TranSkriptorium (tS): procesado, transcripción e indexación de imágenes de texto

Enrique Vidal

Símile, ISSN 2171-6293, n. 49, 2021

La empresa TranSkriptorium (tS) se dedica a la comercialización de los servicios y productos desarrollados por el centro de investigación de la Universitat Politècnica de València en relación con el Reconocimiento de Formas y Tecnologías del Lenguaje Humano (PRHLT). Una de las aplicaciones más conocidas del reconocimiento de formas es el reconocimiento óptico de caracteres (OCR), si bien su uso no resulta muy práctico en el caso de documentos manuscritos históricos, ya que requiere una separación clara de los caracteres individuales. Para este tipo de documentación se han empleado tecnologías HTR, las cuales utilizan un enfoque holístico que evita analizar cada carácter por separado y pretende interpretar la palabra dentro de su contexto, acercándose más al proceso cognitivo humano. Profundizando más en esta línea, tS está desarrollando una nueva tecnología de reconocimiento llamada indexación probabilística (PrIx). Esta herramienta analiza los píxeles que componen la imagen de texto, estableciendo la probabilidad de que ese píxel forme parte de una secuencia de caracteres plausibles dentro de una palabra. El índice probabilístico de cada imagen puede contener alrededor de 4000 hipótesis de palabras posibles, lo cual supone que, partiendo de que cada imagen puede contener unas 200 palabras, la densidad media de indexación por cada palabra real es de unas 20 hipótesis. Para entender su enorme eficiencia y utilidad, debe compararse con otras tecnologías

como el OCR, el cual tan solo ofrece una única hipótesis por palabra escrita. La preservación de dichas hipótesis permite conservar las diversas interpretaciones posibles que puede emplear el usuario al realizar búsquedas en el contenido. Asimismo, PrIx resulta muy eficaz a la hora de generar posibles interpretaciones del texto incluso en documentos deteriorados o con tipos de escritura muy complejos y ambiguos. Aunque la idea inicial del desarrollo de PrIx consistía en facilitar la búsqueda, el autor propone otras áreas en las que podría resultar de utilidad tales como la transcripción automática o tareas derivadas del análisis del texto: segmentación y clasificación de grandes unidades archivísticas, extracción de entidades nombradas, etc. En definitiva, la indexación probabilística y la transcripción automática de manuscritos históricos ya es una herramienta actualmente al alcance de bibliotecas, archivos y otros centros de documentación. Asimismo, tS continúa trabajando en el desarrollo de otras muchas aplicaciones innovadoras de este tipo de tecnologías dentro del campo de las humanidades digitales.

https://cobdcv.es/simile/transkriptorium-ts-procesado-transcripcion-indexacion-imagenes-texto/

Resumen elaborado por Elena Esteban Jiménez