Jabirú: hacia una herramienta para corrección asistida de errores

Julio Esteban Rojas Castro
Instituto Tecnológico de Costa Rica
Alajuela, Costa Rica
Email: j.rojas.c123@gmail.com

José Miguel Hernández Víquez
Instituto Tecnológico de Costa Rica
Alajuela, Costa Rica
Email: jmiguelh.19@gmail.com

Jose Paulo Yock
Instituto Tecnológico de Costa Rica
Alajuela, Costa Rica
Email: jyock1996@gmail.com

Aurelio Sanabria Rodríguez
Instituto Tecnológico de Costa Rica
Alajuela, Costa Rica
Email: ausanabria@itcr.ac.cr

***Resumen*—Este artículo tiene el propósito de presentar resultados preliminares de una investigación en proceso hecha por estudiantes del Laboratorio Experimental. El equipo ha estado trabajando en un nuevo módulo para *LibreScan*, un software para digitalización de texto. Con objetivo de incorporar reconocimiento de errores producidos durante el proceso de digitalización, específicamente, durante la etapa de reconocimiento de caracteres(OCR). El resultado de este trabajo es un archivo capaz de ser leído por un software lector de pantalla para ser escuchado por estudiantes con discapacidad visual. El artículo categoriza algunos de los errores más comunes que se producen durante el proceso de OCR y compara diferentes maneras para llevar a cabo el proceso de digitalización de un texto, el cual comprende desde capturar la imagen hasta generar el documento de salida.**

I. Introducción

En el Centro Académico de Alajuela del Instituto Tecnológico de Costa Rica, se crea en el 2013 el Laboratorio Experimental (LabExp), que surge de la necesidad de generar espacios para involucrar a los estudiantes en actividades académicas no contempladas en la malla curricular. Como parte de sus actividades se realizan proyectos de investigación entre los cuáles se encuentra Información Libre y Tecnología (ILT), que busca promover el derecho a la información y el acceso gratuito a la educación mediante el desarrollo de herramientas de Software Libre que garanticen el acceso a cualquier persona que lo necesite.

Como principal resultado del trabajo en este proyecto se encuentra el software *LibreScan*[1]. Este software facilita el acceso de estudiantes con deficiencias visuales. A material educativo de apoyo tal como libros, fotocopias u otro tipo de textos que no tienen una versión digital. Esta herramienta se desarrolla incorporando metodologías de trabajo típicas de proyectos de Software Libre. Permite digitalizar libros sobre la plataforma de hardware *DIY BookScanner*[2] (ver imagen 1), integrando el proceso de digitalización de libros, desde la captura de imágenes hasta la creación de un archivo de salida. Todo este proceso se divide en una serie de etapas, entre ellas la captura de imágenes de cada una de las páginas del documento, luego se filtran las imágenes y se limpian

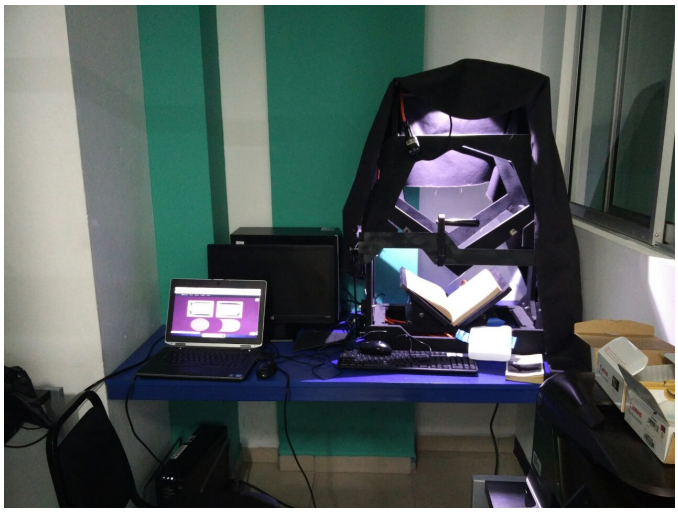

Figura 1. Instalación de prueba del LibreScan junto al DIY BookScanner en la biblioteca de la Sede Interuniversitaria de Alajuela

utilizando *Scantailor*[3], para finalmente extraer el texto de las imágenes mediante el uso de *Tesseract*[4] e integrar todo en un solo archivo de salida con el documento deseado (ver imagen 2).

Actualmente, se trabaja en extender *LibreScan* creando un módulo de corrección guiada de errores –denominado *Jabirú*– (ver imagen 3). El módulo permite editar y corregir errores que se presentan durante la etapa de Reconocimiento Óptico de Caracteres (OCR, por sus siglas en inglés) de forma que le permita al usuario utilizar un lector de textos, en un documento sin errores. Así, una persona con deficiencias visuales puede escuchar un PDF de forma fluida. El proyecto puede causar un impacto significativo en la Sede Interuniversitaria de Alajuela, donde se han presentado casos de estudiantes con diferentes tipos de discapacidades visuales, que para acceder a documentos y fotocopias sin versión digital recurren a la colaboración voluntaria de la bibliotecóloga Lic. Anabelly Tinoco, la cual realiza todo el proceso de digitalización de forma manual. Buscamos que este instrumento brinde igualdad

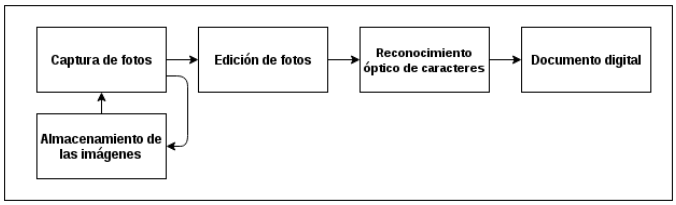

Figura 2. Etapas del proceso de digitalización incorporadas al flujo de trabajo de LibreScan

de condiciones. De la misma manera el desarrollo de esta herramienta nos permite crear una base sobre la cual investigar instrumentos y metodologías utilizadas para la corrección y edición de errores en texto reconocido a partir de imágenes, explorando nuevos algoritmos y técnicas que permitan agilizar este proceso.

II. Errores en el reconocimiento óptico de caracteres

OCR es un proceso en el cual se procesan imágenes e identifican secciones que correspondan con la representación de un carácter. Esta información se extrae a un archivo en un formato específico permitiendo modificarla mediante un software de edición de texto. Al someter una imagen al proceso de extracción de texto, suelen aparecer errores en el reconocimiento. Estos errores aumentan cuando vienen acompañados de una imagen que no posee la nitidez necesaria, ya sea por ser generadas por una cámara de baja calidad o por el estado del material físico en que se encuentra el texto, tal como un libro antiguo con sus hojas dañadas o que ha perdido la calidad de impresión.

Los errores más comunes durante este proceso de digitalización se pueden clasificar en distintas categorías: detección y segmentación de palabras, reconocimiento y segmentación de caracteres y errores por multifactor. El error de *detección y segmentación de palabras* se produce cuando no se identifica correctamente una palabra en el texto, ya sea por una mala definición de imagen, por estar combinados gráficos y textos, errores en detección de espacios, o secciones con espaciado y alineación diferentes al resto del texto. Otra causa es cuando se identifica una letra erróneamente como separada de su palabra, error muy frecuente en fuentes en estilo cursiva.[5], [6]

Los errores en *reconocimiento y segmentación de caracteres* son los producidos cuando el OCR confunde un caracter con algún otro. Estos se pueden subclasificar en inserción, eliminación, sustitución o transposición. La mayoría de errores que se presentan son de sustitución, cuando no se reconoce bien un caracter y se escriben en cambio algunos diferentes, por ejemplo: $n \longrightarrow ri$ o $m \longrightarrow iii$. Los errores de eliminación son los casos donde no se reconocen caracteres y los de inserción se ocasionan cuando se insertan caracteres de más que no existen en el documento original, aunque estos errores no son tan comunes.

Por último, los *errores por multifactor*, son los producidos en diferentes etapas de la digitalización desde la toma de la imagen al OCR. Por ejemplo, por una baja calidad del documento físico, el mal estado del equipo encargado de hacer el proceso, o la distribución compleja del texto en la página. [5], [6]

Como respuesta a estos errores, se han propuesto varias alternativas, como implementar un método con lexemas y n-gramas. Los n-gramas son oraciones de n palabras que tienen sentido semántico. Este método utiliza el algoritmo de Levenshtein. Para encontrar el n-grama que posee la menor cantidad de diferencias con respecto a la agrupación original. Este método brinda de brindar un mecanismo de aprendizaje para el sistema, pues se actualizará el conjunto de n-gramas con la agrupación que el usuario considere correcta durante la corrección asistida del texto. [7] La ventaja de este método es que al utilizar los n-gramas la revisión es semántica. Es decir, se hará una sustitución por lo que se va a sustituir por la palabra que tenga más sentido en la frase. Por otra parte, tiene una desventaja que puede llegar a ser importante, se utiliza una biblioteca externa para el proceso, ya que recompilar una gran cantidad de n-gramas tiene un costo alto en procesamiento y memoria. Por último, esta biblioteca necesita una alta capacidad de almacenamiento o conexión a Internet para ser utilizado como un servicio[6]. Al terminar el desarrollo, se evaluará la efectividad de esta opción utilizando como métrica el *Alphanumeric Word Rate (AWR)*, el cual esta definido como la relación entre las palabras válidas y el total de palabras en el documento[8], para comparar el proceso manual con el asistido por nuestra herramienta, utilizando como referencia documentos digitalizados en la biblioteca institucional que han sido corregidos de forma manual.

Trabajamos solo en español, ya que es el idioma propio de nuestro contexto cultural y nos permitirá tener un impacto directo en nuestra universidad. Sin embargo, esperamos en un futuro incorporar el soporte para nuevos idiomas. Como posibles opciones, inglés y portugués.

III. Alternativas del proceso que realiza *LibreScan*

El proceso de digitalización consiste de una serie de etapas definidas, como: captura de imágenes, limpieza de imágenes, reconocimiento de texto (OCR) y generar un archivo de salida. *LibreScan* tiene la ventaja de proporcionar mecanismos para integrar todas las etapas esenciales del proceso. Con la implementación de Jabirú se pretende integrar tareas de corrección asistida de errores, las cuales están presentes en otras alternativas.

Jabirú esta siendo desarrollado como un módulo independiente, con la intención de integrarse con el software *LibreScan*[1] y complementar la herramienta. Pero además de *LibreScan*, existen alternativas para ejecutar el proceso completo de la digitalización.

Una de las alternativas al *LibreScan* que funciona sobre la misma plataforma – el DIY BookScanner– es *Pi Scan*[9], la cual permite capturar las páginas de los libros y generar un documento con estas fotos. Sin embargo, no ejecuta las otras partes del proceso, las cuales se deben ejecutar de manera manual. Una ventaja es que el programa se ejecuta

en una *Raspberry Pi* por lo que constituye una solución de bajo presupuesto que consume muy pocos recursos. Además, *Spreads*, es otro software que implementa el proceso completo de digitalización con la desventaja de no ser un proyecto activo.[10]

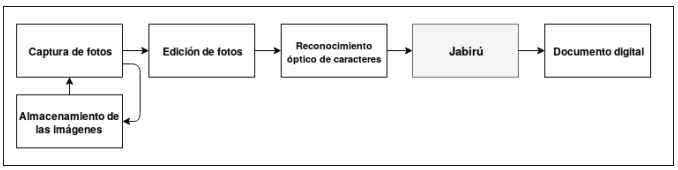

Figura 3. Flujo de trabajo de LibreScan junto al nuevo módulo en desarrollo

III-A. Desarrollo de Jabirú

Jabirú es un módulo del programa *LibreScan* que procesa los archivos con formato *hOCR* generados durante el reconocimiento óptico de los caracteres. Mediante una interfaz gráfica, se muestra al usuario la imagen de la página lado a lado con el texto correspondiente reconocido (Ver figura 4). Con el fin de realizar una corrección guiada utilizando n-gramas, se ayudará al usuario a realizar una corrección semántica del texto para obtener un resultado final más acertado. [11]

Actualmente los algoritmos OCR presentan deficiencias o errores en su resultado, esto se puede deber a la calidad de la imagen, la nitidez de los caracteres e incluso si el documento presenta manchas. Surge entonces la necesidad de aplicar un proceso adicional al documento, con el fin de corregir los posibles errores que se puedan generar durante el reconocimiento. Por lo cual Jabirú surge como una solución acertada para este problema.

Digitalizar un documento con *LibreScan* consiste de una serie de procesos entre los cuales están: tomar fotos, procesar las imágenes y crear el documento final, Jabirú sería el último proceso por aplicar antes de generar el documento final pues requiere una imagen de página con su respectivo archivo *hOCR* que contiene la información obtenida por el OCR. Evitando así un proceso manual del usuario para asegurar la correctitud del documento final generado.

Jabirú se encuentra en fase de desarrollo; aún están pendientes de implementar algunos requerimientos para su correcto funcionamiento entre los cuales se distinguen: mejorar la edición del *hOCR*, e implementar la detección de errores por medio de n-gramas. Por otra parte, se está dando una serie de cambios menores de estándar de código, con el fin de poder integrar a Jabirú con *LibreScan*.

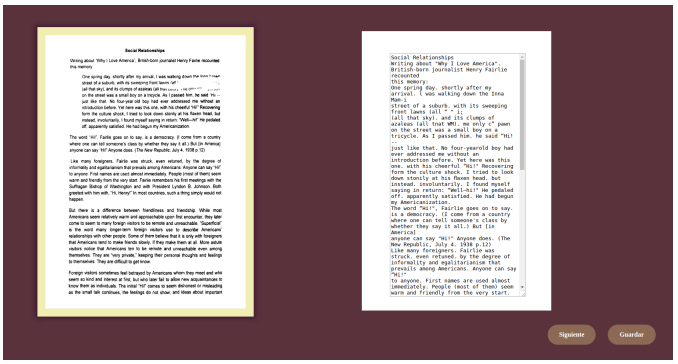

Figura 4. Captura de pantalla de Jabirú

IV. Conclusiones y trabajo futuro

El desarrollo de esta investigación nos permite crear una herramienta de software libre, fácil de utilizar y con la posibilidad de integrarse con *LibreScan*. Actualmente permite la edición manual de archivos *hOCR* que son resultado del proceso de digitalización de textos. La creación de este módulo nos acerca a contribuir a una problemática presente en nuestras universidades, equiparando las oportunidades en el acceso a la educación de personas con discapacidad visual.

Durante el período de ejecución nos topamos con dificultades que limitaron el avance, debido a la mala estimación de la complejidad de la tarea que se trató de resolver, se sobre estimó el tiempo disponible, la lentitud en la comunicación con la comunidad de desarrolladores de *LibreScan*. Todos profesionales graduados que trabajan a tiempo completo, y que horarios disponibles muy limitados. Sin embargo, se mostraron anuentes a reunirse con el equipo actual durante algunos sábados. Los estándares de programación propios del proyecto no se tomaron en cuenta durante el desarrollo, convirtiéndose esto en uno de los principales obstáculos para la completa integración de *Jabirú*. Finalmente el hecho de que no existen bibliotecas para la edición de archivos en formato *hOCR* ni documentación acerca de la estructura interna.

En el ámbito personal, esta investigación nos permitió desarrollar nuevas habilidades complementarias a los cursos de carrera como: trabajo en equipo, redacción de textos científicos, control de versiones de software, metodología de desarrollo de software libre, creación de videotutoriales y documentación técnica en general.

En la siguiente etapa de esta investigación, nos ponemos como objetivo el mejorar el editor de texto, diseñar junto a la bibliotecóloga una interfaz para la corrección asistida de errores, potenciado por el uso del algoritmo de de n-gramas en español para detección automatizada de errores y su corrección. Con lo que esperamos poder comparar la efectividad de nuestra herramienta contra la corrección manual de errores. Nuestro equipo espera poner a funcionar un escáner de este tipo, con el software completamente funcional, en la biblioteca institucional al finalizar el proyecto.

Agradecimientos

Agradecemos el apoyo brindado por el Laboratorio Experimental, la bibliotecóloga Anabelly Tinoco que nos apoya desinteresadamente probando la funcionalidad del software y al equipo de desarrollo original de *LibreScan*: Melvin Elizondo, Daniel Solís, Tony Kong y Diego Ugalde por su constante apoyo durante todo el proyecto.

REFERENCIAS

[1] LabExperimental-SIUA, "LibreScan." https://github.com/LabExperimental-SIUA/LibreScan, 2017.

[2] S. Jonathon Duerig, "Diy book scanner." https://www.diybookscanner.org/, 2015.

[3] N. Craun, "Scan tailor." http://scantailor.org/, 2014.

[4] R. Smith, "An overview of the tesseract ocr engine," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, pp. 629–633, IEEE, 2007.

[5] I. Kissos and N. Dershowitz, "OCR Error Correction Using Character Correction and Feature-Based Word Classification," *Proceedings - 12th IAPR International Workshop on Document Analysis Systems, DAS 2016*, pp. 198–203, 2016.

[6] J. Mei, A. Islam, Y. Wu, A. Moh'd, and E. E. Milios, "Statistical Learning for OCR Text Correction," 2016.

[7] D. a. Tong, Xiang; Evans, "A Statistical Approach to Automatic OCR Error Correction in Context," *Proceedings of the Fourth Workshop on Very Large Corpora (WVLC-4)*, 1996.

[8] P. Estrella and P. Paliza, "Ocr correction of documents generated during argentina's national reorganization process," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, (New York, NY, USA), pp. 119–123, ACM, 2014.

[9] Tenrec-Builders, "pi-scan." https://github.com/Tenrec-Builders/pi-scan, 2016.

[10] Spreads, "Spreads." https://github.com/Spreads/Spreads, 2016.

[11] D. Solís, M. Elizondo, D. Ugalde, T. Kong, J. Gutiérrez, and A. Sanabria, "LibreScan: Software Libre para la digitalización de documentos con escáneres de bajo costo," in *Workshop de software livre*, 2016.