Los documentos históricos son una parte importante de nuestra herencia cultural. Sin embargo, debido a la barrera idiomática inherente en el lenguaje humano y a las propiedades lingüísticas de estos documentos, su accesibilidad está principalmente restringida a los académicos. Por un lado, el lenguaje humano evoluciona con el paso del tiempo. Por otro lado, las convenciones ortográficas no se crearon hasta hace poco y, por tanto, la ortografía cambia según el período temporal y el autor. Por estas razones, el trabajo de los académicos es necesario para que los no expertos puedan obtener una comprensión básica de un documento determinado.

En esta tesis abordamos dos tareas relacionadas con el procesamiento de documentos históricos. La primera tarea es la modernización del lenguaje que, a fin de hacer que los documentos históricos estén más accesibles para los no expertos, tiene como objetivo reescribir un documento utilizando la versión moderna del idioma original del documento. La segunda tarea es la normalización ortográfica. Las propiedades lingüísticas de los documentos históricos mencionadas con anterioridad suponen un desafío adicional para la aplicación efectiva del procesado del lenguaje natural en estos documentos. Por lo tanto, esta tarea tiene como objetivo adaptar la ortografía de un documento a los estándares modernos a fin de lograr una consistencia ortográfica.

Ambas tareas las afrontamos desde una perspectiva de traducción automática, considerando el idioma original de un documento como el idioma fuente, y su homólogo moderno/normalizado como el idioma objetivo. Proponemos varios enfoques basados en la traducción automática estadística y neuronal, y llevamos a cabo una amplia experimentación que ratifica el potencial de nuestras contribuciones –en donde los enfoques estadísticos arrojan resultados iguales o mejores que los enfoques neuronales para la mayoría de los casos–. En el caso de la tarea de modernización del lenguaje, esta experimentación incluye una evaluación humana realizada con la ayuda de académicos y un estudio con usuarios que verifica que nuestras propuestas pueden ayudar a los no expertos a obtener una comprensión básica de un documento histórico sin la intervención de un académico.

Como ocurre con cualquier problema de traducción automática, nuestras aplicaciones no están libres de errores. Por lo tanto, para obtener modernizaciones/normalizaciones perfectas, un académico debe supervisar y corregir los errores. Este es un procedimiento común en la industria de la traducción. La metodología de traducción automática interactiva tiene como objetivo reducir el esfuerzo necesario para obtener traducciones de alta calidad uniendo al agente humano y al sistema de traducción en un proceso de corrección cooperativo. Sin embargo, la mayoría de los protocolos interactivos siguen una estrategia de izquierda a derecha. En esta tesis desarrollamos un nuevo protocolo interactivo que rompe con esta barrera de izquierda a derecha. Hemos evaluado este nuevo protocolo en un entorno de traducción automática, obteniendo grandes reducciones del esfuerzo humano. Finalmente, hemos aplicado el marco interactivo –nuestro nuevo protocolo junto con uno de los protocolos clásicos de izquierda a derecha– a la modernización del lenguaje y a la normalización ortográfica. Al igual que en traducción automática, el marco interactivo logra disminuir el esfuerzo requerido para corregir los resultados de un sistema automático.