Prepoznavanje rukopisnog teksta (HTR) kao revolucija arhivskog svijeta

Kaisa Luhta

Projekt READ - Recognition and Enrichment of Archival Documents (Obogaćivanje i prepoznavanje arhivskih dokumenata) započeo je u siječnju 2016. godine u okviru Horizon 2020. programa Europske unije. Fokus projekta bio je poboljšati dostupnost arhivskog materijala unaprjeđenjem tehnologija prepoznavanja uzoraka, računalnog vida i obrade prirodnog jezika. Na projektu je surađivalo 14 partnera i nekoliko potpisnika Memorandum o razumijevanju (*Memorandum of Understanding*), iz područja računarstva, arhivistike i humanistike.

Najvažniji proizvod projekta READ je istraživačka platforma za manipulaciju digitalnim dokumentima Transkribus. Dokumenti na platformi su privatni s mogućnošću dijeljenja s pojedinim korisnicima. Platforma je alat za transkripciju, obilježavanje i pretraživanje dokumenata, ali i okruženje za korištenje automatskog prepoznavanja strukture i razumijevanja dokumenta. Moguće je koristiti sustav za identifikaciju autora, optičko prepoznavanje znakova (*Optical Character Recognition*, OCR) i upotrebljavati transkribirane dokumente za treniranje prepoznavanja rukopisnog teksta, tj. HTR (*Handwritten Text Recognition*) modela koji je moguće primijeniti na vlastite dokumente.

Za razliku od OCR-a, HTR modeli moraju biti trenirani na specifičnom gradivu za koje će biti korišteni i iz tog razloga je ova tehnologija posebno korisna za velike zbirke. Također, HTR modeli procesuiraju tekst na osnovi linija, a ne kao OCR na osnovi znaka. Broj stranica potrebnih za treniranje pouzdanog modela ovisi o vrsti gradiva koje se želi prepoznati, primjerice, gradivo koje je pisalo nekoliko osoba (više ruku) zahtjeva više stranica za izgradnju modela, dok će za gradivo koje je pisala ista ruka biti potreban manji broj.

Finski državni arhiv je jedan od partnera na projektu READ u kojem je platforma primjenjena na velikoj količini gradiva. Arhiv je projektu ponudio veliku količinu već digitaliziranih dokumenata koji uključuju okružne sudske zapise iz 19. stoljeća (koji su postali glavni fokus primjene HTR modela u okviru Finskog državnog arhiva), poreznih zapisa i dnevnika iz razdoblja Drugog svjetskog rata. Ovo je gradivo izabrano za obradu u projektu na temelju velikog broja zahtjeva za njegovim korištenjem.

Da bi HTR model testnog gradiva bio uspješno istreniran bilo je potrebno izraditi prijepise dokumenata koji će poslužiti kao temeljni tekst (*ground truth*). S obzirom da takvih prijepisa nije bilo u velikom broju za taj je zadatak angažiran vanjski suradnik, a u manjoj mjeri je izvršen i uz pomoć volontera Finskog genealoškog društva i kroz predavanja na finskim sveučilištima.

Za izradu završnog HTR modela sudskih okružnih dokumenata bilo je potrebno izraditi temeljne tekstove za 2.700 stranica jer su ovi dokumenti bili pisani od strane velikog broja osoba čiji se rukopisi međusobno razlikuju. Točnost prepoznavanja završnog modela je od 92 do 95 %.

Nakon završetka READ projekta u lipnju 2019. godine osnovano je SCE READ-COOP udruženje čiji je cilj unaprjeđivanje ovog modela i provođenje do sada uspostavljenih usluga. Finski državni arhiv u siječnju 2019. godine započeo je s projektom „Making a Modern Archive“ (Stvaranje suvremenog arhiva) koji će implementirati tehnologije i radne procese uspostavljen READ projektom. Jedan od glavnih ciljeva novoga projekta je uspostava sučelja za pretraživanje okružnih sudskih dokumenata iz 19. stoljeća prepoznatih HTR modelom izrađenim u okviru READ-a.

Sučelje dopušta korisnicima da pretražuju puni tekst dokumenta i pregledavaju transkribirani tekst usporedno s digitaliziranim datotekama. Ovaj alat koristi koncept uočavanja ključnih riječi (*keyword spotting*, KWS) kojim se pretraživanje ne vrši na transkribiranom tekstu već se koriste podaci o pouzdanosti prepoznavanja (*confidence matrices*) zabilježeni HTR modelom.

Prednost ovakvog načina rada je u tome što se rezultati pretraživanja ne sastoje od samo jednog pojma ili fraze već od više rezultata s odgovarajućim podacima o pouzdanosti prepoznavanja. U ovom slučaju, prepoznavanje ne mora biti apsolutno točno da bi KWS bio od koristi (KWS daje rezultate čak i na dokumentima čija je točnost prepoznavanja samo 80 %).

Javno objavljeni okružni sudski dokumenti su i u svojem elektroničkom obliku izrazito zahtjevni za proučavanje i čitanje te se nadamo da će sučelje za pretraživanje olakšati istraživački posao, ali i da će primjena HTR tehnologije općenito povećati iskoristivost rukopisnih dokumenata u budućnosti. ■

INFO

Transkribus platforma dostupna je na: https://transkribus.eu/

Sudske zapise iz 19. st. u Finskoj pretražite na: https://transkribus.eu/r/kansallisarkisto/en

Više o SCE READ-COOP aktivnostima saznajte na: https://read-coop.eu/