SíMILE
--- Butlletí del COBDCV // 2º Època // ISSN 2171-6293 ---

TranSkriptorium (tS): Procesado, Transcripción e Indexación de Imágenes de Texto

Paraules clau: Enrique Vidal, Indexación de Imágenes, Texto Manuscrito, Transcripción Automática

1. Introducción

La empresa tranSkriptorium IA S.L. (http://www.transkriptorium.com, en adelante abreviada como " *tS* ") se creó recientemente como una "spinOff" de la Universitat Politècnica de València (UPV). Su misión principal es la comercialización de productos y servicios derivados de desarrollos científicotécnicos realizados en el centro de investigación de Reconocimiento de Formas y Tecnologías del Lenguaje Humano, de la UPV (PRHLT, del inglés "Pattern Recognition and Human Language Technologies", https://www.prhlt.upv.es) durante los últimos 20 años. Concretamente, se trata de dar soporte profesional y comercial a diversas tecnologías de procesado, reconocimiento, transcripción e indexación de imágenes de documentos de texto, con especial énfasis en texto manuscrito histórico.

En la última década, estas tecnologías se han consolidado notablemente en el marco de sendos proyectos europeos y nacionales, en los que se han venido resolviendo muy satisfactoriamente retos de dificultad creciente. Cabe destacar tranScriptorium (2013-2015), liderado por PRHLT, que fue el primer proyecto financiado por la Comunidad Europea para el desarrollo de tecnologías de transcripción automática y asistida de imágenes de texto manuscrito. Así mismo, su sucesor, READ (2017-2019), también financiado con fondos Europeos, en el que consolidamos las tecnologías básicas de transcripción y de análisis de maquetación ("layout") e introdujimos con gran éxito una nueva tecnología de indexación y búsqueda, denominada Indexación Probabilística (PrIx). Esta nueva tecnología se consolidó en el proyecto Europeo (con cofinanciación española) Himanis (2015-2017), y más recientemente en dos proyectos genuinamente españoles, financiados por la Fundación BBVA, denominados Carabela (2017- 2019) y

HisClima (2019-2021), en el que se están consolidando nuevas técnicas de extracción de información en registros manuscritos estructurados, como por ejemplo, tablas.

Como resultado tangible de estos proyectos, se pueden consultar los demostradores de las distintas tecnologías que se encuentran en diversas páginas WEB. *1*

Más específicamente, en http://prhlt-carabela.prhlt.upv.es/PrixDemos se puede encontrar una lista de direcciones WEB de todos los demostradores de la tecnología PrIX desarrollados en los últimos 5 años. Entre estos se encuentran las interfaces de búsqueda avanzada de texto libre en cinco grandes colecciones históricas de imágenes de texto manuscrito: *Chancery* (Thrésor des Chartes, 82,000 páginas en francés y latín, siglos XIV-XV), *TSO* (Teatro del Siglo de Oro, 41,000 páginas en castellano, siglos XVI-XVII), *Bentham Papers* (90,000 páginas en inglés, siglos XVIII-XIX), *Carabela* (125,000 páginas en castellano, siglos XV-XIX) y *FCR* (Finnish Court Record Collection, más de 1,000,000 de páginas en sueco, siglos XVIII-XIX).

Todas estas tecnologías han sido licenciadas por la UPV a ***tS*** (participada entre otros por la propia UPV) que, desde 2021, las ofrece en forma de servicios y productos comerciales a diversas entidades públicas y privadas del sector de la documentación, y muy especialmente a todos los archivos históricos españoles y europeos.

2. Sistemas Inteligentes, Aprendizaje Automático y Texto Manuscrito

El reconocimiento de formas (RF) es una disciplina científica clásica, con la que están estrechamente vinculadas otras más recientes como el aprendizaje automático (AA) y, en general, los sistemas inteligentes (SI).

Una de las aplicaciones más antiguas de RF es el reconocimiento óptico de caracteres (en inglés, "optical character recognition", OCR). Las tecnologías tradicionales de RF para OCR pronto dieron lugar a sistemas de gran utilidad para la transcripción e indexación de documentos de texto impreso o mecanografiado. Estas tecnologías requieren una separación previa de los caracteres individuales que componen cada palabra del texto, lo que es generalmente fácil si el texto es impreso o mecanografiado, pero es prácticamente inviable cuando se trata de un texto manuscrito cursivo, en que la separación entre caracteres es casi inexistente o inconsistente, e incluso la separación entre palabras es a menudo confusa y/o caprichosa.

SíMILE
--- Butlletí del COBDCV // 2º Època // ISSN 2171-6293 ---

2.1. Transcripción Automática

El texto manuscrito también ha sido objeto de amplios estudios en RF y AA. Pero solo recientemente se están encontrando aproximaciones metodológicas (mucho más complejas que las técnicas de OCR) que empiezan a dar resultados aceptables para la transcripción automatizada de texto manuscrito en general. A estas nuevas tecnologías se las suele denominar por sus siglas en inglés, HTR ("handwritten text recognition"). Muchos de los desarrollos científico técnicos que han permitido este importante avance se han realizado en el marco de los proyectos del equipo de PRHLT que actualmente da soporte a la empresa ***tS.***

Hay que tener en cuenta que, para un lector humano, la lectura de texto escrito a mano generalmente requiere habilidades cognitivas complejas: para poder interpretar la información contenida en una imagen de texto manuscrito en forma de una secuencia de caracteres y de palabras, es generalmente necesario entender lo que se está leyendo. El contexto de cada carácter y palabra suele ser crucial para poder discernir cuál es la palabra escrita que se está viendo en la imagen, y así saber cuáles son los caracteres que la forman. Las tecnologías HTR se basan en métodos holísticos que, lejos de analizar cada carácter por separado (lo que por otra parte sería imposible en muchos casos), tratan de aproximar el complejo proceso cognitivo humano que acabamos de comentar. Es por esta razón por la que, a diferencia del OCR, un sistema HTR puede considerarse propiamente un sistema inteligente.

2.2. Indexación Probabilística y Búsqueda de Texto Libre en Imágenes

Los sistemas HTR actuales son capaces de transcribir con precisión útil imágenes de manuscritos sencillos (maquetación sencilla y regular, caligrafía simple y uniforme, correcta conservación del documento y buena calidad fotográfica). Pero, lamentablemente, estas condiciones simplificadoras raramente se dan en la mayoría de colecciones de manuscritos históricos, los cuales suelen mostrarse altamente esquivos para estos sistemas de HTR. Por eso, en PRHLT, y ahora en ***tS,*** también estamos desarrollado otra nueva tecnología, que a menudo se considera "disruptiva", denominada indexación probabilística (PrIx, del inglés, probabilistic indexing).

SíMILE
--- Butlletí del COBDCV // 2º Època // ISSN 2171-6293 ---

Para cada imagen de texto, se crea una especie de mapa de calor de palabras. Cada píxel de este mapa indica la mayor o menor probabilidad de que ese píxel forme parte de una o, generalmente, de muchas posibles palabras o secuencias de caracteres plausibles. Para una imagen típica, el índice probabilístico que representa este mapa contiene alrededor de 4.000 hipótesis de palabras o secuencias de caracteres posiblemente escritas en la imagen, con sus correspondientes probabilidades y posiciones en la imagen. Como en una imagen suele haber alrededor de 200 palabras realmente escritas, esto corresponde a una *densidad media de indexación* de unas 20 hipótesis de palabra por cada palabra real. Esto constituye una enorme diferencia con respecto a la transcripción automática que se obtiene mediante un sistema HTR (u OCR), que solo proporciona una única hipótesis por cada palabra transcrita (o sea unas 200 hipótesis de palabra por página, lo que equivale a una densidad de indexación de 1.0).

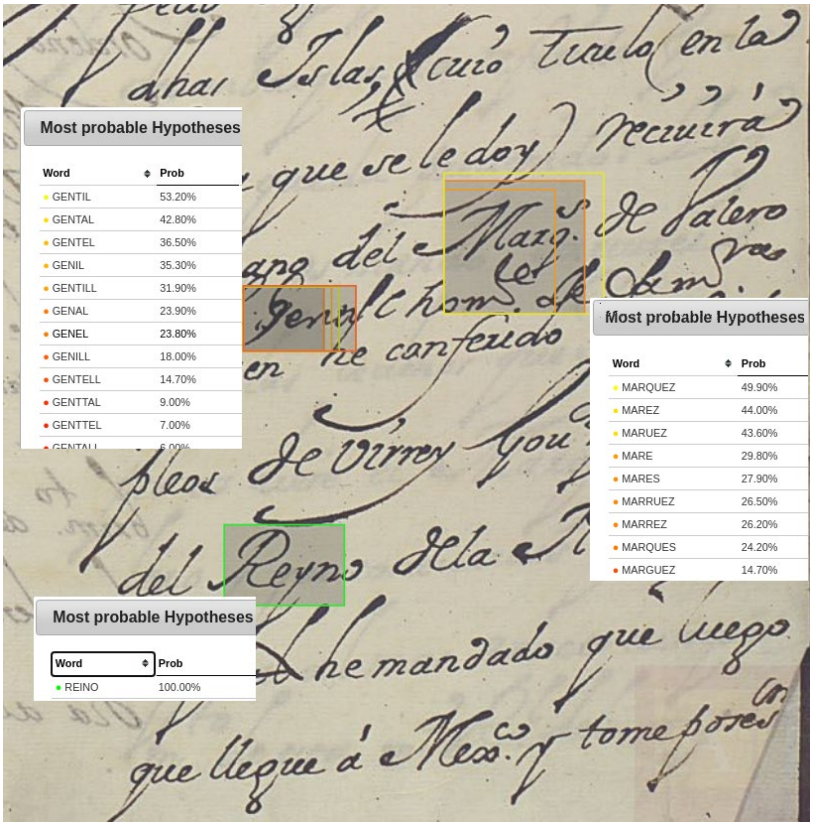

Figura 1: Ejemplos de densidad adaptativa de indexación probabilística (PrIx). Se muestran hipótesis de palabras en tres puntos de la imagen: GENTIL, con más de 30 hipótesis, de las que la primera es la correcta; REINO, con una única hpótesis que es correcta; MARQUES, con nueve hipótesis, de las que la octava es la correcta.

SíMILE
--- Butlletí del COBDCV // 2º Època // ISSN 2171-6293 ---

En la mayoría de colecciones de manuscritos históricos, alrededor del 30 % de las palabras (y a veces muchas más) que se obtienen mediante transcripción automática suelen ser erróneas. Si se usaran solo esas palabras para indexar la colección, los resultados de la mayoría de búsquedas de información serían decepcionantes. Por el contrario, en un índice probabilístico, si una palabra está claramente escrita, sin ambigüedades lingüísticas y la imagen es de buena calidad, el número de hipótesis que se indexan es muy bajo; quizás una sola hipótesis en muchos casos. Sin embargo, en partes deterioradas de documentos, con tipos de escritura complejos y/o ambigüedades lingüísticas (causadas, por ejemplo, por el uso de abreviaturas y/o arcaísmos), se pueden llegar a indexar hasta varios centenares de hipótesis por palabra real. Gracias a esta densidad adaptativa de indexación, es posible finalmente detectar información textual precisa, incluso en las peores condiciones de la documentación analizada.

En resumen, los índices probabilísticos tratan de preservar la incertidumbre inherente a la interpretación como texto de los trazos que se observan en las imágenes; de esta forma se evita perder posibles interpretaciones que quizás puedan parecer poco plausibles a primera vista, pero que pueden ser justamente las que interesen cuando se busque información en esas imágenes.

La figura adjunta muestra un ejemplo de hipótesis indexadas en tres puntos concretos de una imagen de texto manuscrito histórico.

3. Segmentación, Clasificación y Extracción de Información

Nuestra motivación inicial para el desarrollo de las técnicas de PrIx fue facilitar la *búsqueda* de información textual en grandes colecciones de imágenes de texto manuscrito. Como se ha comentado en la sección 1, este objetivo está plenamente alcanzado (y es actualmente ofrecido por ***tS*** como uno de sus productos y servicios estándar). Sin embargo, la información que contiene el PrIx de una imagen de texto permite otras muchas aplicaciones. Quizás la más directa es la transcripción básica (que se obtiene directamente como subproducto del PrIx). Pero además permite desarrollar técnicas innovadoras para abordar tareas de analítica de textos y *"big data"*, que hasta ahora eran impensables para imágenes no transcritas de texto manuscrito.

Una de estas tareas, de gran interés en archivos y bibliotecas, es la *segmentación* de grandes unidades archivísticas (por ejemplo, legajos) en subunides homogéneas (como, por ejemplo, expedientes o registros). Esta tarea, que requeriría considerable esfuerzo humano, es importante para poder realizar una correcta precatalogación y metadatación de la colección considerada.

Otro problema relacionado es la *clasificación* de unidades de archivo en función de sus contenidos textuales o, más concretamente, en función de sus "tipologías". El proceso de segmentación que acabamos de comentar puede ir acompañado de una clasificación de cada segmento o subunidad según su tipología; por ejemplo, un legajo que puede contener cientos de expedientes notariales, se podrá segmentar automáticamente y a cada expediente se le asignará la tipología más probable (tal como poder, carta de pago, venta, arrendamiento, testamento, etc.).

Finalmente, otra potencial aplicación interesante de los PrIx es la *extracción de información textual*; desde *"entidades nombradas"*, como nombres de personas, lugares, oficios, fechas, cantidades numéricas, etc., hasta *"palabras descriptivas"*, es decir, palabras "importantes" que aparecen prominentemente en una unidad de archivo y que distinguen a dicha unidad de otras de la misma colección.

4. Conclusión

La transcripción automática y la indexación probabilística de textos históricos es ya una realidad al alcance de archivos y bibliotecas. Además, de estas tareas básicas, que ***tS*** oferta ya como producto o servicio estándar, hay otras muchas aplicaciones innovadoras, basadas en Sistemas Inteligentes y técnicas de Aprendizaje Automático, que actualmente se encuentran en avanzado estado de desarrollo en el Centro PRHLT y que pueden ya ser ofertadas por ***tS*** mediante convenios específicos, dependientes de las tareas y colecciones consideradas.

Bibliografía

1. http://transcriptorium.eu/demots/htr, http://transcriptorium.eu/indexing-and-searching-based-on-keyword-spotting, http://prhlt-carabela.prhlt.upv.es/tld/ y https://www.prhlt.upv.es/wp/research-areas/htr-showcase (ver "Demonstrations").

SíMILE
--- Butlletí del COBDCV // 2º Època // ISSN 2171-6293 ---

Autor: Enrique Vidal, *es Profesor Emérito de la Universitat Politècnica de València y ha sido durante varias décadas co-director del grupo de investigación PRHLT, que actualmente es un centro propio de dicha Universidad. Es coautor de más de trescientas publicaciones científicas en las áreas de Reconocimiento de Formas, Aprendizaje Automático, Interacción Multimodal y aplicaciones en tratamiento automatizado del lenguaje, el habla y la escritura. En estas áreas y aplicaciones ha dirigido diversos proyectos de gran envergadura, incluyendo varios europeos y uno español del prestigioso programa Consolider Ingenio 2010. El Dr. Vidal es fellow de la International Association for Pattern Recognition (IAPR) y su índice h según Google Scholar es 51.*