CODICE: Un nuevo enfoque metodológico para el Procesamiento Inteligente de Documentos*

J.M. López-Carnicer[1], A. Martínez-Rojas[1], J.G. Enríquez[1], A. Jiménez-Ramírez[1], and Jesús M. Sánchez-Oliva[2]

[1] Departamento de Lenguajes y Sistemas Informáticos. Escuela Técnica Superior de Ingeniería Informática. Avenida Reina Mercedes, s/n. 41012, Sevilla.
{jlopez16,amrojas,jgenriquez,ajramirez}@us.es

[2] Servinform, S.A. Parque Industrial PISA, Calle Manufactura, 5, 41927 Mairena del Aljarafe, Sevilla, Spain. jmsanchezo@servinform.es

Abstract. La automatización de los procesos organizativos suele implicar el tratamiento de documentos en los que se emplean distintas técnicas de procesamiento. La Inteligencia Artificial (IA) y la Automatización Robótica de Procesos (RPA) se usan cada vez más para mejorar este procesamiento. Para ello, existen diferentes metodologías y técnicas para resolver problemas relacionados con la integración y uso cohesionado de dichas tecnologías en el campo del procesamiento inteligente de documentos (IDP). Este trabajo presenta el proyecto CODICE, que aborda la necesidad de crear una metodología para pipelines IDP, definiendo cómo incorporar la asistencia de IA y RPA, y una arquitectura que de soporte a ésta.

Keywords: Intelligent Document Processing · Methodology · Pipeline.

1 Introducción

El avance de la digitalización y la automatización de procesos ha hecho posible el procesamiento de grandes conjuntos de documentos. Esto se debe principalmente al desarrollo del (1) procesamiento inteligente de documentos (IDP, de sus siglas en inglés) [2], que trata de mejorar los resultados obtenidos en el procesamiento de documentos tradicional aplicando inteligencia artificial (IA), y (2) a la automatización robótica de procesos (RPA, de sus siglas en inglés) [8], que facilita la automatización de la ejecución del flujo de trabajo y permite asistir al IDP en la selección de documentos [3] o tareas de entrada y salida. En los últimos años han surgido diversas herramientas en la industria que aúnan soluciones IDP dentro de contextos RPA [7]. Sin embargo, son soluciones a medida que dificultan adaptar la tecnología para asumir las particularidades de cada problema IDP. Además, definen un control limitado sobre sus pipelines (i.e.,

* Esta investigación ha sido apoyada por el proyecto NICO (PID2019-105455GB-C31) del Ministerio de Ciencia, Innovación y Universidades, el proyecto CODICE (EXP 00130458/IDI-20210319 - P018-20/E09) del CDTI y el Programa de becas FPU, del Ministerio de Educación y Formación Profesional de España (FPU20/05984).

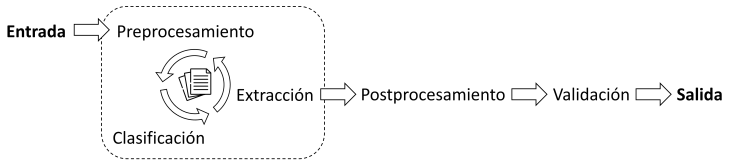

Fig. 1. Diagrama de ciclo de vida de soluciones IDP.

secuencias de trabajo ejecutadas para conseguir un objetivo de procesamiento de documentos). Por otro lado, como se extrae del análisis de [1], la mayoría de las propuestas de la comunidad científica se centran en algoritmos o técnicas aplicadas a tareas específicas de IDP. Como por ejemplo, la extracción de entidades o la segmentación [5]. En este sentido, no se abordan pipelines completos de IDP o, por otra parte, se focalizan en sectores concretos, como el bancario [6]. En este contexto, se presenta el proyecto de investigación CODICE, aún en desarrollo, liderado por la empresa Servinform S.A. junto con el grupo de investigación ES3[3]. CODICE se centra en dar solución a este problema mediante la definición de una metodología de desarrollo de pipelines IDP integrados con componentes configurables de IA y con RPA. El resto del documento se organiza de esta manera. La Sec. 2 describe la metodología propuesta. La Sec. 3 presenta la herramienta que soporta a esta. Finalmente, la Sec. 4 presenta los resultados preliminares de la propuesta y abre una discusión relacionada con estos.

2 Marco metodológico

La metodología propuesta para desarrollar un pipeline IDP comprende 7 fases diferenciadas incluyendo una fase de entrada y otra de salida (cf. Fig. 1): (1) **Entrada** de documentos en el proceso. Para ello, se utilizan robots RPA que obtienen la información a procesar y la envían al servicio IDP, o la almacenan en un contenedor de archivos accesible por este; (2) **Preprocesamiento**, que comprende diversas acciones de mejora para limpiar y transformar el documento de cara a optimizar su tratamiento, e.g., segmentación, conversiones de formato, obtención de metadatos, etc.; (3) **Clasificación** del documento. El clasificador actuará sobre el documento preprocesado y su elección dependerá de la aplicación, la preparación de los datos y el contexto. Pueden existir clasificadores basados en modelos de Machine Learning o Deep Learning, en reglas o plantillas, etc; (4) **Extracción** de información relevante. Se procesa el documento con técnicas como el reconocimiento óptico de caracteres (OCR), detección de entidades, expresiones regulares, etc.; (5) **Postprocesamiento**, que comprende la ejecución de tareas que necesita el documento ya procesado y estructurado, e.g., la ofuscación, anonimización, analítica, o toma de decisiones en base a su contenido; (6) **Validación** a partir de pruebas sobre los datos obtenidos. Esta puede ser: *intrínseca*, utilizando reglas de negocio o deterministas, *extrínseca*, mediante acceso a sistemas externos al proceso, o *híbrida*, incluyendo la intervención humana; (7) **Salida**, que es la respuesta a la solicitud de procesamiento

[3] Servinform S.A. (www.servinform.es) y ES3 (www.es3.us.es)

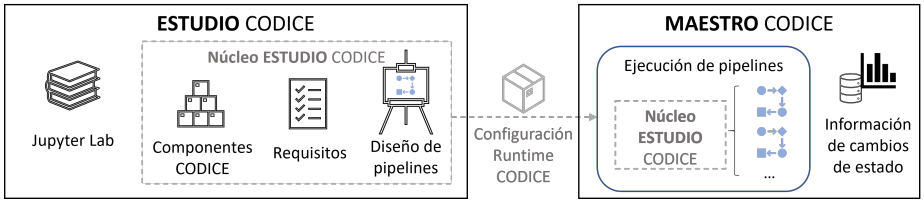

Fig. 2. Arquitectura de la plataforma.

de documentos. Para ello, el pipeline puede (7.1) responder al RPA que realizó la solicitud tras el procesamiento del documento, (7.2) almacenar la información resultante en un contenedor de archivos accesible por el RPA, o (7.3) actuar directamente sobre los sistemas de almacenamiento de datos como bases de datos, CRM, etc. La acción a realizar puede variar en función a la decisión obtenida como salida de algunos de los componentes de postprocesamiento.

Cabe destacar que podrán existir ciclos de preprocesamiento-clasificación-extracción (cf. Fig 1). Esto es debido a que en ocasiones hay que iterar varias veces sobre el ciclo para conseguir el resultado esperado, o que se llevan a cabo distintas clasificaciones, como *extraer el logo* y *clasificar por empresa*, y tras esto, *clasificar por tipo de documento* para derivar al departamento correspondiente. Hay que tener en cuenta que para la implementación de cada una de estas fases es necesaria una arquitectura de soporte que es lo que se define, a continuación, en la herramienta CODICE.

3 Herramienta que soporta el método

Como se muestra en la Fig. 2, CODICE propone una plataforma que soporta la metodología descrita en la Sec. 2, permitiendo además el control completo sobre los pipelines en ejecución. La arquitectura de esta plataforma se divide en dos módulos: (1) *Estudio*, que es un entorno de desarrollo web basado en JupyterLab[4], que permite el desarrollo de forma visual de pipelines IDP exportables. Éste utiliza componentes propios o de terceros y consta por defecto de una batería de componentes categorizados según la fase de la metodología que abordan, que permiten, entre otras funcionalidades, el tratamiento documental (e.g., OCR, detección de entidades, etc.) y el entrenamiento y explotación de modelos de IA. (2) *Maestro*, basado en el framework Django[5], que establece un control completo sobre los pipelines IDP, permitiendo la ejecución de sus múltiples instancias y la comunicación entre sus componentes. Se recopila, así, información sobre los cambios de estados de las instancias ejecutadas.

4 Resultados preliminares

Este trabajo propone una metodología, apoyada por la plataforma CODICE, para facilitar y simplificar el desarrollo de pipelines IDP, así como su integración con los procesos RPA, aumentando el grado de automatización de extremo a

[4] https://github.com/jupyterlab/jupyterlab
[5] https://www.djangoproject.com/

extremo de los mismos. Esta propuesta se valida con un caso de uso real dentro de la operativa de la empresa Servinform. Este consiste en la clasificación y extracción de datos de facturas de proveedores, para su posterior contabilización en el sistema de información SAP. Dicho caso fue probado usando 1.500 facturas, el equivalente a la operativa manual llevada a cabo en un mes por una sección acotada de la empresa. Se llevó a cabo la creación de un pipeline IDP utilizando la metodología CODICE propuesta, junto a su implementación en la plataforma CODICE, obteniendo un 99,36% de scoring (i.e., grado de acierto comparando el dato real al aportado por el pipeline CODICE) en pruebas de laboratorio y un 97,10% en su aplicación real. Este scoring permite alcanzar un nivel de errores menor a las operaciones llevadas a cabo por humanos [9] y obteniendo una mejora del tiempo medio del proceso similar al de soluciones que solo implementan RPA [4], significando una mejora prometedora en la calidad de los resultados. Debido a políticas de confidencialidad y seguridad de la empresa implicada, no se aportan más datos sobre la plataforma ni las pruebas realizadas. Se definen como líneas futuras un almacén de datos que soporte los estados de los documentos en cada una de las fases, además de validar la plataforma en casos de uso diferentes. A partir de diciembre del 2022, fecha de finalización del proyecto I+D que apoya esta iniciativa de valor, a CODICE se le dará continuidad, implantándolo como estrategia metodológica y plataforma de desarrollo y gestión de pipelines IDP en Servinform S.A. y su área de Consultoría e Innovación.

Referencias

1. Baviskar, D., Ahirrao, S., Kotecha, K.: A bibliometric survey on cognitive document processing. Library Philosophy and Practice pp. 1–31 (2020)
2. Esposito, F., Ferilli, S., Basile, T.M.A., Di Mauro, N.: Intelligent document processing. In: 8th ICDAR. pp. 1100–1104. IEEE (2005)
3. Ling, X., Gao, M., Wang, D.: Intelligent document processing based on rpa and machine learning. In: Chinese Automation Congress. pp. 1349 – 1353 (2020)
4. Martínez-Rojas, A., Sánchez-Oliva, J., López-Carnicer, J., Jiménez-Ramírez, A.: Airpa: An architecture to support the execution and maintenance of ai-powered rpa robots. In: BPM. pp. 38–48 (2021)
5. Oliveira, S.A., Seguin, B., Kaplan, F.: dhsegment: A generic deep-learning approach for document segmentation. In: 2018 16th ICFHR. pp. 7–12. IEEE (2018)
6. Oral, B., Emekligil, E., Arslan, S., Eryiğit, G.: Information extraction from text intensive and visually rich banking documents. Information Processing & Management **57**(6), 102361 (2020)
7. Ray, S., Villa, A., Tornbohm, C., Rashid, N., Alexander, M.: Magic quadrant for robotic process automation. Tech. rep., Gartner, Inc., 7 2020. (2020)
8. Ribeiro, J., Lima, R., Eckhardt, T., Paiva, S.: Robotic process automation and artificial intelligence in industry 4.0–a literature review. Procedia Computer Science **181**, 51–58 (2021)
9. Sutherland, C.: Framing a constitution for robotistan–racing with the machine for robotic automation. hfs research, ltd., october 2013 (2020)