Resumen

Actualmente existen grandes colecciones de documentos manuscritos en librerias de todo el mundo. La gran demanda de estos recursos ha llevado a la creación de librerías digitales para facilitar la preservación y el acceso electrónico a estos documentos. Sin embargo, la transcripción de las imágenes de estos documentos no está siempre disponible con tal de permitir la busqueda rápida y eficaz a los usuarios, o extraer patrones y datos estadísticos automáticamente. Esta tesis presenta una nueva aproximación para la transcripción asistida por ordenador (CAT) de documentos de texto manuscrito usando sistemas de reconocimiento de texto manuscrito (HTR).

El objetivo de las aproximaciones CAT es, completar de manera eficar una tarea de transcripción mediante la colaboración hombre-máquina, ya que el esfuerzo requerido para generar una transcripción manual es alto, y las transcriptiones obtenidas automáticamente por sistemas estado del arte aún no llegan a la precisión requerida. Esta tesis se centra en una aplicación especial de CAT, que es la transcripción de documentos manuscritos antiguos cuando el esfuerzo de usuario es limitado, y en consecuencia, el documento no puede ser revisado completamente. En esta aproximación, el objetivo es generar la mejor transcripción posible usando el esfuerzo de usuario disponible. Esta tesis ofrece una guia completa del proceso de CAT desde la extracción de características hasta la interacción de usuario. Primero, se propone una aproximación estadística para generalizar la transcriptión interactiva. Dado que su aplicación directa es inabordable, se han realizado una serie de asunciones para aplicarla en dos tareas distintas: la transcripción interactiva de documentos de textos manuscritos y la detección del formato del documentos de texto.

A continuación, se describe el proceso de digitalización y anotación de dos documentos manuscritos antiguos reales. Este proceso se llevo a cabo dada la escasez de recursos similires y la necesidad de datos anotados con tal de comprobar todas las herramientas y técnicas desarrolladas en esta tesis. Estos dos documentos fueron escogidos cuidadosamente con tal de representar las típicas dificultades que se encuentran al emplear técnicas HTR. Se presentan resultados de referencia en estos dos documentos obtenidos con un sistema estándar para servir de referencia. Finalmente, estos documentos se han hecho públicos y accesibles libremente a la comunidad. Hay que tener en cuenta que todas las técnicas y métodos desarrollados en esta tesis se han evaluado en estos dos documentos antiguos.

Seguidamente, se estudia y verifica de manera exhaustiva una aproximación CAT para HTR cuando el esfuerzo de usuario es limitado. El objetivo final de aplicar CAT se consigue mediante la unión de tres procesos separados. Dado el reconocimiento automático de un sistema HTR. El primer proceso consiste en localizar palabras (posiblemente) incorrectas y emplear el esfuerzo de usuario disponible en supervisarlas y corregirlas (si es necesario). Dado que la mayoría de las palabras no se van a supervisar ya que solo hay una cantidad limitada de esfuerzo de usuario, solo unas pocas serán seleccionadas para su supervisión. El sistema presenta al usuario un pequeño subconjunto de estas palabras elegi-

das por una estimación de su correctitud, o para ser más préciso, eligidas de acorde a su nivel de confianza. A continuación, el segundo proceso empieza una vez estas palabras de baja confianza han sido revisadas. Este proceso actualiza el reconocimiento del documento teniendo en cuenta las correcciones, lo cual mejora la calidad de las palabras que no han sido revisadas por el usuario. Finalmente, el último proceso adapta el sistema a partir de la última transcripción parcialmente supervisada (y posiblemente imperfecta) que se ha obtenido. En esta adaptación, el sistema escoge de manera inteligente que palabras correctas de la transcripción son usadas en la adaptación. Consecuentemente, el sistema adaptado reconocera mejor futuras transcripciones. Los experimentos de transcripción usando esta aproximación CAT que se han realizado muestran que esta aproximación es más eficaz cuando el esfuerzo de usuario aplicado es bajo.

La última contribución de esta tesis es un método para equilibrar la calidad de transcripción final y el esfuerzo de supervisión aplicado cuando se emplea la aproximación CAT previamente descrita. En otras palabras, este método permite al usuario controlar la cantidad de errores en las transcripciones obtenidas con una aproximación CAT. La motivación de este método es permitir a los usuarios decidir la calidad final deseada en los documentos, ya que una transcripción parcialmente erronea puede ser suficiente para entender el contenido, y el esfuerzo requerido para obtener esta transcripción puede ser significativamente menor que el de obtener una transcriptión manual completa. Consecuentemente, el sistema estima el esfuerzo de usuario mínimo requerido para alcanzar la cantidad de error definida por el usuario. La estimación del error se realiza calculando por separado el error causado por cada palabra reconocida, para después pedir al usuario que revisa aquellas donde hay más errores.

Además, se presenta un prototipo interactivo que integra la mayoria de las técnicas interactivas presentadas en esta tesis. Este prototipo se ha desarrollado para ser usado por expertos en paleografía, que no poseen ningún trasfondo en tecnologías HTR. Después de ser ajustado por experto en HTR, el prototipo permite a los transcriptores anotar un documento manualmente o utilizar la aproximación CAT presentada. Todos los procesos automáticos, como el reconocimiento, se ejecutan en segundo plano abstrayendo al transcriptor de los detalles internos del sistema. El prototipo fue probado por un experto transcriptor y demostró ser adecuado y eficiente para su finalidad. El prototipo está disponible libre y publicamente mediante una licencia GNU (GPL).