DEPARTAMENTO DE ELECTRÓNICA
PROGRAMA DE POSGRADO

CICLO CONFERENCIAS Y SEMINARIOS

Curso Académico 2022-2023

23 de febrero de 2023

(15:00 a 16:00 h)

Sala 1 Dpto. Electrónica

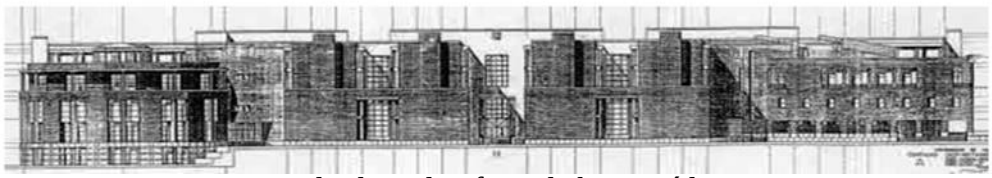
Key Information Extraction in Purchase Documents using Deep Learning and Rule-based Corrections

Dr. Roberto Arroyo Contera

Senior Data Scientist and IIT Lead. Nielsen IQ

Deep Learning (DL) is dominating the fields of Natural Language Processing (NLP) and Computer Vision (CV) in the recent times. However, DL commonly relies on the availability of large data annotations, so other alternative or complementary pattern-based techniques can help to improve results. In this paper, we build upon Key Information Extraction (KIE) in purchase documents using both DL and rule-based corrections. Our system initially trusts on Optical Character Recognition (OCR) and text understanding based on entity tagging to identify purchase facts of interest (e.g., product codes, descriptions, quantities, or prices). These facts are then linked to a same product group, which is recognized by means of line detection and some grouping heuristics. Once these DL approaches are processed, we contribute several mechanisms consisting of rule-based corrections for improving the baseline DL predictions. We prove the enhancements provided by these rule-based corrections over the baseline DL results in the presented experiments for purchase documents from public and NielsenIQ datasets.

Mención hacia la Excelencia (MEE -20110165)



Incluida en la oferta de bonocréditos.