Resum

Actualment existeixen grans colleccions de documents manuscrits en llibreries de tot el mn. La gran demanda d'aquests recursos ha portat a la creaci de llibreries digitals per tal de facilitar la preservaci i access electrnic a aquests documents. No obstant, la transcripci de les images d'aquests documents no est sempre disponible per tal de permetre una cerca rpida i efica als usuaris, o d'extraure patrons i dades estatstiques automticament. Aquesta tesi presenta uno nova aproximaci per a la transcripci asistida per ordinador (CAT) de documents de text manuscrit emprant sistems de reconeiximent de text manuscrit (HTR).

L'objectiu de les aproximacions CAT es, completar de manera efica una tasca de transcripci mitjanant la colaboraci home-mquina, ja que l'esfor requerit per generar una transcripci manual es alt, i les transcripcions obtingudes automticamente per sistemes estat del art encara no arriben a la precisi requerida. Aquesta tesi es centra en una aplicaci especial de CAT, que es la transcripci de documents manuscrits antics quan l'esfor d'usuari es limitat, i en consequencia, el document no pot ser revisat completament. En aquesta aproximaci, l'objectiu es generar la millor transcripci posible emprant l'esfor d'usuari disponible. Aquesta tesi ofereix una guia completa del proces de CAT desde l'extracci de caracterstiques fins a l'interacci d'usuari.

Primer, es proposa una aproximaci estadstica per a generalitzar la transcripci interactiva. Donat que la seua aplicaci directa es inabordable, s'han realitzat una serie d'assumcions per tal d'aplicar-la en dos tasques diferents: la transcripci interactiva de documents de texts manuscrits i la detecci del format de documents de text.

A continuaci, es descriu el proces de digitalitzaci i anotaci de dos documents manuscrits antics reals. Aquest procs s'ha portat a terme donat el nombre escs de recursos similars i la necessitat de dades anotades per tal de comprobar totes les ferramentes i tcniques desarrollades en aquesta tesi. Aquests dos documents han estat escollits amb cura amb l'objectiu de representar les tpiques dificultats que es troben al utilitzar tcniques HTR. Es presenten resultat de referncia en aquests dos documents obtinguts amb un sistema estndar per tal de servir de referncia. Finalment, aquests documents s'han fet pblics i accessibles lliurement a la comunitat. Hi ha de tindres en compte que totes les tcniques i mtodes desarrollats en aquesta tes s'han evaluat en aquests dos documents antics.

Seguidament, s'estudia i verifica de manera exhaustiva una aproximaci CAT per HTR quan l'esfor d'usuari es limitat. L'objectiu final d'aplicar CAT s'aconsegueix mitjanant l'unio de tres processos separats. Donat el reconeiximent automtic d'un sistema HTR. El primer procs consisteix en localitzar paraules (possiblement) incorrectes i emprar l'esfor d'usuari disponible en supervisar-les i corregir-les (si es necessari). Donat que la majoria de les paraules no es van a supervisar ja que sols hi ha una quantitat limitada d'esfor d'usuari, sols unes poques sern selleccionades per una estimaci de la seua correctitut, o per a ser ms precs, sellecionades d'acord amb el seu nivell de confiana. A continuaci, el segon procs comena una vegada aquestes paraules de baixa confiana han estat revisades. Aquest procs actualitza el reconeiximent del document tenint en compte les cor-

recions, el qual millora la qualitat de les paraules que no han estat revisades per l'usuari. Finalment, l'ltim procs adapta el sistema a partir de l'ltima transcripci parcialment supervisada (i possiblement imperfecta) que s'ha obts. En aquesta adaptaci, el sistema escolleix de manera intelligent que paraules correctes de la transcripci son utilitzades en l'adaptaci. En consequencia, el sistema adaptat reconeixer millor les futures transcripcions. Els experiments de transcripci realitzats utilitzant aquesta aproximaci CAT mostren que aquesta aproximaci es ms efica quan l'esfor d'usuari aplicat es baix.

L'ltima contribuci d'aquesta tesi es un mtode per a equilibrar la qualitat de transcripci final i l'esfor de supervisi aplicat quan s'utilitza l'aproximaci CAT previament descrita. En altres paraules, aquest mtode permeteix al usuari controlar la quantitat d'errors en les transcripcions obtses amb una aproximaci CAT. La motivaci d'aquest mtode es permetre als usuaris decidir la qualitat final desitjada en els document, ja que una transcripci parcialment errnia pot ser sufficient per a entendre el contingut, i l'esfor requerit per obtindre aquesta transcripci pot ser significativament menor que el d'obtindre una transcripci manual completa. Com a resultat, el sistema estima l'esfor d'usuari mnim requerit per alcanar la quantitat d'error definit pel usuari. L'estimaci del error es realitza calculant per separat l'error causat per cada paraula reconeguda, per a desprs demanar al usuari que revis aquelles on hi ha ms errors.

A ms, es presenta un prototip interactiu que integra la majoria de les tcniques interactives presentades en aquesta tesi. Aquest prototip s'ha desarrollat per a ser utilitzat per experts paleogrfics, que no poseixen cap coneiximent de les tecnologies HTR. Desprs de ser ajustats per experts en HTR, el prototip permet als transcriptors anotar un document manualment o utilitzar l'aproximaci CAT presentada. Tots els procesos automtics, com el reconeiximent, s'executen en segn pla abstraent al transcriptor dels detalls interns del sistema. El prototip va ser probat per un expert transcriptor i desmostr ser adecuat i eficient per a la seua finalitat. El prototip est disponible lliure i publicament mitjanant una llicencia GNU (GPL).