Najważniejsze wiadomości - Od druku po bity: nowe narzędzia służące do masowej cyfryzacji

Finansowani ze środków UE naukowcy stworzyli zestaw zautomatyzowanych narzędzi do rozpoznawania pisma odręcznego oraz przetwarzania go, charakteryzujących się zwiększoną dokładnością i lepszym wyszukiwaniem tekstów zapisanych w formie cyfrowej, pochodzących z muzeów oraz archiwów bibliotek.

GOSPODARKA CYFROWA

"W dzisiejszych czasach materiały, które nie są dostępne w formie cyfrowej są po prostu niewidoczne", twierdzi Hildelies Balk, dyrektor do spraw projektów europejskich w bibliotece Koninklijke Bibliotheek w Hadze (Holandia). "W przypadku bibliotek i archiwów narodowych powyższy problem jest obecnie bardziej nasilony niż kiedykolwiek wcześniej, gdyż większość osób korzysta w dzisiejszych czasach praktycznie wyłącznie z Internetu. Jeśli jakieś materiały nie są dostępne w Internecie, to powyżsi użytkownicy zakładają, że nie są dostępne wcale. W związku z tym biblioteki narodowe, archiwa oraz muzea są obecnie zobligowane do udostępnienia posiadanych przez siebie zbiorów w formie elektronicznej. Musimy jak najszybciej i jak najdokładniej zeskanować i przetworzyć na formę cyfrową ogromne ilości książek dokumentów oraz materiałów drukowanych".

Proces cyfryzacji jest stosunkowo nieskomplikowany. Najpierw należy zeskanować dokument, tworząc obrazy odpowiadającej każdej z jego stron. Na wczesnym etapie rozwoju cyfryzacji na tym kończył się proces konwersji dokumentów drukowanych do formy elektronicznej. Obecnie jednak obrazy uzyskane ze skanera są przetwarzane, zwykle przy użyciu oprogramowania do "optycznego rozpoznawania znaków" ('optical character recognition' - OCR), które pozwala dokonać konwersji tekstu

drukowanego na formę cyfrową. Tekst w formie cyfrowej może być indeksowany i przeszukiwany przy użyciu silników wyszukiwania.

Możliwość przeszukiwania tekstów historycznych sprawia, że zbiory stają się nagle potężnym zasobem kulturowym. Dawniej znalezienie konkretnego dokumentu wymagało udania się do odpowiedniej instytucji. Obecnie wystarczy na przykład odpowiednio dobrać klika słów kluczowych, by uzyskać dostęp do tysięcy dokumentów; dzięki temu możliwe jest znalezienie ogromnej ilości przydatnych zasobów, bez konieczności posiadania uprzedniej wiedzy na dany temat.

Zrozumieć obraz

Czy jednak konwersja tekstu zapisanego na papierze do formatu zrozumiałego dla maszyn jest na tyle dokładna, by można było ufać wynikom uzyskiwanym podczas wyszukiwania? "Pragnęliśmy ulepszyć istniejące narzędzia lub stworzyć nowe rozwiązania stosowane po zeskanowaniu dokumentów, w celu zmniejszenia ilości błędów pojawiających się podczas korzystania z metod OCR", tłumaczy Dr Balk. "Dzięki masowej cyfryzacji powstają ogromne zasoby i spodziewam się, że w niedalekiej przyszłości pojawi się wiele rozwiązań korzystających z tych zasobów lub nawet pozwalających czerpać z nich korzyści finansowe. Musimy być jednak pewni, że cyfrowe wersje tekstów historycznych stanowią dokładne odzwierciedlenie oryginału".

W ciągu ostatnich czterech i pół lat Dr Balk był koordynatorem projektu "Ulepszanie dostępu do tekstów" ('Improving access to text' - Impact), realizowanego w ramach 7PR. Jednym z głównych celów powyższej inicjatywy było zwiększenie dokładności i bezbłędności konwersji tekstów drukowanych do formy elektronicznej poprzez stworzenie szeregu narzędziowych programów komputerowych oraz modułów przetwarzania, które można stosować (niekiedy w sposób sekwencyjny) pracując ze skanami.

Przed zastosowaniem jakichkolwiek technik OCR na zeskanowanym obrazie niezbędne jest jego uprzednie "oczyszczenie". Pracownicy Uniwersytetu w Salford (Wielka Brytania), Krajowego Centrum ds. Badań Naukowych w Atenach oraz moskiewskiej firmy ABBYY, specjalizującej się w technologii OCR, opracowywali szeroką gamę algorytmów do przetwarzania obrazów, pozwalających analizować i korygować skany. Narzędzie 'One tool' analizuje sposób wyrównania znaków na stronie, a następnie prostuje linijki tekstu, które z jakiegoś powodu zostały przekrzywione (np. zlokalizowane były w pobliżu grzbietu książki). Inny algorytm pozwala usunąć czarne i białe piksele zwane "szumem typu sól i pieprz", często pojawiające się w sposób losowy na zeskanowanych dokumentach.

Prawdopodobny znak

Uczestnicy projektu przeanalizowali ponadto różne rozwiązania pozwalającej polepszyć efekty uzyskiwanie dzięki zastosowaniu technologii OCR. Jednym z kluczowych obszarów współpracy były bliskie relacje z producentem i sprzedawcą oprogramowania OCR, firmą ABBYY. "Zdecydowaliśmy się współpracować z firmą ABBYY, gdyż oferowane przez nią oprogramowanie OCR jest powszechnie używane w procesie cyfryzacji przez europejskie biblioteki", twierdzi Dr Balk. "Firma ABBYY udostępniła nam swoje narzędzia i biblioteki programistyczne, natomiast współpraca z naszej strony polegała na wdrożeniu wyników naszych badań w produkty ABBYY. Bardzo satysfakcjonujące było obserwowanie jak wyniki naszych badań wykorzystywane są do ulepszenia produktu, który jest już obecny na rynku".

"Samo w sobie ulepszanie oprogramowania OCR nie było naszym celem", tłumaczy DR Balk, "gdyż jest ono w zaawansowanym stadium rozwoju, jednak specyfika tekstów historycznych sprawia, że oprogramowanie to jest mniej dokładne. Chcieliśmy opracować narzędzia, które uwzględnią powyższy kontekst historyczny".

Przykładowo teksty historyczne często mają skomplikowany układ, w postaci wielu kolumn czy dużych czcionek na początku akapitów. Co więcej, czcionki stosowane w tekstach historycznych są zwykle rzadko spotykane we współczesnych dokumentach. W ramach projektu Impact stworzono zbiór (zwany korpusem) złożony z 50\;000 cyfrowych transkryptów, skompilowany z ponad 500 000 zeskanowanych stron pochodzących z kilku europejskich bibliotek narodowych. Te tak zwane "wzorce" są potwierdzonymi, prawie doskonałymi konwersjami, używanymi do "szkolenia" oprogramowania OCR w zakresie rozpoznawania czcionek oraz radzenia sobie z nietypowymi układami stron, a także do weryfikowania prawidłowości działania aplikacji.

Uczestnicy projektu tworzą ponadto słowniki historyczne, dzięki którym możliwe jest polepszanie jakości konwersji wykonywanej przez programy OCR. Oprogramowanie OCR podczas konwersji analizuje skan, zapisując każdy rozpoznaną literę, następnie układając litery w "słowa" i sprawdzając, czy dane słowo rzeczywiście istnieje; Jeśli dane słowo nie zostanie odnalezione w słowniku, to aplikacja zwykle zgaduje inne, o podobnej pisowni.

Jednak większość programów OCR korzysta ze współczesnych słowników, zawierających współczesne słowa. "Naukowcy pragną czytać faktyczną treść dokumentów, w których pisownia wyrazów jest oryginalna", twierdzi Dr Balk, "jednak szukając danego materiału wolelibyśmy uniknąć uwzględniania 10 czy wręcz 50 wersji danego wyrazu. W związku z powyższym w ramach inicjatywy "skompilowane słowniki" - 'compiled dictionaries' powiązaliśmy ogromną ilość słów i sposobów ich zapisywania dla dziewięciu języków z ich współczesnymi synonimami i ich pisownią. Dzięki temu oprogramowanie OCR będzie mogło zarówno dokonać wiernej konwersji dokumentu, jak i zastąpić historyczne słowa ich współczesnymi odpowiednikami. Powyższy słownik pozwala zwiększyć dokładność, elastyczność

oraz użyteczność procesu cyfryzacji".

Dotyk człowieka

W przypadku masowej cyfryzacji istotne jest, by powyższe narzędzia działały w sposób automatyczny - jeśli mamy do czynienia z milionami stron wymagających cyfryzacji, to niemożliwe jest, by dokładność tego procesu sprawdzali ludzie. Mimo to uczestnicy projektu opracowali nowatorskie technologie, które umożliwiają użytkownikom łatwe i szybkie sprawdzanie efektów zastosowania oprogramowania OCR.

Językoznawcy obliczeniowi z Uniwersytety w Monachium opracowali algorytm , który pozwala określić prawdopodobieństwo poprawności konwersji poszczególnych słów w procesie OCR. Powyższy algorytm uwzględnia czas powstania dokumentu, język, w jakim został napisany, informacje na temat ustalonych wzorców pisowni oraz wiedzę z zakresu historii języków. Dzięki powyższym danym system potrafi przykładowo określić, czy błędnie zapisane słowo jest efektem niepoprawnej pracy oprogramowania OCR (wówczas słowo zostanie wyróżnione w tekście), czy też błędny z punktu widzenia współczesnej pisowni zapis ma swój historyczny odpowiednik.

Naukowcy z firmy IBM Israel Science and Technology opracowali inny system, bazujący na nowatorskim podejściu do technologii OCR. System 'adaptive OCR' , zwany również CONCERT , to dodatkowe, inteligentne i kolaboracyjne narzędzie korygujące. Do uczestnictwa w jego tworzeniu zaproszono ochotników, których rola polega na zwiększaniu dokładności konwersji OCR dzięki korektom błędów wprowadzanym przez ludzi.

"Uczestnicy projektu Impact stworzyli zestaw narzędzi testowanych obecnie przez partnerów projektu, w celu określenia ich wpływu na dokładność i wierność konwersji", zauważa Clemens Neudecker kierownik techniczny projektów unijnych w bibliotece Koninklijke Bibliotheek. "Pragniemy ocenić zarówno ich indywidualny wpływ na jakość konwersji, jak i efekty ich sumarycznej pracy w ramach łańcucha obróbki realizowanej po ukończeniu procesu skanowania. Chcemy się także upewnić, że wszystkie powyższe narzędzia będą ze sobą kompatybilne. W związku z tym udostępniliśmy technologiczną platformę architekturalną , pozwalającą bibliotekom korzystać z narzędzi i przetwarzać cyfrowe wersje dokumentów bez konieczności martwienia się o formaty plików i ich konwersje".

Przewidywany czas ukończenia projektu to czerwiec 2012, jednak kolektywna wiedza uczestników tej inicjatywy oraz ich doświadczenie w zakresie tworzenia narzędzi do cyfryzacji oraz korzystania z nich udostępniane są obecnie szerokiej społeczności zainteresowanej cyfryzacją, za pośrednictwem Centrum Kompetencji Impact .

Projekt IMPACT uzyskał wsparcie finansowe na badania naukowe w wysokości 12,1 milionów euro (całkowity budżet projektu wyniósł 17,1 milionów euro) w ramach podprogramu TIK, będącego częścią Siódmego Programu Ramowego UE (7PR).

Użyteczne odnośniki:

- Strona internetowa projektu "Ulepszanie dostępu do tekstów" - 'Improving access to text'
- Informacje na temat projektu IMPACT w bazie danych CORDIS
- Centrum Kompetencji Impact
- TIK Wyzwanie 4: Biblioteki i treści cyfrowe ('ICT Challenge 4: Digital libraries and content'
- Europeana

Odnośne publikacje:

- hhttps://cordis.europa.eu/article/id/88696-feature-stories-digitising-our-cultural-heritage/pl (Najważniejsze wiadomości - Cyfryzacja naszego dziedzictwa kulturowego)

Powiązane projekty

ARCHIVED

IMProving ACcess to Text

IMPACT

5 Kwietnia 2023

PROJEKT

Ostatnia aktualizacja: 20 Maja 2013

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