ, M. R. (2010). A non-linear semantic mapping technique for cross-language sentence matching. *Lecture Notes in Computer Science, Springer, 7th International Conference on Natural Language Processing (IceTAL)* (pp. 57-66). Iceland.

Banchs, R. E., & Costa-jussˆ, M. R. (2011, June). *A semantic feature for statistical machine translation*. ACL HLT: 5th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5), Portland.

Brown, S., Race, R., & Bull, J. (1999). *Computer-assisted assessment in higher education.* Kogan Page.

Deerwester, S., Dumanis, S., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407.

Dumais, S., Landauer, T., & Littman, M. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR 1996 Workshop on Cross-Lingual Information Retrieval*.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999b). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic J. of Computer-Enhanced Learning*. Retrieved from http://imej.wfu.edu/articles/1999/ 2/04/index.asp

Greenacre, M. (2011). *It had to be U – The SVD song*. Available from http://www.youtube.com/tch?v=JEYLfIVvR9I&feature=player_detailpage.

Hidekatsu, K., Kiyoshi, A., Chiharu, I., Nagatomo, N., & Shinya, W. (2007). Toward a software development model for automatic marking software. *Proceedings of the 35th Annual ACM SIGUCCS Conference on User Services.* Orlando, Florida, USA.

Hofmann (1999). Probabilistic latent semantic analysis. *Proceedings of Uncertainty in Artificial Intelligence, UAI99I* (pp. 289-296).

Hussein, S. (2008). Automatic marking with Sakai. *Proceedings of the 2008 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries: Riding the Wave of Technology.* Wilderness, South Africa.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society.* Mawhwah, NJ. Erlbaum.

Lemaire, B., & Dessus, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, *24*, 305–320.

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research, 29*(4), 495-512.

PŽrez, D., Alfonseca, E., Rodr'guez, P., & Pascual, I. (2006). Willow: Automatic and adaptive assessment of students free-text answers. In *Proceedings of the 22nd International Conference of the Spanish Society for the Natural Language Processing* (SEPLN).

Quah, J.T-S., Lim, L-R., Budi, H., & Lua, K-T. (2009). Towards automated assessment of engineering assignments. *Proceedings of the 2009 International Joint Conference on Neural Networks* (pp. 1411-1418). Atlanta, Georgia, USA.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D. Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*, 337-354.

Salton, G. (1989). *Automatic text processing: The transformation, analysis and retrieval of information by computer*. Addison-Wesley.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Steinhart, D. (2000). *Summary street: An LSA based intelligent tutoring system for writing and revising summaries* (Ph.D. thesis). University of Colorado, Boulder, Colorado.

Su, W., Wang, P., Li, L,. Li, G., Zhao, Y. (2006). MathEdit, a browser-based visual mathematics expression editor. *Proceedings of The 11th Asian Technology Conference in Mathematics (*pp. 271-279.). Hong Kong.

Tyner, K. (1999). *Development of mental representation: Theories and applications.* Lawrence Erlbaum Associates.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). Approximate natural language understanding for an intelligent tutor. In *Proc. of the 12th Int'l Artificial Intelligence Research Symposium* (pp. 172–176). Menlo Park, CA, USA. Wiemer-Hastings, P., & Graesser, A. (2000). Select-a-kibitzer: A computer tool that gives meaningful feed-back on student compositions. *Interactive Learning Environments*, *8*, 49–169.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes, 25*(2-3), 309-336.

Autoras

Mireia Farrús

Marta R. Costa-Jussà