Propuesta de Tesis

Maestría en Ciencias de Datos

- **Tema**: Corrección de transcripciones producidas por LUISA y OCRs
- **Tutor**: Ignacio Ramírez, Aiala Rosá
- **Formación específica recomendada**: Conocimientos de Procesamiento de Lenguaje Natural y Aprendizaje Automático.

Resumen de la propuesta

El estudiante deberá hacer una revisión exhaustiva de la bibliografía para familiarizarse y ponerse al día con los métodos existentes de corrección automática de texto. Deberá interiorizarse de los desarrollos previos llevados adelante en el marco del proyecto LUISA/Cruzar: enfoques aplicados, herramientas desarrolladas, recursos y lenguajes utilizados.

Descripción:

El Archivo Berruti (AB) consiste en más de 2 millones de imágenes de documentos generados por el aparato represivo antes, durante y después de la dictadura uruguaya (período 1972--1991). Las imágenes no son documentos escaneados comunes, sino escaneos de microfilms (rollos de negativos fotográficos diminutos) que a su vez eran fotografías de documentos físicos. Las imágenes tienen la particularidad de ser binarias, es decir, los pixeles son blanco o negro; no hay tonos de gris.

El proyecto Cruzar es un esfuerzo colectivo de docentes de UdelaR, así como de algunos actores externos, cuyo objetivo es la interpretación y análisis histórico del pasado reciente en Uruguay. El subproyecto LUISA/Cruzar engloba al conjunto de aplicaciones cuyo fin es transcribir los textos que se encuentran en las imágenes para luego ingresarlos a una base de datos. Es sobre esta base de datos que otros investigadores (por ejemplo, historiadores, sociólogos) realizan consultas con el fin de obtener información útil sobre el pasado reciente.

La transcripción de textos y, en especial, la transcripción automática de textos utilizando métodos computacionales (OCR -- Optical Character Recognition), genera una tasa relativamente alta de errores. Estos errores incluyen la sustitución de caracteres, y la aparición o falta de algunos de ellos. Claramente, este tipo de errores reducen la utilidad de los datos si no son corregidos. Es por eso que la corrección a posteriori de los textos transcritos es fundamental como paso previo a su ingreso a la base de datos.

Existen diferentes técnicas para la corrección automática de textos en general. Se trabaja con enfoques basados en diccionarios y distancias entre palabras [4], en modelos de lenguaje [1]. Tambien es usual la aplicación de técnicas de traducción automática [3]. Estos enfoques se han

aplicado en este proyecto, pero queda espacio para seguir investigando. Se trabajó con modelos de lenguaje de n-gramas, quedando pendiente el uso de modelos neuronales [2], por otro lado, se aplicaron técnicas de traducción automática estadística, pero no traducción basada en redes neuronales.

En este caso particular, se trata principalmente de corregir errores introducidos por sistemas OCR, que tienen particularidades propias que los diferencian del tipo de errores que una persona puede cometer al escribir. Es por esto que es necesario estudiar el problema de la corrección automática de textos para este caso específico. Una referencia interesante sobre posprocesamiento de OCR es la competencia ICDAR (Competition on Post-OCR Text Correction) [5].

Para tener una idea de lo anterior, un aspecto clave a comprender es que los OCR cometen errores de transcripción en base a imágenes, mientras que las personas lo hacen en base a la sonoridad. Por ejemplo: una persona puede escribir “Pedro se fue a casar” cuando en realidad el texto correcto debía ser “Pedro se fue a cazar”.

Cambiar “s” por “z” es un error relativamente común para un humano, pero no para un OCR, porque los símbolos “s” y “z” no se parecen. Sin embargo, mientras una persona difícilmente escribiría “Pedro se fue a cazer”, eso sería un error muy común para un OCR: los símbolos “a” y “e” suelen confundirse mucho (lo mismo con la “c” y la “o”, la “e” con la “c”, etc.).

Otro aspecto interesante de los OCR es que es posible disponer de la probabilidad de que un OCR confunda una letra cualquiera [4] x por otra y, $P(Y=y|X=y)$. Este tipo de información puede ser explotada a la hora de buscar la palabra correcta que debió transcribir el OCR en caso de un error.

Referencias

[1] Dan Jurafsky and James H. Martin. 2021. Speech and Language Processing (3rd ed. draft). Capítulo 3: N-gram Language Models. [https://web.stanford.edu/~jurafsky/slp3/, último acceso: noviembre 2021]
[2] Dan Jurafsky and James H. Martin. 2021. Speech and Language Processing (3rd ed. draft). Capítulo 7: Neural Networks and Neural Language Models. [https://web.stanford.edu/~jurafsky/slp3/, último acceso: noviembre 2021]
[3] Dan Jurafsky and James H. Martin. 2021. Speech and Language Processing (3rd ed. draft). Capítulo 10: Machine Translation. [https://web.stanford.edu/~jurafsky/slp3/, último acceso: noviembre 2021]
[4] Dan Jurafsky and James H. Martin. 2021. Speech and Language Processing (3rd ed. draft). Anexo B: Spelling Correction and the Noisy Channel. [https://web.stanford.edu/~jurafsky/slp3/, último acceso: noviembre 2021]

[5] Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, Jean-Philippe Moreux. 2019. ICDAR 2019 Competition on Post-OCR Text Correction. 15th International Conference on Document Analysis and Recognition, Sep 2019, Sydney, Australia. Pp.1588-1593. hal-02304334