Transcripción de periódicos históricos: aproximación CLARA-HD

Transcription in historical newspapers: the CLARA-HD approach

Antonio **Menta,** Eva **Sánchez-Salido y** Ana **García-Serrano**

ETSI Informática, C/ Juan del Rosal 16, UNED, 28040 Madrid, Spain

Resumen

Analizar periódicos de los siglos XVIII, XIX y principios del XX exige cierta calidad de las fuentes digitalizadas y la utilización de recursos específicos de dominio o de la lengua. Cualquier aproximación utilizando las tecnologías actuales, se encuentra con que la mayoría de los modelos PLN disponibles para la transcripción o el reconocimiento de entidades están entrenados con textos en "lenguajes actuales". Si además el reto consiste en extraer información de periódicos históricos en español, la complejidad aumenta, ya que la normalización del español es relativamente "moderna" y hay que intentar refinar los modelos de PLN o generar nuevos recursos. En esta presentación del corpus construido desde los textos disponibles en la Hemeroteca Digital de la BNE, Diario de Madrid (1788-1825), se mostrarán los pasos seguidos para su transcripción automática generando un modelo (99% de rendimiento) en el marco del proyecto CLARA-HD. Finalmente se incluyen unas conclusiones iniciales.

English translation. The analysis of historical newspapers from the 18th, 19th, and early 20th centuries requires a certain quality of digitized sources and the use of specific domain or language resources. Any approach using current technologies finds that most of the NLP models available for transcription or entity recognition are trained with texts in "current languages". If, in addition, the challenge consists of extracting information from historical newspapers in Spanish, the complexity increases since the normalization of Spanish is relatively "modern" and it is necessary to try to refine the NLP models or generate new resources. In this demonstration for the corpus built from the BNE Digital Hemeroteca, Diario de Madrid (1788-1825) the steps followed will be shown for its automatic transcription using a defined model (99% performance), within the framework of the CLARA-HD project. Finally, some initial conclusions are included.

Palabras Clave

Transcripción de textos, modelos del lenguaje, recursos lingüísticos.

1. Introducción

La utilización de técnicas de Procesamiento de Lenguaje Natural (PLN) en el tratamiento de documentos textuales, en concreto en el ámbito de las Humanidades Digitales (HD), se ha convertido en una práctica referente en muchos de los proyectos actuales [10]. En los últimos veinte años se han realizado multitud de procesos de digitalización para la conservación de colecciones culturales tanto a nivel local como nacional y europeo. Estos proyectos han generado millones de imágenes que necesitan ser tratadas para la transcripción del texto que contienen, ya sea de forma manual o mediante la aplicación de procesos de reconocimiento óptico de caracteres,

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

EMAIL: amenta@invi.uned.es (A. Menta-Garuz); evasan@lsi.uned.es (E. Sanchez-Salido); agarcia@lsi.uned.es (A. Garcia-Serrano)

ORCID: 0000-0002-3172-2829 (A. Menta-Garuz); 0000-0001-8665-3018 (E. Sanchez-Salido); 0000-0003-0975-7205 (A. Garcia-Serrano)

conocido como OCR (del inglés *Optical Character Recognition*).

La elaboración de corpus históricos está sujeta a múltiples factores, entre ellos su finalidad [9]. Por ejemplo, para el estudio de una lengua actual en general se pretende que el corpus sea proporcional, es decir, que la cantidad de palabras o de textos de cada muestra esté en proporción respecto a su distribución en el total de la población. Sin embargo, este requisito es difícil de conseguir en corpus históricos, ya que a menudo no se conservan suficientes documentos representativos de cada tipo, o incluso se desconocen las proporciones en que deberían aparecer. Por otra parte, la creación del corpus también depende del tipo de consulta que se desee realizar sobre los resultados que proporcione su análisis. En función de las posibilidades de consulta, los corpus son etiquetados mediante marcas declarativas que describen los elementos formales del texto (cursiva, tamaño de la fuente), elementos estructurales (capítulos, páginas) y elementos lingüísticos (entidades, cambios de registro).

La comunidad científica concienciada de la dificultad de tratar documentos históricos, en los últimos años está realizando un esfuerzo en mejorar las herramientas disponibles para su gestión, acceso y consulta [5]. Aquí es donde entran en juego las técnicas de PLN. Estas son capaces de extraer, procesar y relacionar la información que contienen los documentos para su posterior utilización y que sirvan de ayuda a los humanistas en sus reflexiones y análisis [6]. Si además es necesario trabajar con imágenes y textos [1] los sistemas de soporte a la investigación o de apoyo al trabajo del profesional se fundamentan en interfaces de interacción con la información mas complejos [2].

En esta presentación del corpus construido desde los textos disponibles en la Hemeroteca Digital de la BNE, Diario de Madrid (1788-1825)**¡Error! Marcador no definido.**, se justifica, en el apartado segundo, la necesidad de construir corpus de suficiente calidad para el análisis PLN previo al estudio de historiadores o público en general, se muestran los pasos seguidos para su transcripción en el apartado tercero y finalmente se incluyen algunos comentarios sobre este trabajo.

2. Necesidad de corpus de textos históricos de calidad

Las facilidades que ofrece la informática propician la confección de corpus que presentan el mismo texto en diversas modalidades de edición: facsímil (reproducción fotográfica del original), paleográfica (transcripción sin correcciones ni interpretaciones), normalizada (transcripción siguiendo la normativa ortográfica, léxica y sintáctica vigente), crítica (transcripción que pretende reconstruir el texto original) o interpretativa (transcripción que sigue los postulados de la edición paleográfica pero permite corregir ciertos errores para poder explicar el sentido del texto). Ejemplos son el corpus burckhardtsource.org y el proyecto CHARTA[2].

En el estudio del impacto de la tarea de reconocimiento de entidades nombradas (NER, por sus siglas en inglés) en el ámbito de las HD, en [11] se reflexiona sobre las posibilidades de utilizar NER y otros métodos de extracción de información en textos no estructurados y proponen ampliar el debate sobre la forma de utilizar las tecnologías del PLN a la comunidad humanística.

Dentro de las HD, el estudio de las ediciones de periódicos históricos entre el siglo XVIII y principios del siglo XX es un campo idóneo para aplicar estas técnicas debido a la presencia de todo tipo de entidades en ellos y a su evolución temporal a lo largo de los años para recuperar, almacenar y consultar la herencia cultural transmitida. Aun así, su uso directo presenta varios inconvenientes al utilizarlos en textos históricos. La mayoría de los modelos actuales son modelos estadísticos que necesitan un conjunto de datos etiquetados para ser entrenados en el contexto que se quieren utilizar, y estos conjuntos escasean o no son públicos en las HD. Esto repercute en otra dificultad añadida, que es la representación que deben tener los textos para ser utilizados por las técnicas del PLN.

Desde hace años se ha impuesto la utilización de modelos vectoriales de baja dimensión para representar los textos, conocidos como *word embeddings*. Para obtener estos modelos, en la mayoría de las ocasiones es necesario realizar un entrenamiento en una gran cantidad de textos del contexto en el que se quieren utilizar para aprender las relaciones entre las palabras y conceptos. Para obtener una mejor representación

[2] https://www.corpuscharta.es

final se suele realizar un pre-procesamiento de los textos para eliminar información irrelevante (como código HTML y algunos metadatos). Una vez limpio el texto, se utiliza como entrada para generar los *word embeddings*, ya sean estáticos o contextuales como los modelos basados en *Transformers* [12].

Últimamente, las redes neuronales basadas en modelos de lenguaje mejoran la detección de entidades, especialmente desde la publicación del modelo BERT [4] en 2018, o los modelos de lenguajes basados en *Transformers*. En [7] se realiza un estudio del impacto de la salida del OCR en el rendimiento de los modelos basados en BERT en un problema de clasificación de extractos de libros que van desde finales del siglo XVIII a finales del siglo XX. En sus conclusiones mencionan una degradación de los resultados y recomiendan realizar un ajuste fino de los modelos en esta tipología de documentos con anterioridad a realizar la clasificación para hacerlos más robustos a los errores ortográficos. Además, el vocabulario utilizado en siglos pasados dista enormemente del usado hoy en día y es un reto y una motivación para hacer hincapié en la utilización de los modelos de lenguaje basados en redes neuronales.

En definitiva, los intentos de análisis de documentos históricos mediante tecnologías de PLN actuales se encuentran con el problema de que la mayoría de los modelos disponibles están entrenados con textos en "lenguas modernas", y aumenta la complejidad al intentar extraer información de documentos históricos en español, ya que la normalización del español es relativamente "moderna" y hay que refinar los modelos de PLN o generar nuevos recursos.

3. Construcción del modelo de transcripción

La dificultad para aplicar la tecnología actual de PLN en las HD es el origen de los datos, porque la mayoría de las fuentes están almacenadas en imágenes de mala calidad con tipografías antiguas que necesitan de un OCR específico.

Transkribus[3] es una plataforma para la digitalización, el reconocimiento de texto, la transcripción y la búsqueda en documentos históricos. Es resultado de un proyecto europeo y de pago a partir de un cierto límite de uso. Con el registro se obtienen 500 créditos (unas 500 páginas). La herramienta está bien documentada[4] y cuenta con funcionalidades de acceso libre desde el navegador[5] o la aplicación.

Para la transcripción dispone de modelos basados en redes neuronales públicos y entrenados en distintos idiomas y grafías[6], lo que facilita encontrar uno que se aproxime al de los documentos a transcribir. De no ser así, la herramienta permite entrenar uno propio y automatizar la transcripción de nuestros documentos. De hecho, ya disponemos de un modelo entrenado a partir de transcripciones manuales en el proyecto CLARA-HD.

Para ello, se comienza creando una colección y cargando los ficheros que contienen los textos en ella (Figura 1).

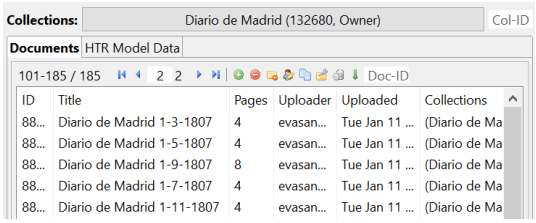

Figura 1. Carga de ficheros.

Para poder transcribir los documentos hay que realizar manualmente el reconocimiento de su estructura (o *layout*), diferenciando las regiones en las que se encuentra el texto (Figura 2). El reconocimiento en general no es perfecto, por lo que en ocasiones habrá que corregir errores o modificar manualmente.

[3] https://readcoop.eu/transkribus/
[4] https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/
[5] https://transkribus.eu/lite/
[6] https://readcoop.eu/transkribus/public-models/

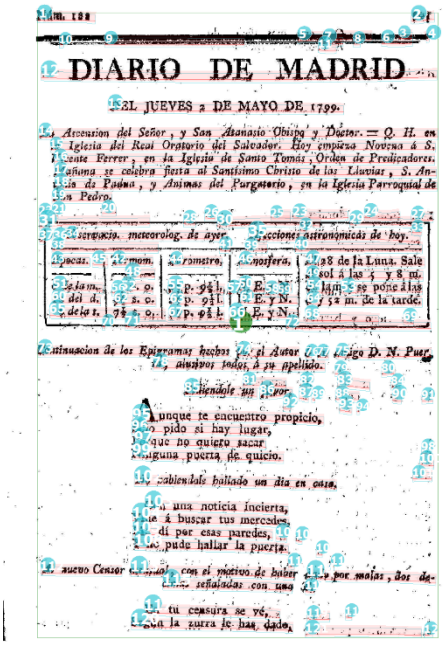

Figura2. Reconocimiento de la estructura.

Una reconocidas las regiones se transcribe el texto, línea a línea manualmente o con la ayuda de un modelo público seleccionado. Es posible que haya que editar la transcripción para corregir errores (Figura 3).

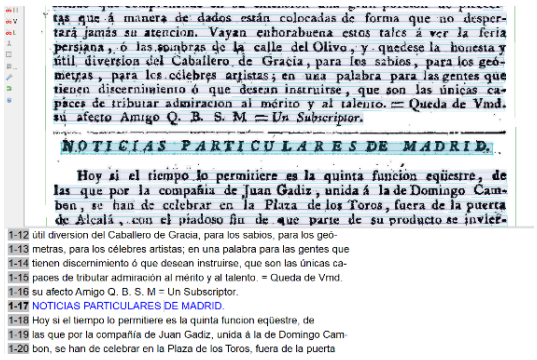

Figura 3. Transcripción manual.

Para automatizar este proceso se ha creado un modelo propio de transcripción a partir de un conjunto de entrenamiento junto con una guía de estilo, realizando los pasos mostrados anteriormente: (1) subida de documentos a la herramienta, (2) reconocimiento manual de la estructura de todas las páginas de los documentos, (3) transcripción de un cierto número de páginas manualmente o con la ayuda de un modelo público y (4) revisión manual final de las mismas, para entrenar nuestro modelo de transcripción.

[7] www.clara-nlp.uned.es

4. Comentarios finales

Se ha presentado cómo construir un corpus con la herramienta Transkribus, entrenando un nuevo modelo de transcripción capaz de reconocer caracteres no vistos por el modelo base, alcanzando una precisión en el reconocimiento de caracteres nuevos del 99%.

En este momento estamos trabajando con historiadores de la UNED interesados en el contenido del Diario de Madrid, para identificar tanto la terminología como los temas de interés para su investigación y evaluar cuánto es soportada por la tecnología PLN utilizada. Una vez identificados los tipos de entidades útiles para los historiadores, se seguirá con la extracción de las menciones de cada tipo, como las localizaciones, las profesiones o palabras complejas de entender.

5. Agradecimientos

Este trabajo parcialmente financiado por el proyecto coordinado CLARA-NLP[7] consta de tres subproyectos para dominios especializados en historia[8], biomedicina [3] y economía [8].

Finalmente, un agradecimiento especial para la participación en este subproyecto de los estudiantes en prácticas V. Sánchez-Sánchez, R. Garcia-Sánchez y A. Rodriguez-Francés.

Referencias

[1] J. Benavent, X. Benavent, E. de Ves, R. Granados, A. García-Serrano, Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches, M. Braschler, D.Harman, E. Pianta (Eds.) CLEF, CEUR Proc., V 1176. 2010.

[2] J. Calle-Gómez, A. García-Serrano, P. Martínez, Intentional processing as a key for rational behaviour through Natural Interaction, Interacting with Computers V 18 N 6, pp:1419-1446, 2006.

[3] L. Campillos-Llanos, A. Terroba, S. Zakhir, A. Valverde, A. Capllonch, Building a comparable corpus and a benchmark for Spanish medical text simplification, Procesamien. del Lenguaje Natural 69, 2022.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep

[8] (PID2020-116001RB-C31), (PID2020-116001RB-C32), (PID2020-116001RA-C33)

Bidirectional Transformers for Language Unders., arXiv preprint 1810.04805, 2018.
[5] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers, CLEF proc. 2020.
[6] A. Garcia-Serrano, A. Menta-Garuz, La inteligencia artificial en las Humanidades Digitales: dos experiencias con corpus digitales, Revista de Humanidades Digitales, v.7, pp: 19-39, 2022.
[7] M. Jiang, Y. Hu, G. Worthey, R. C. Dubnicek, T. Underwood, Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts, CHR 2021: Computational Humanities Research Conference, pp. 266–279, 2021.
[8] A. Moreno-Sandoval, A. Gisbert, H. Montoro, Fint-esp: a corpus of financial reports in Spanish, Multiperspectives in Analysis and Corpus Design, Editorial Comares, pp. 89-102, 2020.
[9] J. Torruella Casañas, Lingüística de corpus: Génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística, Peter Lang Ed., 2017.
[10] M. Toscano, A. Rabadán, S. Ros, E. González-Blanco, Digital humanities in Spain: Historical perspective and current scenario. Profesional de la Información, 29(6), 2020.
[11] S. van Hooland, M. de Wilde, R. Verborgh, T. Steiner, R. Van de Walle, Exploring entity recognition and disambiguation for cultural heritage collections, Digital Scholarship Humanities, V30, N2, 2015.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information Processing Systems 30, 2017.