Articoli di approfondimento - Dalla pagina stampata ai bit: nuovi strumenti per la digitalizzazione di massa

Nell'ambito di una ricerca finanziata dall'UE, è stata sviluppata una suite di riconoscimento testuale e di strumenti di elaborazione automatizzati in grado di migliorare la fedeltà e la ricercabilità di testi digitalizzati provenienti da archivi di musei e di biblioteche.

ECONOMIA DIGITALE

"Ai giorni nostri, i dati non digitali non sono visibili", afferma Hildelies Balk, capo dei progetti europei presso la Koninklijke Bibliotheek all'Aia, nei Paesi Bassi. "Per le biblioteche e gli archivi nazionali, oggigiorno questo problema è ancora più pronunciato rispetto al passato, in quanto la maggior parte delle persone si affaccia solo ora al mondo di Internet. Se qualcosa non è online, si presume che non sia disponibile. Oggi, quindi, le biblioteche, gli archivi e i musei nazionali hanno l'obbligo di rendere disponibili tutte le risorse a livello elettronico. Abbiamo bisogno di effettuare una scansione e una digitalizzazione di massa di libri, documenti e materiali stampati il prima possibile e nel modo più preciso possibile".

Il processo di digitalizzazione è relativamente semplice. Innanzitutto, si procede con la scansione di un documento per la creazione di un'immagine della pagina. Questo è il punto in cui il processo si è interrotto nei primissimi giorni della digitalizzazione. Oggigiorno, tuttavia, l'immagine scansionata viene poi elaborata, tipicamente mediante l'utilizzo di un software di "riconoscimento ottico dei caratteri" (OCR) in grado di estrarre il testo in un formato digitale. Una volta che il testo è stato

digitalizzato in base a queste modalità, l'intero documento è disponibile per l'indicizzazione e accessibile ai motori di ricerca.

La ricercabilità dei testi storici trasforma improvvisamente le collezioni in una potente risorsa culturale. In passato, per ricercare un documento specifico, era necessario recarsi presso un istituto specifico. Oggigiorno, grazie ad una rapida ricerca per parole chiave, ad esempio, è possibile recuperare migliaia di documenti e identificare un enorme volume di fonti importanti finora del tutto sconosciute.

È chiaro come funziona?

Ma i livelli di precisione della conversione da parole stampate a testo leggibile da una macchina sono tali da infonderci fiducia nei risultati di ricerca? "Volevamo migliorare o creare nuovi strumenti a valle della scansione effettiva che avrebbero ridotto gli errori creati dal sistema OCR", spiega il dott. Balk. "Questa digitalizzazione di massa sta generando un'immensa risorsa e, nel prossimo futuro, credo che si assisterà ad una proliferazione di applicazioni che sfruttano e addirittura monetarizzano la risorsa. Ma dobbiamo fidarci del fatto che la versione digitale di un testo storico sia una copia autentica dell'originale".

Negli ultimi quattro anni e mezzo, il dott. Balk ha coordinato il progetto Impact ("Improving access to text") nell'ambito del 7° PQ. Uno degli obiettivi principali dell'iniziativa è stato il miglioramento della precisione e dell'affidabilità del testo in uscita mediante lo sviluppo di una suite di strumenti software e di moduli di elaborazione che potrebbero essere utilizzati (talvolta in sequenza) con le immagini scansionate.

Prima di poter utilizzare qualsiasi OCR su un'immagine scansionata, è necessario dapprima "pulirla". L'Università di Salford nel Regno Unito, il Centro nazionale per la ricerca scientifica "Demokritos" ad Atene e lo specialista della tecnologia OCR ABBYY, con sede a Mosca, hanno lavorato su una serie di algoritmi di elaborazione di immagini che potrebbero analizzare e regolare l'immagine scansionata. I fautori del progetto One tool si sono occupati dell'allineamento di caratteri nella pagina e del raddrizzamento delle linee di testo diventate oblique, probabilmente perché vicine al dorso di un libro. Il progetto Another algorithm è in grado di rimuovere l'aspetto casuale dei pixel bianchi e neri (conosciuti con il nome di "rumore sale e pepe") spesso presenti nelle immagini scansionate.

Un bel carattere

I fautori dell'iniziativa hanno analizzato varie opzioni per il miglioramento dei risultati OCR. Un'importante area di collaborazione era basata su uno stretto partenariato con lo sviluppatore del software OCR e il venditore di ABBYY. "Abbiamo scelto di lavorare con questa azienda, in quanto il suo software OCR è ampiamente utilizzato

nelle biblioteche di tutta Europa per la digitalizzazione", afferma il dott. Balk. "ABBYY ha messo a nostra disposizione il suo kit di sviluppo software e ha lavorato a stretto contatto con noi per integrare la nostra ricerca nel suo software. È stato straordinario vedere come la nostra ricerca contribuisse al miglioramento di un prodotto già in uso".

"Non eravamo interessati al miglioramento dell'OCR in quanto tale", spiega il dott. Balk, "perché questo strumento è abbastanza ben sviluppato, ma la natura dei testi storici può talvolta rendere l'OCR meno preciso. Volevamo sviluppare strumenti che prendessero in considerazione questo contesto storico".

Ad esempio, i documenti storici presentano spesso layout complicati, con colonne multiple e capolettere. Solitamente, questi testi utilizzano anche vari tipi di caratteri non presenti nei materiali moderni. Il progetto Impact ha generato una serie (nota come corpus) di 50 000 trascrizioni digitali estratte da più di mezzo milione di pagine scansionate provenienti da varie biblioteche nazionali europee. Queste cosiddette "verità al suolo" che, come è stato confermato, rappresentano trascrizioni quasi perfette da utilizzare per "formare" il software OCR, allo scopo di renderlo in grado di riconoscere nuovi tipi di carattere o di gestire layout di pagine non usuali, oltre che per testare applicazioni per i loro risultati.

Il progetto ha anche prodotto dizionari storici che possono essere utilizzati dal software OCR per migliorarne le trascrizioni. Poiché l'OCR lavora mediante un'immagine scansionata, esso mette insieme i caratteri riconosciuti per formare "parole", verificandone quindi l'effettiva esistenza; in caso contrario, il software tenterà tipicamente di indovinare le parole ricercando i termini caratterizzati da un'ortografia più simile.

Tuttavia, la maggior parte dei software OCR utilizzerà dizionari moderni contenenti parole moderne. "I ricercatori vogliono leggere il contenuto effettivo dei documenti, con le ortografie originali", afferma il dott. Balk, "ma per ispezionare il documento non si vorranno ricercare 10, e a volte più di 50, ortografie diverse di una parola. A tal fine, abbiamo creato il progetto compiled dictionaries basato su parole arcane di nove lingue e ortografie che sono state mappate su sinonimi e ortografie moderni. In tal modo, l'OCR sarà in grado di trascrivere un documento parola per parola, ma sarà anche possibile utilizzare il dizionario per effettuare la conversione in ortografie moderne. Il dizionario aiuta a migliorare la precisione della digitalizzazione, insieme alla flessibilità e all'utilizzabilità".

Il tocco umano

Con la digitalizzazione di massa, è importante che questi strumenti lavorino automaticamente: le milioni di pagine che richiedono la digitalizzazione rendono impossibile la verifica della precisione delle trascrizioni. Tuttavia, il progetto ha

permesso di sviluppare nuove tecnologie che consentiranno agli utenti di verificare l'output dell'OCR in modo facile e veloce.

I linguisti computazionali presso l'Università di Monaco worked on an algorithm sono in grado di prevedere l'eventuale correttezza delle parole nella trascrizione OCR. L'algoritmo prende in considerazione il periodo di tempo e la lingua originale del documento e delle informazioni in base a modelli consolidati per l'ortografia e la linguistica storica. A partire da tale presupposto, il sistema è in grado di stabilire se le parole contenenti refusi, ad esempio, possono essere errori di OCR (che verranno evidenziati) oppure valide varianti ortografiche storiche.

Gli scienziati provenienti da IBM Israel Science and Technology hanno sviluppato un altro sistema in grado di combinare un nuovo approccio all'OCR. Questo "OCR adattivo" , chiamato CONCERT , aggiunge un sistema di correzione collaborativo intelligente che incoraggia l'impegno volontario teso al miglioramento dei livelli di precisione dell'output OCR automatizzato mediante la correzione degli errori umani.

"Impact ha prodotto una suite di strumenti che i partner stanno attualmente testando allo scopo di valutarne l'impatto sulla precisione e sulla fedeltà della trascrizione", osserva Clemens Neudecker, responsabile tecnico dei progetti europei presso la Koninklijke Bibliotheek. "Non vogliamo solo valutare la portata del miglioramento individuale dell'output, ma anche l'impatto di tali strumenti se combinati in una catena di elaborazione post-scansione. Stiamo anche garantendo l'interoperabilità di tutti questi strumenti, mediante la pubblicazione di una struttura architettonica tecnologica , che consentirà alle biblioteche di utilizzare gli strumenti e di elaborare documenti digitalizzati senza doversi preoccupare dei formati e delle conversioni dei file".

Il progetto terminerà alla fine del mese di giugno 2012. Tuttavia, attualmente, grazie all'iniziativa Impact Centre of Competence , l'esperienza collettiva dei partner e la loro dimestichezza nell'utilizzo e nello sviluppo di strumenti di digitalizzazione sono a disposizione della comunità della digitalizzazione di massa.

Il progetto IMPACT ha ricevuto un finanziamento di ricerca di 12,1 milioni di euro (del budget totale di 17,1 milioni di euro) tramite l'area tematica di ricerca sulle TIC del Settimo programma quadro (7° PQ) dell'UE.

Link utili:

- Sito web del progetto "Improving access to text"
- Scheda informativa del progetto IMPACT su CORDIS
- Impact Centre of Competence
- ICT Challenge 4: Digital libraries and content

- Europeana

Articoli correlati:

- Articoli di approfondimento - Digitalizzare il nostro patrimonio culturale

Progetti correlati

ARCHIVED

IMProving ACcess to Text

IMPACT

5 Aprile 2023

PROGETTO

Ultimo aggiornamento: 10 Luglio 2012

Permalink: https://cordis.europa.eu/article/id/88874-feature-stories-from-the-printed-page-to-bits-new-tools-for-mass-digitisation/it