A mesterséges intelligencia használata a levéltári gyakorlatban

Beszámoló a Magyar Levéltárosok Egyesülete Informatikai Szekciójának szakmai napjáról

A Magyar Levéltárosok Egyesülete (MLE) Informatikai Szekciójának a Nemzeti Kulturális Alap által támogatott *A mesterséges intelligencia használata a levéltári gyakorlatban* című szakmai napjára 2023. november 6-án került sor Budapest Főváros Levéltárában (BFL). A szakmai nap tematikájának meghatározásánál szempont volt, hogy nem levéltárban, de rokon területeken dolgozó szakemberek mutassák be a mesterséges intelligencia levéltári alkalmazhatóságát, új ötletek, megoldások bemutatása révén inspirálva a levéltári szakmát. Délelőtt a plenáris előadásokra került sor a BFL Gárdonyi Albert-termében, ezeken minden MLE-tag részt vehetett, míg 14 és 16 óra között kiscsoportos gyakorlati foglalkozások voltak az előzetesen regisztráltak számára a BFL Oktatótermében és Tárgyalótermében. A plenáris ülésen, amelyen Gerhard Péter, a szekció vezetője elnökölt, négy előadást hallgathattak meg a résztvevők.

Az első előadást Biszak Sándor (Arcanum Adatbázis Kft.) tartotta *Mesterséges intelligencia alkalmazása az Arcanum adatbázisaiban* címmel. Az előadás elsősorban a mesterséges intelligencia gyakorlati alkalmazását mutatta be Magyarország (és egyúttal immár Románia) legnagyobb közgyűjteményi tartalomszolgáltató műhelyénél, ahol elsősorban a keresések optimalizálását igyekeznek támogatni a mesterséges intelligencia segítségével. Neurális hálózatokat az Arcanumnál először az arcfelismerés kapcsán vetettek be. Az Amazon-felhő neurális hálójába töltötték fel nagy munkával az Arcanum által digitalizált 52 millió oldalon található kb. 12 millió „arc"-ot, az eredmény világszerte egyedülálló lett. Az arcok felimertetése még festményeken is ért el eredményeket. Emellett a Főfotó anyagára (ami majdnem kétszázezer képet jelent) alkalmaztak egy automatikus címkeosztályozást. Az Arcanum következő lépése a természetesnyelv-feldolgozás bevetése volt. Néhány évvel ezelőtt a Google elérhetővé tette az ún. BERT-modelljét, a ChatGPT közvetlen elődjét, ez is egyfajta neurális háló, ami alkalmas arra, hogy bármelyik nyelvre alkalmazzuk, csak az kell hozzá, hogy 10-20-30 milliárd szóval és mondattal betanítsuk. A tanítás után ezen a magyar nyelvű adatbázison sikerült felismertetni a szövegállományokban a tulajdonneveket (ilyen pl. Pereg mint község). A régi román nyelvű szövegek keresését megnehezítette az egykori ószláv/cirill írásbeliség, ezért itt az átírást is meg kellett oldani. Ehhez a Tesseract nevű, nyilvánosan hozzáférhető (az elég drága Amazon Textractnál kevésbé jó, de legalább ingyenes) OCR-programot használták. Sok ezer oldal betanítása után ez is használható lett, ami azért is nagy előny, mert a románok nagy többsége ma nem tud cirill betűket olvasni. A következő lépés az MI alkalma-

zása terén az ún. Newspaper Segmentation kialakítása volt, amely óriási méretű régi napilapok cikkenkénti szegmentálását segíti elő. A szegmentáláshoz először a bekezdések sorrendjét kell megállapítani, majd azt, hogy milyen részekre oszthatók az újságoldalak. Így sikerült például szerzők, képaláírások, lábjegyzetek szerint elkülöníteni a találatokat. A szegmentáláshoz DeepLabot használnak, ez egy olyan neurális háló, ami a képpixelek osztályozására szakosodott, és több százezer oldalnyi betanítás után már értelmezni is tudja a képrészeket. Ezeket az eredményeket ma már az Egyesült Államokban is el tudják adni, a 300 millió oldalt bemutató newspaperarchive.com amerikai újságadatbázisban már szintén az Arcanum technológiáit hasznosítják.

A második előadó Répászky Lipót, az MTVA-nál működő Nemzeti Audiovizuális Archívum vezetője volt, aki *Hány arca van egy személynek?! MI a Nemzeti Audiovizuális Archívumban* címmel tartotta meg előadását. Ő elsősorban az 1910-ig visszanyúló, 13 millió fényképet tartalmazó fotóadatbázis kereshetővé tételéről beszélt (emellett jelentős az újságok, a rádióadások és a TV-felvételek archívuma is az MTVA-nál). Korábban a levéltárosi, kulcsszóalapú feldolgozás egy fotó leírása esetén átlagosan 12 percnyi időt vett igénybe, ha a levéltáros egyáltalán felismerte, mi van a képen. Ezzel a módszerrel 12 ember munkáját számításba véve e fotóadatbázis feldolgozása 412 évig tartott volna. Az elmúlt időszakban ily módon már feldolgozott 324 000 kép alkalmas viszont arra, hogy összehasonlítható legyen az ember és az MI munkateljesítménye. Először a Microsoft irányából indultak el, de végül három MI-modellt is ráeresztettek az adatbázisra, majd ezeket összerakták, egész hasonló eredménnyel, mint az emberi leírás (kivéve a nagyon rossz minőségű képeket). A három MI-rendszer egymás javaslatait validálja, így nem szükséges embernek is validálni az új felismertetéseket. Az MI segítségével ráadásul egyben leírhatjuk, mi látható a képen, nem csak az egyes kulcsszavakat, ami jobban hasonlít a természetes emberi keresőkifejezésekhez. 17 000 felismert személlyel végezték az arcfelismerés betanítását, az adatbázis mára már 41 000-re bővült, ezzel szemben az egyetlen arcfelismeréssel foglalkozó kolléga napi tíz arcot tud azonosítani. Az életkori sajátosságokat úgy tudja legjobban lekövetni az MI, ha legalább tízévente rendelkezik egy felismert arccal az illetőről. A betanítás kapcsán az előadó azt jelezte problémának, ha véletlenül „fals" személyt „tanult meg" az MI, ilyenkor nehéz a visszalépés. Az előadó végezetül a ChatGPT alkalmazhatóságáról beszélt az MTI híradatbankja alapján, a kérés az volt a ChatGPT felé, hogy vegye ki a helyszíneket, eseményeket, dátumokat egy hírből, majd adja vissza egy standard sémában, konkrétan JSON struktúrában. Ezzel nagyon könnyűvé vált a továbbfeldolgozás. Tanulságul pedig azt fogalmazta meg az előadó, hogy nem az alkalmazott MI-modell a lényeg, hanem a jól megfogalmazott kérdés.

Az első két előadáshoz kapcsolódó hozzászólások, majd a kávészünet után következő előadó, Szűcs Kata Ágnes (OSZK) *Gépi kézírásfelismertetés a közgyűjteményekben* címmel tartotta meg (Mihály Eszterrel és Varga Emesével közösen készített) előadását. Először az (eleinte a Petőfi Irodalmi Múzeum, 2022 márciusa óta pedig az) OSZK égisze alatt működő Digitális Bölcsészeti Központ, illetve annak digitális forráskiadási platformja, a dhupla.hu került bemutatásra. Az itt létrejövő digitális korpuszok sokrétűen haszno-

síthatók: annotációk, a szemantikus web, lekérdezések, statisztikák, adatvizualizációk létrejöttét segíti elő, nem utolsósorban pedig tanítókorpuszként használhatóak MI-alapú technológiákhoz. A dhupla.hu bemutatáskor egyrészt Móricz Zsigmond egy levelét mutatta be az előadó, amelyen már annotációk, tulajdonnevek kiemelése is láthatók, illetve megtekinthető a levélbe lerajzolt csizma Petőfi Irodalmi Múzeumban őrzött eredeti példányának fényképe is. Szűcs Kata Ágnes bemutatott egy, Kiss József levelezése alapján készült adatvizualizációt is, amelyen a költő kapcsolati hálója látható. A dHUpla workflowja a következőképpen épül fel: a digitalizált kéziratból egy gépi átirat, amiből pedig egy TEI XML annotált szöveg készül, ez található a megjelenítés mögött, ami egy GitLab alapú háttérszolgáltatás alatt fut, tehát a megjelenő facsimile kép mögött még nagyon sok információ rejtőzik, a felhasználóknak ezek kinyerését is meg kell ismerniük. A workflow második lépése, a gépi átirat a Transkribus segítségével készül. A Transkribus digitalizált kéziratok filológiai központú, de akár egyszerűsített szövegkiadására is alkalmas, átfogó, nyelvtől független platform, ami bárhonnan elérhető. A projekt egy európai uniós pályázatból 2013-ban indult, a mögötte 2016-tól álló európai szövetkezet, a READ (Recognition and Enrichment of Archival Documents), illetve 2019-től READ-COOP SCE ma 32 országból kb. 150 tagot számlál. 2023 tavaszától az OSZK is szavazati joggal rendelkező tag a konzorciumban. A tagok értesülnek a frissítésekről és tesztelhetik is azokat. A Transkribus nemcsak a kézírásfelismerésről szól, ők is építettek egy olyan ökoszisztémát, ami a szkenneléstől a publikálásig támogatja az online forráskiadást. Elsősorban azonban a Transkribus kézzel írott és nyomtatott dokumentumok MI-modellek tanításán alapuló automatikus átírását támogatja, emellett a már átírt dokumentumok kereshetőségét, a szerkezet és a tartalom címkézését, illetve a dokumentumok különböző formátumokba történő exportálását teszi lehetővé. Jelenleg két platformon működik, egy asztali és egy webes alkalmazásban, és most még az asztali tud többet, de ezt hamarosan kivezetik, és csak a webes interface marad meg. Lehetőség van továbbá API-k futtatására is, és akkor nem kell egy külső szerveren tárolni a dokumentumokat. A workflow a szegmentálással kezdődik, amelyben megállapításra kerül az adott dokumentum képi elrendezése, a szövegterületek, a sorok és az alapvonalak megállapítása. Ezután került sor a kézírásfelimertetésre. A HTR (Handwritten Text Recognition) hasonló eredményt ad, mint az OCR (Optical Character Recognition), mégsem ugyanaz a technológia, a HTR kézírásos vagy kora újkori nyomtatott szövegeket is fel tud ismerni, mert nem az egyes karaktereket, hanem a teljes sort veszi figyelembe. Ehhez egyéni MI-modellek tanítása is lehetséges, szemben a fix OCR-rel. A kézírásfelismerés is a neurális hálózatokhoz tartozik, ami a gépi tanuláson belüli mélytanulás egyik formája, mintája az emberi agy felépítése. Ennek működésbe hozásához kell egy tanuló korpusz, ami alapján meg tudja jósolni a rendszer a lehetséges jövőbeni kimeneteleket. A jelenlegi magyar nyelvi modell Kiss József levelezésén alapul, de tervezik ennek továbbépítését, kibővítését. A Transkribus esetében a fejlesztők szeretnének áttérni a jelenlegi kreditalapú, oldalanként fizetős modellről az előfizetéses modellre, amelyben intézményenként egyedi lehetőségek lennének a kézírás-felismertetés költségeinek megállapítására.

A délelőtti plenáris szekció utolsó előadását Sidó Zsuzsa és Szekrényes István (ELTE Digitális Örökség Nemzeti Laboratórium) *A DH–LAB fejlesztései közgyűjteményi állományok feldolgozására* címmel tartotta meg. Sidó Zsuzsa elmondta, hogy az Innovációs és Technológiai Minisztérium (ITM) kezdeményezésére 2020-ban jött létre összesen 18 darab laboratórium, az ELTE Digitális Örökség Nemzeti Laboratórium főleg a bölcsészet- és társadalomtudományok területéért felelős. A DH-LAB-ot egy konzorcium alkotja, amelynek tagjai a Magyar Nemzeti Levéltár, a Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézete, illetve a Miskolci Egyetem. A laboratórium célja, hogy kidolgozza a nemzeti kulturális örökség mesterségesintelligencia-alapú kutatásának, oktatásának és közzétételének lehetőleg legszélesebb körű módszertanát, és a kialakított kompetenciák mind közgyűjteményi, mind piaci területen hasznosuljanak is. Projektek és témacsoportok is vannak, a mesterségesintelligencia-témacsoport vezetője Nemeskey Dávid, de szoros a kapcsolat az ELTE illetékes tanszékével is. Foglalkoznak mélytanulásra épülő magyar nyelvmodellekkel, „born digital curation"-nel is, de adatgazdász („data steward") képzést is indítanak. Magyarországi és határon túli közgyűjteményekkel (pl. Móra Ferenc Múzeum, Kereskedelmi és Vendéglátóipari Múzeum) indultak már közös kézírás-felismertetési pilot projektek. Közgyűjteményi projektjeik közül egy esettanulmány került bemutatásra, a Kolozsvári Állami Magyar Színház jelmezterveinek digitalizálási projektje, amelyben a Magyar Nemzeti Levéltár és a Forum Hungaricum működött közre.

Szekrényes István azzal folytatta az előadást, hogy a DH-LAB-nál a HTR mellett foglalkoznak automatikus leiratozással (OPEN AI/Whisper alapú, nagy mennyiségű hanganyagon előre tanított beszédfelismerő modellek magyarra finomhangolásával), illetve MI alapú, kézi javítással finomhangolható optikai karakterfelismeréssel is. Ez utóbbihoz ők is a Tesseract rendszerhez készítettek saját tanítóanyagot, illetve modelleket. HTR-fejlesztési projektjük célja, hogy csak kézírásos formában elérhető gyűjteményekről olyan webszolgáltatást hozzanak létre, amellyel azokból kereshető, strukturált dokumentumok tömeges mennyiségben legyenek generálhatóak, az eddig is elérhető fizetős szolgáltatások helyett szabadon és ingyenesen felhasználható eszközökre építkező, egyúttal – Rest API-n keresztül – integrálható alternatíva révén. A tanítóanyagot ugyan ők is a Transkribussal hozták létre, de a cél az, hogy a csatlakozó közgyűjtemények ingyenesen használható eszközhöz jussanak, ezért nyílt forráskódú eszközök használhatóságát vizsgálták meg. Egyelőre a TrOCR, a Microsoft által fejlesztett, nyílt forráskódú, ingyenesen használható keretrendszer vált be a legjobban, ez nem feltétlenül kézírásfelismertetésre való, de lehet úgy finomhangolni, hogy a kézírásra is megfelelően működjön. A finomhangolást saját tanítóanyagon végezték, a legnagyobb modelljük Arany János levelezése és hivatali iratai alapján készült, ez kb. öt szerzőtől 900 oldalnyi kézirat, a tesztelések során 5,86%-os CER (karakterhiba-arány – ez jobb, mint a Transkribus aránya!) és 22,36%-os WER (szóhiba-arány) lett az eredmény. A másik modelljüket a Rerum Ungaricarum Libri korpusz alapján („Brutus", egy szerző, 200 oldal, de a nyomtatott szöveget megközelítő minőségű kódexszerű kézírás) készítették el, itt még jobb eredményeket értek el (2,03%-os CER és 9,54%-os WER). A harmadik modelljük

a Magyar Nemzeti Levéltártól kapott 300 oldalnyi, egy szerzőtől származó anyakönyvi iratokon alapult, itt 2,77%-os CER és 11,67%-os WER lett az eredmény. Terveik között szerepel e három modell összehangolása is. A szövegkép szegmentálásához két eszközt implementáltak, az egyik a Transkribusban is használt, az alapvonalakat detektáló P2Pala, ez PAGE-XML-kimenetet produkál, de csak az alapvonalakat ismeri fel, ezért nem minden kézírásnál alkalmas a sorok detektálására. A másik ilyen eszköz a Kraken blla szegmentáló eszköze, ami szövegblokkokra és sorokra szegmentál, ez JSON kimenetet produkál. A Kraken maga is egy teljes HTR-keretrendszer, viszont nem latin betűs írásokra lett optimalizálva, de a szövegblokkok és sorok elkülönítésére ez a modulja nagyon is alkalmas. A kialakított rendszerük egyik kimeneti formája az Alto-XML, illetve fejlesztés alatt van egy PAGE-XML kimeneti forma is, továbbá tervezik a JSON kimenetet is. Letölthető emellett a szegmentálást bemutató jpg, illetve kétrétegű pdf-kimenet is.

Az utolsó két előadáshoz kapcsolódó vita, amelyen a Magyar Nemzeti Levéltár tapasztalatairól is szó esett, majd az ebédszünet után került sor a kiscsoportos gyakorlati foglalkozásokra. A regisztráltak két turnusban, egymást váltva ismerkedhettek meg egyrészt Mihály Eszter és Varga Emese útmutatásai alapján a Transkribus használatával, másrészt Sidó Zsuzsa és Szekrényes István segítségével a DH-LAB demófelületével, ahol a Kraken blla szegmentáló eszköze és a TrOCR eredményei kerültek bemutatásra.

Gerhard Péter