SISTEMA DE AUTO SOPORTE PARA EL MEJORAMIENTO DE LAS PRÁCTICAS DOCENTES, PROYECTO FONDEF D11I1009[ix]

Self-Support System to Help Improve Teachers' Practices

Dartnell, Pablo[a]. Araya, Roberto[a]. Bahamondez, Manuel[a]. Hernández, Josefina[a]. Aguirre, Carlos[a]. Castillo, Roberto[a]. Rodríguez, Patricio[a]. Palavicino, María Angélica[a]. Jiménez, Abelino[a]. Van der Molen, Johan[a].

[a]Centro de Investigación Avanzada en Educación, Universidad de Chile; correos electrónicos: dartnell@ciae.uchile.cl, roberto.araya@ciae.uchile.cl, manuel@bahamondez.com, josefina.hernandez@ciae.uchile.cl, carlosaguirre@automind.cl, castillo.n.r.e@gmail.com, prodriguez@ciae.uchile.cl, map@ciae.uchile.cl, jvandermm@gmail.com, abjimenez@gmail.com.

Resumen

El siguiente proyecto de investigación y desarrollo presenta un sistema de retroalimentación continuo que permite al docente de Matemáticas mejorar sus prácticas de enseñanza al transcribir automáticamente el discurso del profesor de una sala de clases a texto y analizar automáticamente la transcripción de cada clase para identificar prácticas docentes y contenidos, logrando un nivel parecido al de un experto humano enfrentado a la misma tarea y con las mismas fuentes de información.

Palabras clave: *Práctica Docente, Retroalimentación, Aplicación Móvil, Aprendizaje de Máquina*

Abstract

This research presents a Self-Support System that helps mathematics teachers detect and improve the different practices they use in their classes. This is performed, by recording all spoken audio in a classroom, transcribing this audio to written text, and analyzing each text in search of 6 different mathematical contents and 17 teacher practices, achieving the same results of a human expert exposed to the same task with the same information.

Keywords:*Teacher Practices, Feedback, Mobile Application, Machine Learning*

INTRODUCCIÓN Y CONTEXTO

Lograr sólidos aprendizajes escolares en matemáticas, y ciencias en general, es de fundamental importancia para el desarrollo futuro y por esto, constantemente se buscan formas de fortalecer y enriquecer la labor docente. Desde los inicios del siglo veinte, investigadores han recopilado información sobre la interacción entre el docente y sus alumnos, por ejemplo, contando la cantidad de preguntas que hace el/la profesor/a a sus estudiantes o calculando la proporción de palabras hablas por el docente respecto a las palabras habladas por los estudiantes (Stevens, 1912). Desde 1946 se ha recogido información estadística sobre diferentes práctica pedagógicas de diversos métodos, como filmando las salas de clases mientras los profesores están en acción (National Education Association, 1946). Sin embargo, a pesar de que algunas prácticas específicas pueden ser útiles con estudiantes de manera individual (por ejemplo, este alumno aprenderá más en este momento si le doy retroalimentación o si lo dejo entender el contenido por su cuenta), la búsqueda de prácticas generales que optimicen la efectividad, eficiencia y atención de los estudiantes es un desafío mayor (Koedinger, et al. 2013).

[ix]Proyecto financiado porXIX Concurso de Proyectos FONDEF I+D 2011

Es difícil calificar apropiadamente una clase de manera automática, debido a la gran cantidad de variables presentes y la diversidad de contextos en la que ella pudo situarse. Sin embargo, la información oportuna respecto a los diversos factores que han estado presentes en sus clases previas puede permitirle a un profesor ajustar sus estrategias pedagógicas de acuerdo al contexto particular en que realiza su labor docente. Es así que proveer al profesor de retroalimentación puede fortalecer y enriquecer la práctica de la enseñanza. Existen diversos métodos para entregar retroalimentación al profesor, tales como la observación en aula o la revisión de videos de sus clases. En particular, actualmente las prácticas docentes son generalmente monitoreadas por observadores entrenados que utilizan rúbricas para asesorar interacciones que se dan en las salas de clases, o revisar videos previamente grabados de las clases (National Board Resource Center, 2010; Pianta, 2003; TIMSS, 2003). Sin embargo, este proceso es lento, tedioso y puede presentar errores, haciendo que sea prácticamente imposible de utilizar para entregar de forma masiva y permanente a los docentes sugerencias inmediatas de estrategias de enseñanza en el asesoramiento del desempeño y necesidades de aprendizaje de los estudiantes (Araya, et al, 2011).

Teniendo en cuenta aquello, el presente proyecto busca producir una herramienta generadora de información y retroalimentación constante e inmediata, con el fin de apoyar la labor de perfeccionamiento docente a través de la autoobservación. Con este objetivo en mente, dado que se busca generar un sistema accesible, se ha propuestoun sistema que comience por la captura del audio del profesor en la clase a través de una aplicación para "Smartphone", lo que permite contar con el contenido verbal de la clase para su estudio y análisis posterior. El sistema funciona de la siguiente forma: **1)** La aplicación registra la grabación del audio de una clase, **2)** Dicha grabación se transcribe automáticamente a texto, y **3)** Finalmente, el texto se analiza para identificar y cuantificar la presencia de los factores en observación. En esta oportunidad se considera la observación de seis contenidos matemáticos, presentes en el currículum nacional, y la de diecisiete prácticas docentes, escogidas por existir amplio acuerdo en su importancia como indicadores de calidad (Araya et al. 2011, Koedinger et al. 2013). El algoritmo de identificación y clasificación ha de ser entrenado antes de poder ser utilizado de manera autosuficiente, es por esto que en primera instancia se utiliza la colaboración de expertos que identifican y clasifican textos de clases de matemáticas que serán utilizados para el entrenamiento del sistema.

Una vez superada la fase de entrenamiento, se comparan los resultados de la identificación y clasificación generada por el algoritmo en nuevos textos y se observa el desempeño en comparación con los expertos. Es deseable que el sistema entregue niveles de acuerdo con los expertos similares o superiores a los exhibidos entre dichos expertos. En este reporte daremos cuenta de las primeras pruebas hechas con el sistema de auto soporte, entregando información respecto al acuerdo entre la información producida por el sistema y las clasificaciones de las mismas clases que expertos han realizado de manera independiente.

METODOLOGÍA

Datos

Para crear y entrenar el modelo de transcripción automática y el modelo de clasificación de la aplicación, 93 profesores registraron el audio de un total de 866 clases para este proyecto. Todos los registros de audio de clases fueron grabados, transcritos y clasificados a través de la aplicación móvil del proyecto. De cada audio de clase se utilizó un extracto de cinco minutos de duración, escogido de manera tal que fueran los cinco minutos con más palabras habladas de toda la clase (a través de la aplicación móvil, de forma automatizada). Se experimentó la captura del audio de profesores en salas de clases con diversos medios, para obtener audio de buena calidad, sin necesidad de procedimientos demasiado engorrosos. Finalmente se decidió que la mejor solución es registrar los audios usando micrófonos alámbricos de manos libres conectados al celular, por lo que la aplicación móvil no permite otra forma de registrar las clases.

Los audios de cinco minutos fueron transcritos manualmente, y luego expertos identificaron, en cada uno de dichos segmentos de clase, la presencia o ausencia de cada práctica y contenido buscado, apoyándose tanto en los textos como los audios para las clasificaciones. De los 866 registros de audio utilizados para entrenar el sistema, 309 clases fueron transcritas y clasificadas por tres expertos contratados para ello, y luego las siguientes 557 fueron transcritas y clasificadas por docentes que estaban utilizando la aplicación.

Para las transcripciones, los transcriptores escuchaban el audio, y debían ir corrigiendo la transcripción automática preliminar propuesta por una versión inicial del sistema, dejándola exactamente igual al audio de la clase.

Los contenidos matemáticos eran: álgebra, aritmética, ecuaciones, fracciones, geometría y proporcionalidad; y las prácticas eran (observar que son diferentes, pero no necesariamente mutuamente excluyentes entre ellas): aclaración de conceptos, realización de cálculos, clasificaciones de conceptos o contenidos, crear buen ambiente en el aula, dinamismo en la exposición, generar espacio para la discusión, trabajo con ejemplos, presentación de explicaciones, hacerle preguntas a los estudiantes, dar instrucciones, interacción con estudiantes, interpelación a estudiantes, nombrar a los estudiantes por sus nombres, desarrollo de razonamiento matemático, reforzamiento de contenidos o habilidades previamente cubiertas, y valoraciones negativas y positivas de los aportes o trabajo de los estudiantes.Para cada categoría, tanto de contenidos como de prácticas docentes, usando unaplataforma web especialmente construida para estos efectos, los clasificadores debían indicar, para cada texto a clasificar, si la categoría estaba presente o no.

Para la transcripción de los audios a texto, se contó con el apoyo de colaboradores del Instituto de Tecnologías del Lenguaje de Carnegie Mellon University (CMU) en el uso del software libre de captura y transcripción de voz, *CMUSphinx* versión 3 (desarrollado por ellos). El software fue adaptado para generar un modelo de lenguaje Castellano chileno, para lo cual se le incorporó información de fonemas asociados a nuestras formas locales de pronunciación y vocabulario, fundamentalmente asociado a las clases de matemáticas.

El identificador automático de contenidos y prácticas fue construido con el algoritmo de aprendizaje automático *Random Forest* (Breiman, 2001). Para esto, primero a cada texto se le eliminan las palabras que no representan contenido, sino más bien juegan un rol formal en el lenguaje, tales como artículos, pronombres, conectores, etc., y luego se agrupan las palabras según su raíz. Posteriormente, para cada categoría de clasificación, el algoritmo *Random Forest* cuenta cuántas veces se repiten las diferentes raíces en cada texto por separado, y a partir de las clasificaciones manuales, le asigna acada raíz de cada texto un peso (importancia), generando así una bolsa de palabras que se asocian a cada categoría de clasificación.

RESULTADOS

Transcripción Automática

Para el *Sphinx* entrenado, su efectividad se mide a través de la diferencia entre las transcripciones de los párrafos hechas por *Sphinx* y aquellas correctas (revisadas por las personas), mediante el indicador *Word Error Rate* (*Basic Concepts of Speech*, n.d.), que consiste en contabilizar y sumar 3 tipos de errores posibles, que son 1) Número de eliminaciones: texto borrado de la posición que corresponde; 2) Número de sustituciones: texto equivocado (sustituido por otro) en la posición que corresponde; y 3) Número de agregados: texto inexistente adicional. Todo esto se expresa como porcentaje sobre el tamaño del trozo de texto correctamente transcrito (transcripción humana), y el porcentaje de acierto logrado es 36.46%.

Identificación de contenidos matemáticos y prácticas docentes

El desempeño de un clasificador automático depende de su capacidad de replicar las identificaciones realizadas por el/los experto(s) en el conjunto de testeo. Para evaluar la clasificación automática, se utilizó el indicador *Alternative Chance-Corrected Coefficient* (AC1) (Blood & Spratt, 2007) que mide el nivel de concordancia entre los expertos y el clasificador automático, tomando valores entre 0 y 1, donde 0 es el menor nivel de concordancia y 1 el máximo. En particular, el AC1 es un índice que trata de corregir el acuerdo por azar.

Como fue mencionado anteriormente, las primeras 309 clases fueron clasificadas por 3 expertos cada una, y las restantes clases fueron clasificadas por una o más personas. Por ende, para los textos que clasificaron los tres expertos, primero se entrenó el sistema con la clasificación de cada experto por separado, utilizando 214 clases para entrenar y 95 para testear. Los niveles de concordancia entre los expertos y el clasificador automático (AC1) de los 3 entrenamientos se muestran en la Figura 1 y Figura 2.

Por consiguiente, se consolidaron las clasificaciones manuales de los tres expertos de manera tal que: (1) si dos o más expertos dicen que una categoría está presente en un texto, entonces se considera presente, mientras que (2) si solamente 1 o ninguno de los 3 dice que está presente, entonces no lo está. Así, se generó una "Clasificación Experta Consolidada", con la que se entrenó también el sistema. En la Figura 1 y Figura 2 se muestra el AC1 obtenido del entrenamiento del sistema con esta consolidación de los juicios expertos, junto a los asociados a cada experto por separado.

Se continuó entrenando el sistema con todos los textos disponibles, aunque no estuvieran clasificados por tres expertos cada uno. Para esto se consideró una categoría como presente en un texto si más de la mitad de los clasificadores de dicho texto considera que la categoría en cuestión está presente. Los niveles de concordancia entre la Base Final de Entrenamiento y el clasificador automático, con 866 textos clasificados por 53 personas en total, se presentan en la Figura 3 a continuación. Por último, utilizando también el indicador AC1, se calculó el acuerdo entre los tres Expertos al clasificar, de a pares y entre los tres. En la Figura N°4 se muestran los valores de AC1 obtenidos, junto a los resultados del entrenamiento final.

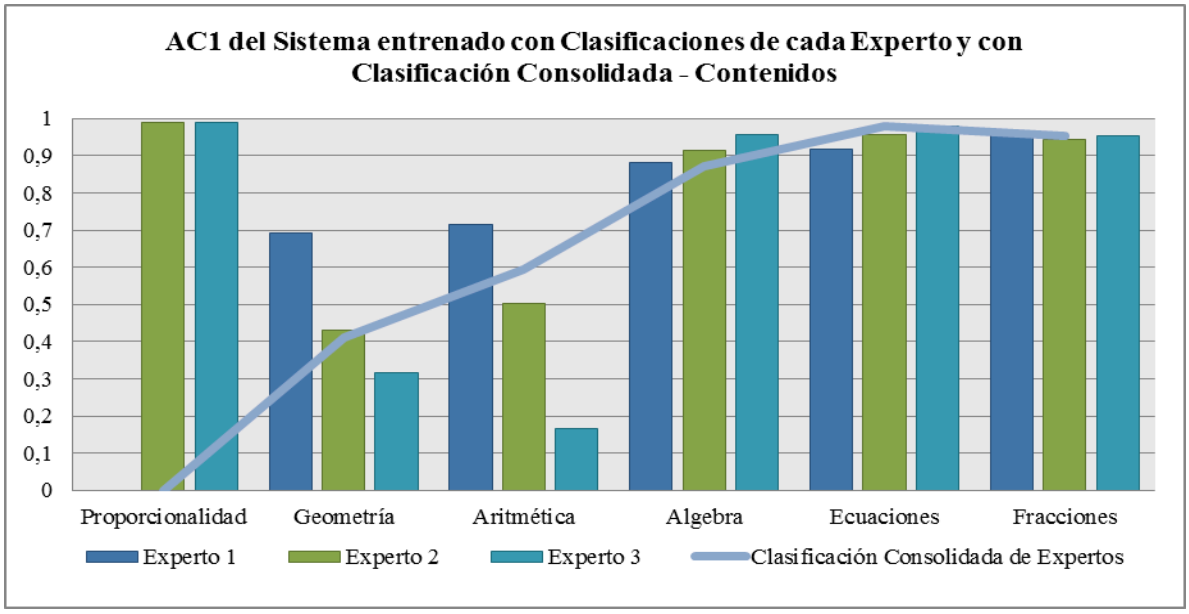

Figura 1. AC1 del Sistema entrenado con Clasificaciones de cada Experto vs. Clasificación Consolidada de los tres Expertos, para Contenidos

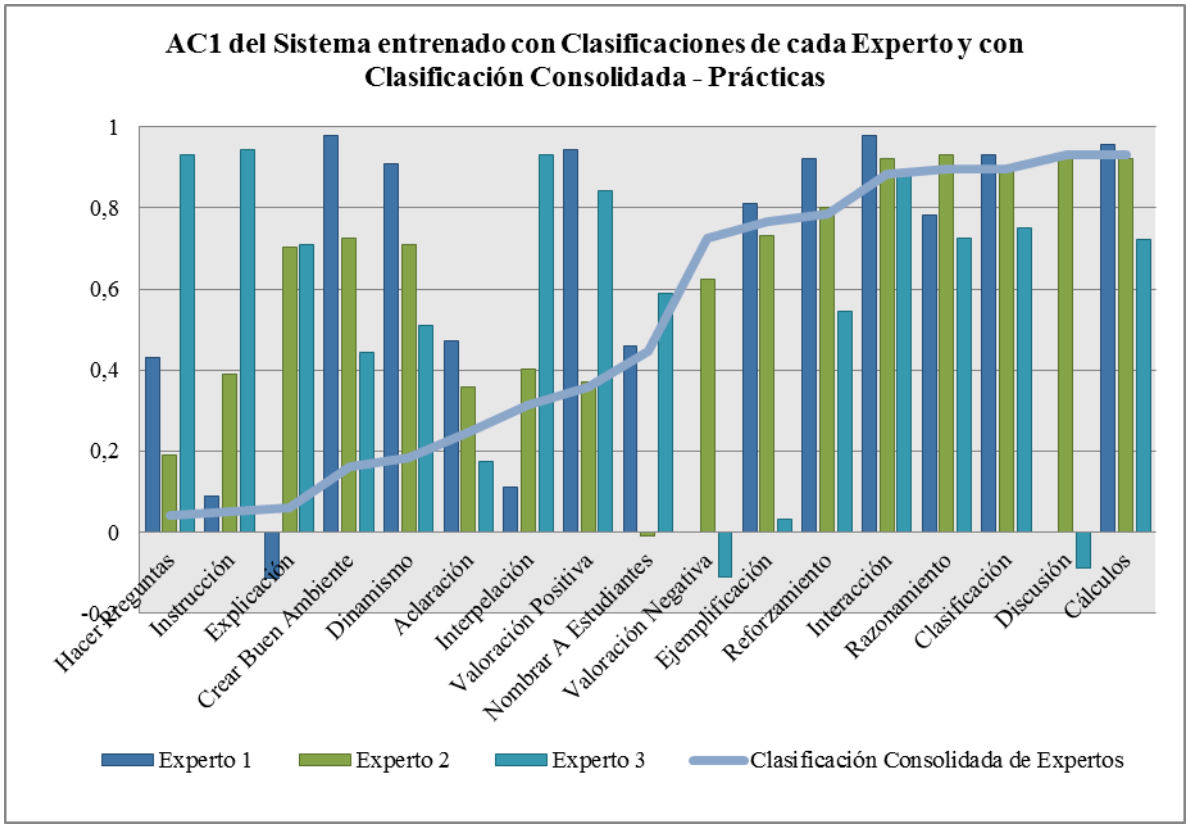

Figura 2.AC1 del Sistema entrenado con Clasificaciones de cada Experto vs. Clasificación Consolidada de los tres Expertos, para Prácticas

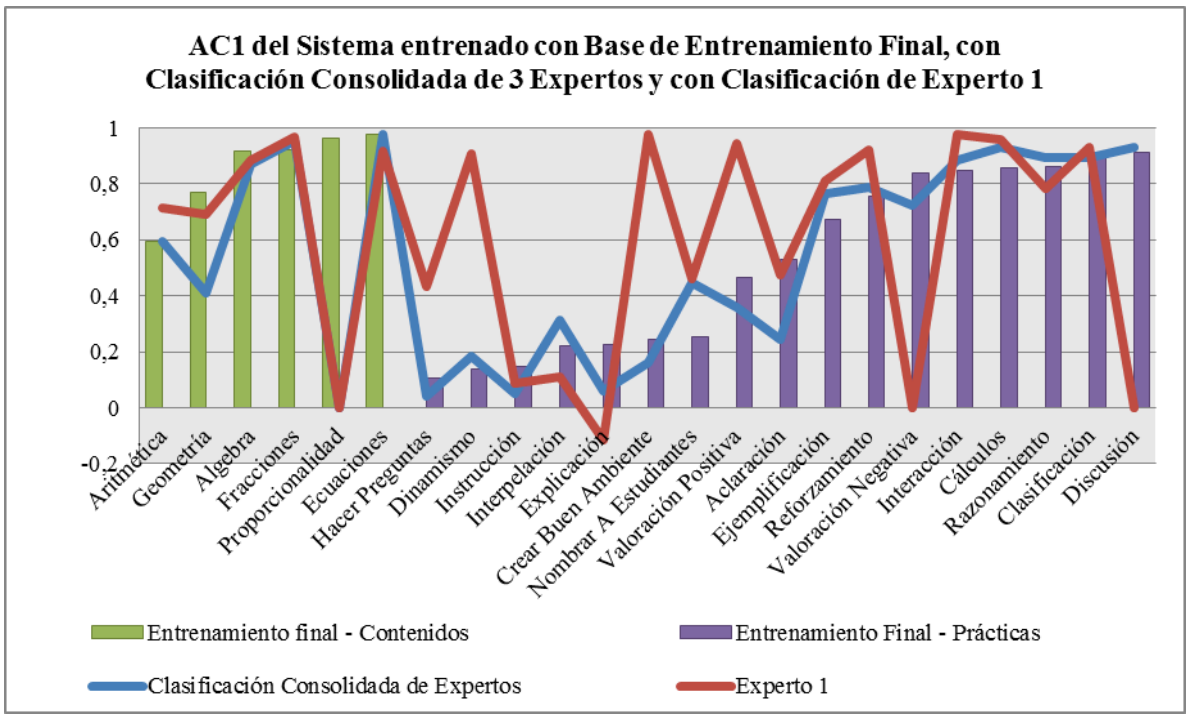

Figura 3.AC1 del Sistema entrenado con Base de Entrenamiento Final, con Clasificación Consolidada de los tres expertos y con Clasificación de Experto 1.

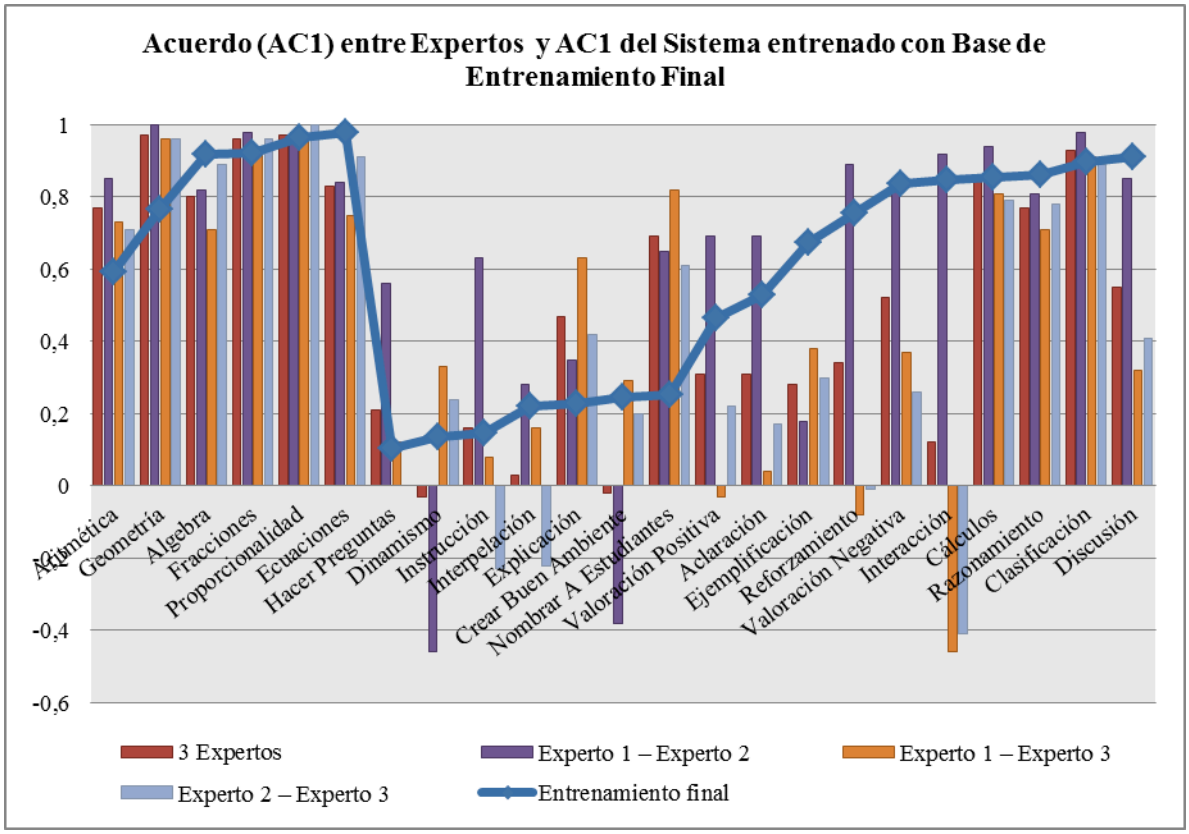

Figura 4. Acuerdo entre expertos y AC1 del sistema entrenado con Base de entrenamiento Final

DISCUSIÓN

Considerando que no teníamos ningún argumento para preferir el juicio de uno de los expertos en particular frente a los otros, primero se entrenó el sistema con cada clasificación por separado para ver el desempeño del clasificador automático según los tres expertos. En la Figura 1 y Figura 2 se observa que el Experto 1 tiene el mayor AC1 para 12 de las 23 categorías, mientras que los Expertos 2 y 3 tienen el mayor AC1 solamente para 3 y 7 categorías respectivamente (y hay una categoría donde los Expertos 2 y 3 tienen el mismo AC1). Por otra parte, se calculó el acuerdo entre los expertos. En los Contenidos, el Experto 1 con el Experto 2 están tan de acuerdo como el Experto 2 con el 3, y en las prácticas, se observa una gran variación en los valores encontrados, habiendo incluso varios valores negativos y otros que están bajo 0.4. Lo anterior se debe a la dificultad que conlleva definir exactamente qué es y cómo detectar una práctica en una clase, y por ende, los expertos inevitablemente interpretan las prácticas de acuerdo a su experiencia personal y no de forma completamente objetiva.

Para tratar de lograr un desempeño más estable e independiente de quién sea el experto que clasifica las clases, se decidió entrenar el sistema con la "Clasificación Experta Consolidada". A partir de la Figura 1 se observa que el entrenamiento de la clasificación de contenidos con la Clasificación Consolidada de los tres expertos tiene un valor promedio de AC1promedio similar al promedio del AC1 obtenido con el entrenamiento de los tres Expertos por separado, y en cuanto a las prácticas, la Figura 2muestra que el valor del AC1 del entrenamiento con la Clasificación Consolidada de los tres expertos es más bajo que el obtenido al entrenar el sistema con los Expertos 1, 2 y 3 por separado, pero al utilizar la Clasificación Experta Consolidada, estamos entrenando el sistema con datos más robustos y objetivos, dada la subjetividad en la detección de prácticas docentes en las clases.

Por consiguiente, dado que los tres expertos clasificaron solamente una parte de las clases disponibles para entrenar el sistema, se decidió seguir entrenando el sistema con todos los textos

disponibles, aunque no estuvieran clasificados por tres expertos cada uno. Lo anterior aumentó los datos de entrenamiento de 214 a 624, y los de testeo de 95 a 242. Así, de acuerdo a los resultados presentados en la Figura 3, el clasificador automático mejoró significativamente su desempeño, sobre todo en aquellas categorías más difíciles de detectar.

En particular, considerando que, hasta el momento el mejor desempeño se había obtenido al entrenar el sistema con las clasificaciones del Experto 1, de la Figura 3 se observa que el entrenamiento final tiene mejor AC1 que el Experto 1 en 11 categorías; y mejor AC1 en 12 categorías respecto a la Clasificación Consolidada de los tres Expertos. Lo anterior es importante porque el sistema en su etapa de producción se deberá seguir reentrenando constantemente a partir de la nueva información que los docentes entregan al utilizar la aplicación grabando y clasificando más clases, y es inviable que un Experto se mantenga de manera permanente encargado de clasificar todas las nuevas clases que se espera ingresar a futuro. Por ende, saber que el clasificador automático obtiene mejores resultados al entrenarlo con clasificaciones realizadas por múltiples personas utilizando la aplicación, hace que este proyecto sea sostenible en el tiempo.

Por otra parte, a partir de la Figura 4 se observa que el sistema entrenado puede clasificar automáticamente los textos y lograr una concordancia con las clasificaciones de los expertos, igual o mejor a la concordancia que logran los expertos entre sí. Dado que la base final de entrenamiento es creada a partir del promedio de los acuerdos entre expertos clasificadores, no es realista pensar que un autómata pueda estar más de acuerdo con los expertos que los expertos entre ellos, y por ende un buen resultado es aquel que se asemeja al acuerdo que logran los expertos entre sí, porque indica que el sistema se puede comportar como un experto más en la clasificación.

Al observar los resultados del entrenamiento final de la Figura 3, de las 23 categorías a clasificar, 5 tienen un valor de AC1 sobre 0.9, 10 están sobre 0.8 y 16 de las 23 categorías están sobre 0.4. Al separar por categorías de Contenidos y Prácticas, tenemos que 4 de los 6 Contenidos están sobre 0.9 y las 6 categorías de contenidos están sobre 0.55. Por último, 8 de las 17 prácticas superan 0.6 y 10 superan 0.4. Estos resultados muestran que sí es posible clasificar tanto contenidos como prácticas de forma automática a partir de un texto. Los resultados son mejores en la identificación de Contenidos que Prácticas. Esto se explica por el hecho de que, a diferencia de un contenido, la presencia o no de una práctica docente puede manifestarse desde muchas dimensiones (por ejemplo, tono de voz, gestos, etc.), siendo el texto hablado solo una de ellas, es probable que se necesite mucho más entrenamiento del sistema para reconocer prácticas docentes con los mismos niveles de AC1 (acuerdo con expertos) que los contenidos, usando solo texto. Asimismo, dada la mejoría en el desempeño al casi triplicar la cantidad de datos de entrenamiento, esperamos que a mayor cantidad de registros de audios, mejor será el desempeño del clasificador automático.

Notemos que existen otros indicadores que pueden ser usados para estudiar la coincidencia o no de los resultados con los juicios expertos. Uno de ellos es el de Kolmogorov-Smirnov (KS) (Arnold & Emerson, 2011). Si bien este es un indicador muy usado en la literatura, en el caso en que las categorías que se reportan presentes (o ausentes) aparecen sólo en porcentajes pequeños de la muestra, el AC1 se comporta mejor. En nuestro caso, para las prácticas docentes, 7 categorías fueron reportadas presentes en promedio en el 46% de los datos, pero otras 10 de ellas el porcentaje promedio de presencia es sólo del 17%, lo que explica entonces nuestra selección del AC1 como indicador. Si bien no lo reportamos en detalle en este manuscrito, cabe señalar que los menores valores de AC1 en nuestros datos tienen asociados en cambio los mejores valores de KS.

Para futuras investigaciones, y por completitud, nos gustaría mejorar y perfeccionar el transcriptor automático para aumentar los aciertos, aunque esto no es esencial dado que nuestros resultados actuales son menores a lo esperado debido a la inexactitud en transcribir las palabras que no tienen relación directa con los contenidos y prácticas buscados, tales como pronombres, artículos, etc. Con esto en mente se está trabajando en pasar el sistema a Sphinx 4, con lo que esperamos un mejor

desempeño. Así también, en el trascurso del desarrollo del proyecto se ha logrado aumentar la cantidad de audios transcritos y clasificados, con lo que el entrenamiento del sistema, tanto enel módulo de transcripción automática como en la clasificación de contenidos y prácticas, ha mejorado sus resultados a medida que aumentan los audios, como efecto de disponer de más datos para el entrenamiento de los algoritmos de clasificación. Por ende, a medida que los algoritmos se entrenen con más clases, el desempeño del clasificador automático de prácticas y contenidos también debería mejorar.

CONCLUSIONES

Se ha podido comprobar que el sistema de auto soporte es capaz de entregar información respecto a contenidos y prácticas docentes en clases de matemáticas que puede ser similar a la proporcionada por expertos respecto a los mismo ítems. Se ha visto además que a medida que el sistema es alimentado con más clases, aprende y la calidad de la retroalimentación continúa mejorando lo que permite esperar que con la masificación de su uso se incremente aún más la calidad de la información que entrega.

AGRADECIMIENTOS

Este trabajo contó con el apoyo de Rita Singh y Bhiksha Raj del Language Technologies Institute at CMU. Agradecemos también a Anita Tobar por su colaboración en la incorporación de fonemas para el sistema de transcripción automática y a Patricio Calfucura por su ayuda en el proceso de toma de datos. El desarrollo de este proyecto ha sido posible gracias al financiamiento del Programa Fondef de Conicyt, a través del Proyecto D11I1009, y al proyecto FB0003 del Programa de Investigación Asociativa de Conicyt.

REFERENCIAS

Araya, R.; F. Plana; Dartnell, P.; Soto-Andrade, J.; G. Luci; E. Salinas; M. Araya (2012) "Estimation of teacher practices based on text transcripts of teacher speech using a support vector machine algorithm". *British Journal of Educational Technology*, 43 (6), 837–846.

Arnold, Taylor B.; Emerson, John W. (2011).Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. *The R Journal*, 3 (2), 34–39.

Basic Concepts of Speech.(n.d.). Obtenido del sitio web de Carnegie Mellon University Sphinx, http://cmusphinx.sourceforge.net/wiki/tutorialconcepts

Blood E, Spratt KF (2007). Disagreement on agreement: two alternative agreement coefficients. *SAS Global Forum 2007*.

Bouchet-Valat, Milan (08 de Agosto de 2014). Documentation for R package SnowballC. [Sitio web]. Obtenido de https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf

Breiman, Leo (2001). "Random Forests".*Machine Learning*45 (1), pp 5-32.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology*, 61(1), 29–48.

Koedinger, K.; Booth, J.; Klahr, D. (22 de Noviembre de 2013).*Instructional Complexity and the Science to Constrain It*. Science Magazine, 342, 935-937

National Board Resource Center (2010).A quality teacher in every classroom. Sanford, California: Stanford University.

National Education Association (1946).Research bulletin. December 1946, pp. 146–148.

Stevens, R. (1912). *The question as a measure of efficiency in instruction*. New York: Bureau of Publications. Teachers College, Columbia University. pp. 11, 15–17.

Wild, Fridolin (07 de Mayo de 2015). Documentation for R package LSA.[Sitio web]. Obtenido de https://cran.r-project.org/web/packages/lsa/lsa.pdf