

Diana Pérez Marín, Enrique Alfonseca Cubero, Pilar Rodríguez Marín

Dpto. de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid

{diana.perez, enrique.alfonseca, pilar.rodriguez}@uam.es

## 1. Introducción

En los últimos años hemos asistido a una auténtica revolución en el campo de la enseñanza. La popularización de los cursos e-learning con la posibilidad de estudiar desde cualquier ordenador conectado a Internet y en cualquier momento ha abierto innumerables posibilidades a los estudiantes que ya no dependen del profesor sino que controlan su propio ritmo de aprendizaje.

Desde los inicios de la historia de la pedagogía se ha destacado la importancia de la evaluación en la educación. Por lo tanto no es de extrañar que el empuje de la educación on-line traiga consigo el florecimiento de un campo asociado: la Evaluación Automática por ordenador (*Computer Assisted Assessment*, CAA o *e-assessment*) que tiene como objetivo estudiar cómo el ordenador se puede usar para evaluar eficazmente el proceso de aprendizaje.

Se puede fijar el inicio de este campo en los años setenta cuando se empezaron a usar tests que, aunque los estudiantes rellenaban manualmente, se evaluaban usando reconocedores ópticos permitiendo una evaluación casi instantánea. Sin duda, ha habido una gran evolución desde entonces, y actualmente los ordenadores se aplican a muchos más tipos de evaluación, algunos de los cuales son: evaluación basada en iguales (*peer assessment*) en la que los ordenadores proporcionan programas para que los estudiantes puntúen a otros estudiantes de su mismo nivel; evaluación cara-a-cara (*face-to-face assessment*) en el que un sistema CAA se usa en la clase para moderar interactivamente debates sobre la lección a estudiar; evaluación basada en confianza (*confidence-based assessment*) en el que los estudiantes no sólo tienen que responder a la pregunta sino también declarar cómo de seguros están de que han dado la respuesta correcta; y evaluación bajo demanda (*assessment on demand*) en el que los estudiantes deciden cuánto quieren practicar.

No solamente los ordenadores se aplican a más tipos de evaluación, sino que también pueden puntuar más tipos de preguntas, desde las típicas de elección múltiple o de señalar sobre un esquema o dibujo la solución, a las preguntas abiertas. Muchos autores coinciden en que las preguntas de elección múltiple no son capaces de evaluar completamente si el estudiante ha comprendido los conceptos expuestos en la lección;

# ¿Pueden los ordenadores evaluar automáticamente preguntas abiertas?

**Resumen:** Tradicionalmente la sección de evaluación de la mayoría de los cursos on-line se basaba únicamente en preguntas de elección múltiple. Sin embargo, según la opinión generalizada de muchos investigadores, educadores y psicólogos restringirse exclusivamente a preguntas cerradas no permite evaluar completamente las habilidades cognitivas de los estudiantes. Esto ha tenido como consecuencia la creación del campo conocido como Evaluación Automática de preguntas abiertas. Lo que se plantea en este artículo es si realmente funciona. Esto es, ¿cuán fiable es un ordenador como evaluador automático de respuestas en texto libre escritas por estudiantes? Para dar respuesta a este interrogante hemos revisado la historia de este campo y visto como en los últimos años, estos sistemas han empezado a usarse como comprobadores de las notas puestas por los profesores, para asegurar una correcta evaluación. En todo caso no se puede olvidar que los ordenadores (al menos por ahora) no son más que máquinas sin sentido común ni inteligencia propia, lo que les impide enfrentarse con éxito a respuestas demasiado originales que quizás estén bien pero se salen de lo comúnmente establecido como correcto. El desafío está propuesto y, en nuestra opinión, mantener los objetivos realistas será lo que consiga que este campo avance, con paso firme, despejando cualquier incertidumbre sobre su validez.

**Palabras clave:** educación a distancia, evaluación por ordenador, preguntas abiertas, recursos docentes, técnicas de procesamiento de lenguaje natural.

por lo tanto, ampliar la Evaluación Automática por ordenador a preguntas que esperan una respuesta en texto (ya sea corto o un ensayo) ha recibido una gran atención por parte de la comunidad científica internacional en los últimos años. Instituciones educativas y gubernamentales como la *Scottish Qualifications Authority* (SQA), el centro CAA en el Reino Unido o el departamento de educación de EE.UU. han expresado su apoyo a este campo. De hecho, exámenes como el *Graduate Management Admissions Test* (GMAT), *Test of English as a Foreign Language* (TOEFL) o el *Graduate Record Examination* (GRE) ya están incluyendo preguntas de este tipo en su versión electrónica.

La motivación de la existencia de este campo está bien justificada, pero la cuestión ahora es su viabilidad. Para responder a esta pregunta, se comenzará haciendo un repaso de la historia de los sistemas de evaluación automática de respuestas en texto libre. A continuación se comentarán las principales puntos a favor de usar estos sistemas, los retos que aún hay que superar, y, finalmente, se responderá el interrogante planteado en las conclusiones.

## 2. Revisión histórica

La evaluación automática de respuestas en texto libre por ordenador es una disciplina con casi cuarenta años de existencia como se puede observar en la **figura 1**.

El pionero de este campo es Ellis Page, un profesor de inglés, que en 1966 fue la primera persona en creer posible la evaluación automática de respuestas en texto libre. Page

desarrolló su sistema al que llamó **Project Essay Grader** (PEG) y que se basaba en buscar ciertas características en la respuesta del estudiante como el número de palabras de longitud mayor a tres que usaba, el número de palabras distintas, etc. Este enfoque no convenció a la comunidad científica y su sistema fue muy criticado durante las siguientes décadas. De hecho, la investigación en el campo se paró casi por completo durante casi treinta años. No fue hasta principio de los años noventa, cuando un grupo de investigadores del centro *Educational Testing Service* (ETS) empezaron a aplicar técnicas de *Procesamiento de Lenguaje Natural* (PLN), que resurgió de nuevo el interés por este campo.

El sistema que desarrollaron se llamó **Educational Testing Service I** (ETS I). Aunque sólo se concentraba en el estilo de la respuesta (sin considerar el contenido todavía) y trabajaba sobre frases cortas (menos de 20 palabras), fijó el inicio del sistema **E-rater** [3], basado en un enfoque híbrido que combina técnicas de PLN y estadísticas. El proceso de evaluación con **E-rater** comienza con la creación del modelo contra el que se va a comparar la respuesta en texto libre del estudiante. Para ello, hay una fase inicial de entrenamiento en la que a **E-rater** se le introducen al menos doscientos textos relacionados con el tema a evaluar. Cuando el estudiante introduce su respuesta en el sistema, ésta es analizada sintáctica y retóricamente para identificar sus estructuras más relevantes y hacer un análisis de las características que según el modelo creado en el entrena-

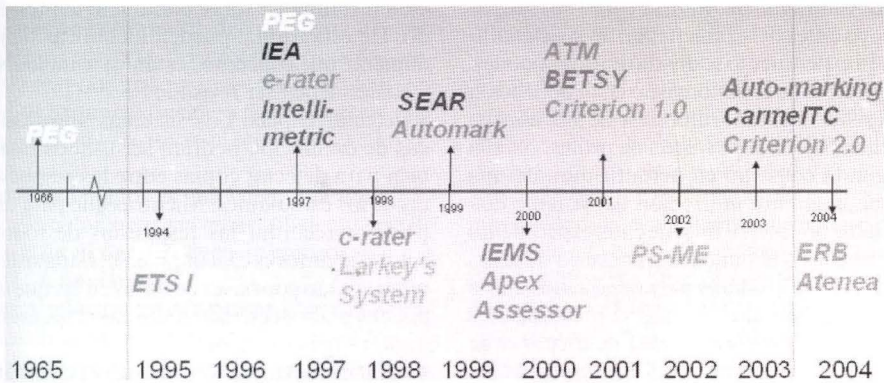


Figura 1: Línea temporal de los sistemas de evaluación automática de respuestas en texto libre.

miento, finalmente proporcione la nota que obtiene el estudiante.

Desde finales de los noventa se produjo un auge espectacular con avances que marcan la evolución de sistemas inicialmente sólo para investigación hacia auténticas aplicaciones comerciales. Por ejemplo, *E-rater* se usa con éxito desde 1999 en la evaluación del examen GMAT (con un porcentaje de acuerdo con las notas del profesor del 97%), ha sido comprado por más de 200 instituciones y usado por más de 50.000 usuarios

A partir de entonces, cada año aparecen del orden de dos o tres nuevos sistemas para evaluar automáticamente respuestas en texto libre basándose en distintas técnicas de PLN. ETS complementa *E-rater* con *C-rater* que se concentra en evaluar el contenido de la respuesta del estudiante. Igual que con *E-rater* es necesario inicialmente construir el modelo de referencia contra el que comparar las respuestas de los estudiantes. Con la diferencia de que en el caso de *C-rater* el número de textos de entrenamiento puede ser inferior ya que *C-rater* no trata de forma global toda la respuesta sino que se concentra en información específica que debe contener la respuesta del estudiante para ser considerada correcta. *C-rater* normalmente se usa integrado en sistemas *e-learning* para evaluar las preguntas de revisión que aparecen en la sección final de cada lección. En particular, en un estudio realizado a gran escala para evaluar 170.000 respuestas cortas a 19 preguntas de comprensión lectora y 5 preguntas de álgebra, *C-rater* obtuvo un 85% de acuerdo con las notas puestas por el profesor [2].

Otros sistemas que aparecen a finales de los noventa son: el sistema de Larkey y el *Intelligent Essay Assessor* (IEA). El sistema de Larkey se basa en técnicas de categorización de textos (*Text Categorization Techniques*, TCT). Los profesores deben indicar las características a buscar en las respuestas de los estudiantes, de forma que en función de su presencia o ausencia éstas sean clasificadas como correctas o incorrectas. Este sistema ha sido usado en ensayos

sobre estudios sociales, legales y físicos consiguiendo un 60% de acuerdo total con las notas de los profesores y un 100% de acuerdo con un punto de diferencia [7]. *Intelligent Essay Assessor* fue desarrollado en la universidad de Colorado y utiliza Análisis de Semántica Oculta (*Latent Semantic Analysis*, LSA). El primer paso de la evaluación consiste en una fase de entrenamiento para construir el modelo LSA a partir de un conjunto de documentos relacionados con el tema a evaluar. De esta forma, la respuesta del estudiante se transforma en la representación LSA que se compara con el modelo LSA construido en la fase anterior. IEA se ha aplicado a textos de psicología, medicina e historia alcanzando hasta 90% de acuerdo total con los profesores [6].

En los últimos años se han presentado: *Intelligent Essay Marking System* (IEMS) que usa reconocimiento de patrones en la respuesta del estudiante para evaluarla y se ha aplicado a resúmenes realizados por estudiantes de tercer curso de Ingeniería Mecánica sobre un texto de ciencia ficción consiguiendo un 80% de correlación [9].

*Automated Text Marker* (ATM) que emplea Extracción de Información para buscar conceptos y sus dependencias en el texto. Los profesores han introducido previamente unas plantillas con los conceptos que se deben buscar y las relaciones entre ellos, y posteriormente mediante reconocimiento de patrones se comparan los conceptos y las dependencias encontradas en las respuestas de los estudiantes con las modelo de la plantilla del profesor. Aún no se han publicado experiencias del uso de ATM [4].

*Paperless school free-text marking engine* (PS-ME) basado en el uso de PLN, requiere un entrenamiento inicial con al menos treinta textos evaluados a mano por profesores incluyendo tanto respuestas correctas (puntuación alta) como incorrectas (puntuación baja) y tiene en cuenta la frecuencia de adjetivos, adverbios y patrones sintácticos dentro de la respuesta del estudiante para compararlos con las respuestas modelo de los profesores. El sistema se ha aplicado en el

campo académico al examen GCSE (*General Certificate of Secondary Education*) y *National Curriculum Grade* en Inglaterra, y en el comercial ha sido utilizado por editores. Los autores destacan su preocupación en cuanto a que el sistema puntúa mejor las respuestas más parecidas a las respuestas modelo mientras que los profesores normalmente puntúan mejor una respuesta original de un alumno que una basada en un texto modelo. Además, también advierten de la dificultad de la selección de dichos textos modelo [8].

**Automark** que emplea técnicas de PLN y de Extracción de Información para realizar un análisis del texto de la respuesta del estudiante según plantillas predefinidas completadas por los profesores con lo que se considera correcto para superar la pregunta. El sistema ha sido usado para evaluar las respuestas de estudiantes de 11 años al examen nacional de ciencia en Inglaterra consiguiendo correlaciones entre 93% y el 96% con las notas puestas por los profesores [10].

*Schema Extract Analyze and Report* (SEAR) que también emplea técnicas de Extracción de Información para comparar plantillas predefinidas completadas por los profesores con las plantillas que se generan a partir de las respuestas de los estudiantes. El sistema se ha aplicado a ensayos en historia consiguiendo una correlación del 30%. Según sus autores el principal problema encontrado ha sido la construcción de las plantillas predefinidas de los profesores [5].

**Auto-marking** que combina análisis sintáctico, Extracción de Información y reconocimiento de patrones, comparando también la plantilla generada con el estudiante con la predefinida por el profesor. El sistema se ha aplicado a respuestas del examen GCSE de Biología en Inglaterra con un 88% de acuerdo exacto con las notas puestas por el profesor. Los autores destacan la limitación del sistema de tratar con inferencias realizadas por los estudiantes y en general cualquier información que no aparezca en las plantillas predefinidas [12].

**CarmelTC** que emplea redes bayesianas y árboles de decisión. Mediante un sistema automático se identifican las características que se consideran como correctas en las respuestas de los estudiantes. A partir de ellas se puede clasificar el texto y de esta forma asignarle una calificación. El sistema se ha usado con 126 ensayos de física, con un 90% de precisión y 80% de cobertura [11].

**Atenea**, que combina técnicas de PLN y estadísticas [1]. Para cada pregunta almacenada en el sistema se pide a varios profesores distintos que introduzcan la respuesta correcta que debe introducir el alumno para superar la pregunta. Estas respuestas correctas o referencias quedan almacenadas en

la base de datos del sistema, de forma que cuando el alumno introduce su respuesta, ésta se compara con las referencias apropiadas. Para facilitar la comparación se aplican varias técnicas tanto a la respuesta del estudiante como a las referencias: lematización para ignorar discrepancias morfológicas; filtrado de palabras no significativas como preposiciones, determinantes, etc.; desambiguación de sentidos para palabras polisémicas y/o reconocimiento de sinónimos. Finalmente, la respuesta procesada del estudiante es comparada con las referencias mediante el algoritmo llamado ERB (*Evaluando Respuestas con Bleu*). La salida del algoritmo es un valor numérico entre 0 y 1, que posteriormente se escala al rango de notas proporcionado por el profesor. Además, ERB también marca en la respuesta del estudiante las coincidencias de trigramas (bloques de tres palabras) con verde oscuro, de bigramas (bloques de dos palabras) con verde y unigramas (palabras) con verde claro. Es decir, el brillo del color verde como color de fondo de la respuesta del alumno está correlacionado con la densidad de n-gramas de la respuesta del alumno que aparecen en las referencias escritas por el profesor. La figura 2 muestra un ejemplo de página de feedback generada por Atenea en la que se observa que la nota que ha obtenido el alumno ha sido un 1.0 y la gradación de verdes en el texto de la respuesta corregida del estudiante. Es importante destacar que no es necesaria una fase de entrenamiento previo ni que los profesores introduzcan las referencias ya evaluadas. Únicamente se pide a los profesores que califiquen a mano las preguntas cuando se quiere comprobar el correcto funcionamiento de Atenea y estudiar la correlación entre las notas automáticas y las del sistema (correlación que ha alcanzado un valor de hasta un 56%). Además, Atenea ha sido usada por 32 estudiantes de la asignatura de Sistemas Operativos de Ingeniería de Telecomunicaciones de la Escuela Politécnica Superior de la Universi-

dad Autónoma de Madrid como sistema de apoyo antes del examen final de la asignatura. El experimento se dividió en dos sesiones, la primera de ellas consistió en la utilización del sistema durante una clase de una hora en el aula bajo la supervisión de profesores y la segunda consistió en permitir durante una semana la libre utilización del sistema disponible on-line. Al final de cada experimento se pasó a los alumnos una encuesta de satisfacción. De los datos proporcionados en las encuestas se concluye que a los estudiantes les pareció atractiva la idea de disponer de un sistema que les evalúa de forma interactiva y que además les presenta retroalimentación instantánea (característica que puntuaron como una de las más interesantes del sistema). De hecho, 29 de los 32 estudiantes no sólo dijeron que usarían el sistema para reforzar conceptos antes del examen final sino que también lo recomendarían a otros estudiantes tanto de la misma asignatura como de otras asignaturas. Los tres estudiantes que no la recomendarían argumentaron que el sistema tiene un método de corrección muy estricto y que creen que debería incorporar más inteligencia artificial.

### 3. Puntos a favor

■ **Conseguir un proceso de evaluación más justo y eficiente:** El ordenador no se cansa, aburre, irrita o pierde la concentración. Además, es capaz de examinar y analizar el texto en mucho más detalle que un humano y es libre de cualquier tipo de prejuicio.

■ **Permitir que los profesores se concentren en la preparación de sus clases y realización de más ejercicios:** La cantidad de tiempo que dedica cualquier profesor a evaluar a sus alumnos es muy elevada. Gracias al uso de mecanismos automáticos de corrección se permite que los profesores se concentren en otras tareas.

■ **Proporcionar más feedback a los estudiantes:** Puntuar de forma automática permite que los estudiantes reciban informa-

ción instantánea sobre cómo han contestado. De esta forma, pueden identificar rápidamente qué conceptos están ya asimilados y cuáles deben ser revisados.

■ **Detectar copias:** Los sistemas automáticos de evaluación podrían ser usados también para detectar copias entre las respuestas de los estudiantes. Mientras que para un profesor recordar las respuestas de todos sus estudiantes es casi imposible, estos sistemas cuentan con bases de datos en las que se pueden ir almacenando todas las respuestas y así ser más sencillo.

■ **Retar a los estudiantes:** La idea es que si un estudiante es capaz de obtener buena nota engañando al sistema, en realidad, este estudiante se merece la nota conseguida, puesto que descubrir el mecanismo de evaluación requiere un gran dominio del tema en estudio.

■ **Reducir costes:** Una vez pasada la fase de implantación del sistema, y en escuelas en las que se pueda utilizar en varios departamentos, la reducción en el tiempo que dedican los profesores a corregir y la posibilidad de generar aleatoriamente distintas secuencias de las preguntas sin tener que imprimir varias versiones del examen abaratan el proceso.

■ **Amenizar el proceso de evaluación:** Mientras que tradicionalmente el estudiante debe responder y luego esperar varios días hasta que el profesor le devuelva la corrección, con los sistemas automáticos el estudiante se ve inmerso en un proceso interactivo. El sistema va eligiendo qué pregunta realizarle e inmediatamente proporcionándole feedback y una nueva pregunta para que continúe respondiendo. Así se consigue mantener al alumno más tiempo atento y animado para continuar con las siguientes preguntas.

### 4. Retos

■ **Para algunas personas el ordenador no resulta creíble en su papel de evaluador automático:** Puesto que, independientemente de la técnica o técnicas que usen estos siste-

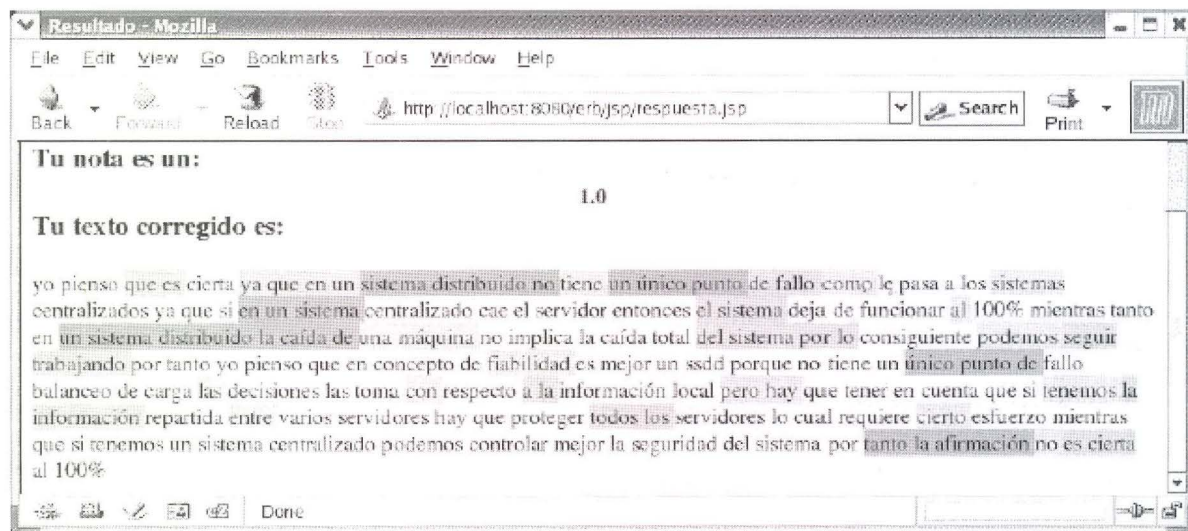


Figura 2. Ejemplo de pantalla de feedback del sistema Atenea.

mas, la gran mayoría acaba finalmente comparando la respuesta del alumno con la respuesta correcta del profesor, aquellas respuestas que aunque correctas se salgan fuera de lo que se considera normal y aceptable, al ser más creativas u originales, serán rechazadas.

■**El ordenador es una “máquina tonta”:** El ordenador sigue sin tener suficiente sentido común ni una inteligencia comparable a la de un humano.

■**La relación estudiante-profesor se debilita:** Si el profesor deja de ser quién lee las respuestas de los estudiantes y las evalúa, entonces pierde información sobre el progreso de sus alumnos y qué puntos debe ir reforzando en las clases.

■**La evaluación que puede realizar un ordenador está limitada:** Hay que considerar que algunos sistemas podrían no ser capaces de adaptarse a las necesidades particulares de un individuo, clase, profesor o institución.

■**Sobre su aplicación en exámenes de selección:** Se podrían clasificar las pruebas de evaluación en dos grandes tipos: las pruebas de entrenamiento (*formative assessment*) y las pruebas de selección (*summative assessment*). La utilidad de estos sistemas en pruebas de entrenamiento es demostrable. Sin embargo, no está tan claro su eficacia en pruebas de selección. De hecho, ni siquiera es evidente que se puedan utilizar, ya no sólo porque un pequeño error del mecanismo de evaluación significaría que un alumno queda fuera o repite curso, sino por posibles problemas legales a la hora de defender la puntuaciones así conseguidas ante un tribunal.

■**Problemas de seguridad y fiabilidad:** Tanto el software como el hardware han mejorado mucho en las últimas décadas y ahora trabajamos con ordenadores más resistentes y fiables. Sin embargo, todavía existe el miedo generalizado ante fallos no sólo del programa o del hardware, sino ante problemas de seguridad.

## 5. Conclusiones

La evaluación automática por ordenador tiene muchas posibilidades de aplicación, y en los últimos años se ha trabajado mucho en su investigación. Hoy en día ya se usan muchos de estos sistemas de evaluación automática de respuestas de texto libre comercialmente, y algunos creadores de estos sistemas se plantean la completa automatización del proceso de evaluación.

Sin embargo, su aplicación aún no se ha generalizado, muchos profesores ni siquiera conocen de su existencia y algunos de los que los conocen se muestran reticentes a aceptar que los “fríos y tontos” ordenadores puedan ser capaces de evaluar, sin ayuda ni supervisión, respuestas escritas por humanos. Además, también está de fondo el miedo a fallos informáticos y problemas de seguridad, especialmente en departamentos no técnicos en los que ni los profesores ni los técnicos

están familiarizados con los sistemas informáticos.

En todo caso, la mayoría de los retos técnicos existentes son superables, ya que nos encontramos en pleno proceso de popularización del uso de los ordenadores en las aulas, lo que irá facilitando el camino a la posibilidad de evaluación automática, y en concreto a la evaluación automática de respuestas en texto libre. Son más difíciles los problemas relacionados con la desconfianza que inspira la falta de sentido común y de inteligencia de los ordenadores, para reconocer muchas respuestas distintas pero válidas como correctas. Por ello hay que distinguir entre la evaluación orientada al entrenamiento del alumno de la evaluación orientada a puntuar alumnos en procesos de selección.

Para la evaluación orientada al entrenamiento, estos sistemas son clave, puesto que se pueden integrar en la sección de evaluación de los sistemas e-learning, y permitir a los profesores dar más ejercicios a los alumnos que son corregidos inmediatamente. Además, los alumnos no tienen que esperar hasta que el profesor se los pueda devolver corregidos, sino que reciben feedback instantáneo que les indica qué puntos deben revisar para su examen final.

Para la evaluación orientada a la selección, aún estos sistemas se deberían usar con cuidado, puesto que a pesar de que algunos sistemas están obteniendo correlaciones con las puntuaciones dadas por los profesores comparables a las que tienen dos profesores al evaluar el mismo conjunto de preguntas, el riesgo es alto al ser estos procesos decisivos para los alumnos: quién pasará al próximo año, quién entrará en la universidad y en qué facultad... Esto no contradice que se puedan usar y que de hecho, ya se estén usando, como segunda puntuación en los exámenes de mayor relevancia. Sirven de comprobación a las notas que están dando los profesores y únicamente cuando la nota del profesor es muy distinta a la del sistema, ésta tiene que ser revisada por un segundo profesor, lo que garantiza una mayor transparencia y justicia del proceso de evaluación.

La tecnología existe, los sistemas están consiguiendo buenos resultados y el campo está recibiendo el apoyo de importantes centros y organizaciones educativas, por lo que es de esperar un futuro prometedor para esta área en crecimiento con interés de parte de los investigadores, educadores y ejecutivos siempre que los objetivos se mantengan realistas y nadie intente reemplazar completamente la figura del profesor con un ordenador.

## 6. Agradecimientos

Este trabajo ha sido financiado por el proyecto TIN2004-0314 del Ministerio de Educación y Ciencia española.

## Referencias

- [1] E. Alfonseca, D. Pérez. *Automatic Assessment of Short Questions with a BLEU-inspired Algorithm and shallow NLP*, en *proceedings* de la conferencia internacional "Avances en Procesamiento de Lenguaje Natural", Alicante (España), Springer Verlag, LNCS 3230, págs. 25-35, 2004.
- [2] J. Burstein, C. Leacock, R. Swartz. *Automated evaluation of essays and short answers*, en *Proceedings* de la conferencia internacional de Evaluación Automática por Ordenador (*Computer Assisted Assessment*, CAA), Loughborough (Inglaterra), 2001.
- [3] J. Burstein, M. Chodorow, C. Leacock. *Criterion online essay evaluation: An application for automated evaluation of student essays*, en *Proceedings* de la decimoquinta conferencia internacional de Aplicaciones Novedosas de Inteligencia Artificial (IAAI), Acapulco (Méjico), 2003.
- [4] D. Callear, J. Jerrams-Smith, V. Soh. *CAA of short non-MCQ answers*. En *proceedings* de la quinta conferencia internacional de Evaluación Automática por ordenador (*Computer Assisted Assessment*, CAA), Loughborough (Inglaterra), 2001.
- [5] J. Christie. *Automated essay marking for content – does it work?* En *Proceedings* de la séptima conferencia internacional de Evaluación Automática por ordenador (*Computer Assisted Assessment*, CAA), Loughborough (Inglaterra), 2003.
- [6] P. Foltz, D. Laham, T. Landauer. The intelligent essay assessor: applications to educational technology, *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2), 1999. <<http://imej.wfu.edu/articles/1999/2/04/index.asp>>.
- [7] L.S. Larkey. *Automatic essay grading using text categorization techniques*. En *proceedings* de la veintiún anual e internacional ACM SIGIR conferencia en Investigación y Desarrollo en Recuperación de Información, págs. 90-95, Melbourne (Australia), 1999.
- [8] O. Mason, I. Grove-Stephenson. *Automated free-text marking with paperless school*, en *Proceedings* de la sexta conferencia internacional de Evaluación Automática por ordenador (*Computer Assisted Assessment*, CAA), Loughborough (Inglaterra), 2002.
- [9] Y. Ming, A. Mikhailov, T. Kuan. *Intelligent essay marking system*, Learners Together, NgeeANN Polytechnic, Singapore, 2000.
- [10] T. Mitchell, T. Russell, P. Broomhead, N. Aldridge. *Towards robust computerised marking of free-text responses*, en *proceedings* de la sexta conferencia internacional de Evaluación Automática por ordenador (*Computer Assisted Assessment*, CAA), Loughborough (Inglaterra), 2002.
- [11] C. Rosé, A. Roque, D. Bhembe, K. VanLehn. *A hybrid classification approach for analysis of student essays*, en *proceedings* del workshop de Tecnologías del Lenguaje Humano y la Asociación Americana de Lingüística Computacional (HLT-NAACL) en creación de herramientas educativas usando PLN, Sapporo (Japón), 2003.
- [12] J. Sukkarieh, S. Pulman, N. Raikes. *Auto-marking: using computational linguistics to score short, free-text responses*, en *proceedings* de la vigésimonoventa conferencia de la asociación internacional de evaluación educativa (IAEA), Manchester (Inglaterra), 2003.