VALÓS TÉRBEN – AZ ONLINE TÉRÉRT

Networkshop 31: országos konferencia

2022. április 20–22.
Debreceni Egyetem

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2022

NETWORKSHOP 2022

A kötet megjelenését támogatta az
Energiaügyi Minisztérium

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2022. április 20–22. Debreceni Egyetem, konferencia előadásainak közleményei

ISBN 978-615-82243-0-7
DOI: 10.31915/NWS.2022

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2022

Borítókép: freepik.com

TARTALOMJEGYZÉK

DOI: 10.31915/NWS.2022.30

Mesterséges intelligencia, digitális bölcsészet, kulturális örökség: trendek és eredmények

Palkó Gábor
ELTE Digitális Bölcsészet Tanszék, Digitális Örökség Nemzeti Laboratórium
palko.gabor@btk.elte.hu
ORCID: 0000-0002-4394-8577

A XXI. század első évtizedeiben a kulturális, tudományos és oktatási területen két párhuzamosan zajló, egymáshoz szorosan kapcsolódó trendnek lehetünk tanúi. Egyik oldalról a mesterséges intelligencia (MI) beláthatatlan ütemben ír át és vált le különféle bevett kutatási, vizsgálati módszertanokat, de túlzás nélkül állíthatjuk, az MI a mindennapi életünkre is egyre nagyobb hatást gyakorol. A másik oldalról pedig – részben a kulturális örökség digitalizálása révén, részben pedig a digitálisan létrejövő (born digital) anyagok termelődésének óriási volumene következtében – szó szerint – beláthatatlan nagyságrendű adatterek és -hálózatok jönnek létre.

Ha a digitálisan hozzáférhető dokumentumok mennyisége többszörösen meghaladja a források feldolgozása során következtetéseket megfogalmazni képes szakember befogadóképességét, és ha a digitális dokumentumok özöne számítógép segítsége nélkül már nem tekinthető át, akkor a források közreadása, kutathatóvá tétele, illetve a jövő generáció digitális kompetenciáinak formálása terén a kutatásnak, az oktatásnak és az archiválásnak eleve számolnia kell a gépi intelligencia teremtette új helyzettel.

A mesterséges intelligencia alapú nyelvelemző eszközök fejlesztéséhez szükség van nagy tömegű – történeti forrásokat, sajtóanyagot, médiatermékeket, szépirodalmat és web 2.0-es forrásokat egyaránt tartalmazó – magyar nyelvű anyagra. A heterogén szövegkorpusz egyrészt a többcélú nyelvmodellek készítéséhez szolgál alapul, másrészt olyan bölcsészeti, társadalomtudományi vagy éppen piaci horizontú kutatásokat, fejlesztéseket tesz lehetővé, amelyek az oktatásban, a tudományos diskurzusban, valamint a nagyközönség számára készülő szolgáltatásokban is hasznosíthatók.

Ahhoz, hogy nagy méretű és kitűnő minőségű magyar korpuszok, az ezekből előállított tanítóanyag, továbbá magas színvonalú programozási architektúra álljon rendelkezésre, digitális bölcsészek, számítógépes nyelvészek, informatikai szakértők összehangolt munkájára van szükség. De még ez sem elég: elengedhetetlen a kutatóhelyek, az egyetemek, a kulturális örökséget megőrző intézmények (a „memória-intézmények"), a piaci és kormányzati szereplők szoros együttműködése.

Ebben a rövid előadásban arra vállalkozom, hogy ennek a nemzetstratégiai szempontból elodázhatatlan együttműködésnek a kulturális örökség és a bölcsészettudomány szempontjából legfontosabb gyakorlatait és követendő példáit bemutassam. Előadásom első felében a mesterséges intelligencia, illetve a digitális kulturális örökség kultúrtechnikái közül emelem ki azokat, amelyek – meglátásom szerint – a legnagyobb jelentőséggel bírnak a jövőre nézve. Az előadás második felében pedig az együttműködés intézményi formáira térek ki.

Mélytanulásra épülő magyar nyelvmodellek fejlesztése

Az okoseszközök elterjedése révén az a tény, hogy a hangzó és írott nyelvi megnyilvánulásainkat a mesterséges intelligencia alapú alkalmazások nemcsak értelmezik, de reagálnak is rájuk, mára általános tapasztalattá vált. Azt azonban, hogy ez a változás milyen hatással lesz a kultúra létrehozására és közvetítésére, senki sem tudja.

Az a tény, hogy a magyar nyelvű MI fejlesztésre szorul, és ez hatalmas költségekkel jár, sokak számára csak az ITM és az OTP Bank közös szuperszámítógép-projektjének bemutatása kapcsán vált világossá.[1] Az elmúlt években a mélytanulásos technológiára épülő nyelvmodellek fejlesztése terén több kutatóhely is jelentős eredményeket ért el, és ezek a teljesítmények lassan beépülnek a kutatási és kulturális célokra fejlesztett szolgáltatásokba. Csak két példa: a Nyelvtudományi Intézet és a Pécsi Tudományegyetem (PTE) közös terméke, a HILBERT, melynek létrehozását a Microsoft Hungary segítette, jelentős korpuszépítési és számítástechnikai kapacitást mozgósított; a DH-LAB munkatársa, Nemeskey Dávid által készített HuBERT a magyar nyelvre jelenleg a legpontosabb eredményeket produkáló modell. A hírek szerint viszont Feldmann Ádám (PTE) május elején egy új nyelvmodellt mutat be a Budapest ML Fórumon, a gépi tanulással és mesterséges intelligenciával foglalkozó konferencián.

Az ITM-OTP projekthez visszatérve: ez a fent említett együttműködés eminens példa lehet arra, amikor a magyarországi kutatóhelyek és piaci szereplők egy jól körülhatárolt, közös cél érdekében egyesítik szellemi és fizikai erőforrásaikat, belátva, hogy jó minőségű eredmény csakis így érhető el.

Miért fejlesztünk gigantikus nyelvmodelleket? A bankoktól a bölcsész kutatócsoportokig miért van szükségük nyelvmodellekre azoknak, akik emberi megnyilvánulásokat elemeznek és az emberek számára érthető szövegeket generálnak? A válasz elég egyszerű: a mesterséges intelligenciára épülő új technológia információkinyerési és szövegalkotási képességei messze meghaladják a korábbi technológiák lehetőségeit. Az MI néhány év alatt nemcsak a nyelvtechnológiát forradalmasította, de olyan versenyhelyzetet teremtett, amelyben Magyarország nem engedheti meg magának a lemaradást, mert az kulturális és piaci értelemben egyaránt hatalmas veszteséget jelentene. Nem kell magyaráznunk: a magyar nyelvre optimalizált modellek kiemelkedő minősége csak magyar szakemberekkel garantálható.

A létrejövő modellek felhasználása széles körű, egyaránt szolgálja a nemzeti örökség feltárását, a határon túli örökség integrálását, valamint a közvetlen, innovatív piaci megoldások fejlesztését, a beszéd- és kézírás-felismerés, szövegkivonatolás, szemantikus keresés, ügyfélkapcsolat-automatizálás stb. területén.

Egy valamiről azonban kevesebb szó esik a nyelvmodellek fejlesztésében, ahol többnyire a hatalmas számítástechnikai kapacitás kerül csak reflektorfénybe. Az, hogy egy olyan, csekélyebb méretű digitális adatforrásokkal rendelkező nyelv esetén, mint a magyar, a modell tanításához szükséges nyelvi nyersanyag, a korpusz előállítása rendkívüli erőforrásokat igénylő munka, amely nélkül azonban az egész projektum végkimenetele kétes. (Ezért vagyunk különösen büszkék arra, hogy az ITM-OTP projektben a Digitális Örökség Nemzeti Laboratóriumot bízták meg annak a tanítókorpusznak a felépítésével, amely minden eddiginél nagyobb mennyiségű magyar nyelvű szöveg összegyűjtését feltételezi.)

Milyen forrásokból állítható elő ez a bizonyos tanítókorpusz? A szövegek döntő többsége webes eredetű, born digital, vagyis digitálisan született: nyilvános weboldalakról származik. A repozitált gyűjteményi tartalmak, digitálisan született vagy digitalizált könyvek és dokumentumok is részei azonban ennek a minden képzeletet felülmúló méretű és komplexitású korpusznak, ami természetesen nem csak arra való, hogy hatalmas nyelvmodelleket tanítsanak rajta. A korpusznak magának is a kutatás tárgyának kell lennie, hiszen olyan mintázatok felismerését teszi lehetővé, amelyek korábban kivehetetlenek lettek

1 Ld. pl.: Átadták az ITM és az OTP Bank együttműködésével elkészült szuperszámítógép első egységét. https://kormany.hu/hirek/atadtak-az-itm-es-az-otp-bank-egyuttmukodesevel-elkeszult-szuperszamitogep-elso-egyseget

volna. Mindez persze csakis a számítógéppel támogatott olvasás, a „distant reading” (Franco Moretti) horizontján értelmezhető.

Nincs tér itt a korpusz- és szolgáltatásépítés jogi kérdéseit részletesen tárgyalni. Egyetlen megjegyzés: a hazai digitális bölcsészeti kutatások, és általában valamennyi MI alapú fejlesztés számára kedvező fejlemény, hogy a magyar jogalkotó implementálta a szerzői jogi törvénybe azt az EU által kezdeményezett módosítást, amely a szabad felhasználás körébe vonta a szöveg- és adatbányászatot.[2]

De visszatérve a gigantikus magyar nyelvű korpuszok kutatásához: ezek értelmes feldolgozása, kutathatóvá tétele elképzelhetetlen a szemantikus technológiák fejlesztése nélkül. Ezek felhasználják a fent említett, már részben rendelkezésre álló mélytanulásos nyelvmodelleket, de messze nem azonosak velük.

Szemantikus technológiák a távoli olvasás szolgálatában

A digitális kulturális örökség szövegtere a folyamatos digitalizáció és a digitálisan született anyagok mennyiségének gyorsuló növekedésével elérte azt a kritikus tömeget, ami a hagyományos lassú vagy közeli olvasáson alapuló adatgyűjtés számára áttekinthetetlen. Igen korlátozott megoldást jelent a szabad szöveges keresés a dokumentumok szövegeiben, különösen az agglutináló nyelvek esetében, hiszen a releváns találatok gyakran elvesznek a kutató számára irreleváns szöveghelyek tengerében. A probléma kézenfekvő megoldása a névelemfelismerő és névelem linkelő algoritmusok használatában rejlik, ezek azonban a magyar nyelvre jelenleg kevésbé hatékonyak, másrészről pedig ezen technológiák komoly informatikai ismereteket feltételeznek, egészen addig, amíg be nem épülnek a szélesebb közösséget megcélzó szolgáltatásokba.

Az egyik legkomplexebb ilyen szemantikus technológia a wikification (wikifikálás). A repozitált anyagok (pl. OCR-ezett dokumentumok vagy a webaratásból származó cikkek) egy úgynevezett wikifikációs eljáráson mennek keresztül, melynek során a szövegobjektumokból kinyert szöveg egyes szavaihoz és szókapcsolataihoz a Wikidata tudástár elemeit rendeljük hozzá. Léteznek többnyelvű hasonló eszközök, de ezek a magyar nyelvre nem hatékonyak.

Egy pilot keretében a SZTAKI és a Társadalomtudományi Kutatóközpont a már rendelkezésre álló wikification algoritmusok felhasználásával metaadatokat rendelt repozitóriumi tartalmakhoz. (Micsik András, Gárdos Judit).[3]

A Digitális Örökség Nemzeti Laboratórium konzorciumi együttműködésben saját, magyar nyelvre optimalizált wikifikáló eszközt fejleszt, amelyet a holnapi napon mutatunk be a Networkshopon.[4]

A névelem-felismerés, vagyis amikor egy szöveg egyes szavairól, kifejezéseiről megállapítjuk, hogy földrajzi, személy- vagy egyéb tulajdonnevek-e, vagy az érzelemfelismerés, amikor a szöveg mondatainak érzelmi töltetét a számítógép ismeri fel, a wikifikáció, amikor a szöveg szavait egy ontológia elemeihez kötjük: mindez nehezen ér el a nagyközönséghez, ameddig nem integrálják népszerű szolgáltatásokba. Mindez elmondható a konkrét gyűjteményi anyagokon finomhangolt szövegfelismerő eszközökről is. Ez a következő kultúrtechnika, amelyre előadásomban röviden kitérek.

---

2 Vö. pl. https://ekk.org.hu/2021/10/05/az-uj-szerzoi-jogi-torveny-a-konyvtaraknak-jobb-1-resz/

3 A kutatók eredményeiket itt ismertették: https://milab.tk.hu/hirek/2021/06/2021-junius-17-gardos-judit-es-micsik-andras-eloadasa-a-tk-milab-speaker-series-sorozat-kovetkezo-allomasakent

4 Nemeskey Dávid Márk, Palkó Gábor – Szemantikus névelem-azonosítás (NEL) magyar nyelvű szövegeken (a HuWikifier bemutatása)

Írott szövegek MI alapú feldolgozása

Finomhangolt karakterfelismerés (Optical Character Recognition – OCR)

A gyűjteményi és a kutatási szférában az esetek többségében olyan eszközökkel és munkamenetek során állítanak elő digitális szövegobjektumokat, amelyek végeredményeként MI feldolgozásra kevéssé alkalmas produktumok állnak elő (pl. képi és vagy szövegréteg szintjén gyenge minőségű PDF fájlok). Olyan munkamenet kidolgozására van szükség, amely már a gyűjteményi feldolgozás során tekintettel van a gépi szövegfeldolgozás igényeire. Az egyes nyelvekre optimalizált alapmodellek használata mellett szükség van ugyanakkor a gépi tanulás alapú karakterfelismerésre is. Le kellene váltani azt a gyakorlatot, hogy a könyvtárak és más gyűjtemények anélkül választanak eszközt, illetve alakítanak ki munkamenetet a nyomtatott anyagok OCR-ezésére, hogy mérlegelnék a technológia előnyeit és hátrányait, összevetnék a különféle szoftverek, különféle szoftverbeállítások eredményeit és a saját gyűjteményükre optimalizált eszközöket és munkameneteket dolgoznának ki. Egy dokumentumtípus vagy gyűjteményrész feldolgozásánál kis mennyiségű szöveg kézi javítása is az egész korpusz radikális szövegminőség-javulásával jár.

Kézírás-felismertetés (Handwritten Text Recognition – HTR)

A digitális örökség diskurzusában a könnyen feldolgozható és közzé tehető nyomtatott, illetve digitálisan született anyagok mellett a „valódi" – vagyis kézzel írt – kéziratok háttérbe szorulnak, hiszen általános, nem az adott kézre vagy gyűjteményrészre hangolt modellekkel nem tehetők kereshetővé. Szerencsés fejlemény, hogy létezik egy könnyen kezelhető kézírás-felismerő keretrendszer, amelybe az OSZK munkatársainak jóvoltából már magyar kézírásra tanított kész modell is elérhető.[5] Az eszköz hátránya, hogy a szoftver nagyobb gyűjtemények feldolgozásánál igen költséges, azért csak néhány intézmény használja. Szükség van tehát ingyenesen elérhető, magyar nyelvre optimalizált, nagy gyűjteményi egységek feldolgozására is használható kézírás-felismerő eszközre.[6]

A szemantikus szövegfeltárás és a gépi tanulással optimalizált írásfelismerés munkamenetének széles körű elérhetővé tétele a Digitális Örökség Nemzeti Laboratórium egyik fő feladata, de erre később, az intézményi fejlemények kapcsán térek vissza. Előbb még egy negyedik kultúrtechnika, a „born digital curation" témáját emelném ki.

A digitálisan született anyagok kezelése („born digital curation")

A magyar kulturális örökség nagy volumenű adatvesztést szenved el amiatt, hogy a magyar közgyűjtemények és kutatóhelyek többsége nem rendelkezik sem kompetenciával, sem infrastruktúrával, de még legtöbbször tervekkel sem a digitálisan született anyagok archiválására és kezelésére. A digitálisan született anyagok kezelése nem csak gyűjteményi vagy tudományos feladat, de piaci szereplők is végezhetik azt. A különféle, adott esetben elavult hordozókon tárolt, és/vagy elavult formátumú digitálisan rendelkezésre álló anyagok kezelése speciális szaktudást és eszközkészletet igényel. Nemzetközi összefogásban

5 A Transkribus szolgáltatás első magyarországi modellépítéséhez l.: Bobák, Barbara és Gábori Kovács, József (2019) Kézírásfelismerés Arany János levelein. In: Networkshop 2019. HUNGARNET Egyesület, Budapest, pp. 38-44. A PIM (majd OSZK) DBK projektjéről: Szűcs, Kata Ágnes (2021) Automatikus kézírás-felismertetés Kiss József levelezésén. In: Online térben az online térért : Networkshop 30: országos online konferencia. 2021. április 6-9. Eötvös Loránd Tudományegyetem. HUNGARNET Egyesület, Budapest, pp. 73-80.

6 L. ehhez a jelen konferencián: Lévai Dániel, Szekrényes István, Palkó Gábor: *Kézírásfelismerés saját modellek létrehozásával a Digitális Örökség Nemzeti Laboratórium szuperszámítógépén*

eszközök, szabványok és jó gyakorlatok sora jött létre, és ezek jelentős része ingyenes, nyílt eszköz.

Maga a probléma nem ismeretlen a hazai közgyűjteményi diskurzusban. Bár a levéltárak a tömegesen keletkező digitális iratok megőrzésére fejlett technológiákkal rendelkeznek, az egyedi dokumentumok és az adatkészletek feldolgozásában rosszabb a helyzet.

2017-ben az ELTE és a PIM közreműködésével szerveztem workshopot és konferenciát neves külföldi vendégelőadók részvételével, majd a PIM Digitális Bölcsészet Központ tavaly novemberben műhelykonferenciát tartott a témában. Igen örvendetes, hogy az OSZK-ba átkerült Központ – folytatva a fél évtizede megkezdett munkát – az e-mailek archiválásáról és kutathatóvá tételéről szervezett tutoriált.[7]

A Digitális Örökség Nemzeti Laboratórium keretei között a Magyar Nemzeti Levéltár, a Miskolci Egyetem, a BTK Irodalomtudományi Intézet és az ELTE közös projektet indított, amelynek célja a digitálisan született kutatási anyagok archiválása és közzététele a nyílt tudományosság jegyében. A holnapi napon elhangzó előadás az egyik pilot projektet ismerteti, melynek keretei között olyan, a digitális bölcsészet legveszélyeztetettebb anyagai között számontartható adatbázisok szakszerű kezelésének problémáit és nemzetközi jó gyakorlatait mutatja be, mint a kutatási célú adatkészleteket tároló relációs adatbázis (SQL).[8] Újabb pilot keretei között pedig egy megszűnt határon túli magyar nyelvű hírportál, a manna.ro webportál szakszerű archiválását és kutathatóvá tételét teszteljük.

És ezzel át is térnék előadásom rövidebb második részére, amely a téma szempontjából releváns intézményi fejlemények közül emel ki néhányat.

Open Science, Open Data

Magyarországon a kutatási adatok kezelésével kapcsolatos diskurzus, a nyílt tudományosság jegyében az elmúlt években rendkívüli módon megélénkült.

Megalakult a kutatási adatok kezelését támogató Research Data Alliance (RDA) globális szervezet magyar csoportja, a Research Data Alliance Hungarian National Node (HRDA). Alapító tagjai a Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI), az MTA Könyvtár és Információs Központ (MTA KIK), a HUNgarian Open Repositories (HUNOR) és a Kormányzati Informatikai Fejlesztési Ügynökség (KIFÜ) - és nemrégiben a DH-LAB is csatlakozott a szervezethez. A KIFÜ pedig az EOSC – Európai Nyílt Tudományos Együttműködésben képviseli Magyarországot. Meetupok és workshopok során hallhattunk a kutatási adatok összefüggésében a FAIR alapelvekről, az Open Science fontosságáról és lehetőségeiről, illetve az adatrepozitóriumok jellemzőiről. (A napokban érkezett a meghívó a Nyílt tudományos fórum újabb, már negyedik workshopjáról!)

Egyre több intézmény alkalmaz vagy képez ki data steward-ot (adatgazdászt), és mind több kutatóhelyen merült fel egy nem pusztán publikációk, de különnemű kutatási adatelemek kezelésére is alkalmas – akár a szemantikus hálózatok kezelését is lehetővé tevő – intézményi vagy intézményközi adatrepozitórium létrehozásának terve. Ebbe a folyamatba a Digitális Örökség Nemzeti Laboratórium több szinten is bekapcsolódik.

Még a kutatási infrastruktúrák kiépítésénél is fontosabb talán a kutatók és az archívumi szakemberek képzése, ha a nyílt tudományosság jegyében a kutatási adatok felelős és

7 L. a PIM workshopjának programját itt: https://pim.hu/hu/esemenyek/digitalisan-letrejott-born-digital-keziratok-kezelese

8 Dióssy Anna Laura, Alföldi István: Digitálisan született kutatási anyagok megőrzése: a relációs adatbázis mint born-digital objektum

intelligens kezelésének széles körű előmozdítása a célunk. A DH-LAB ezért az ELTE-n 2022-ben adatgazdász képzést indít.

A szakképzés célja olyan szakemberek képzése, akik a könyvtártudomány, a digitális bölcsészet és a könyvtári informatika határmezsgyéjén a digitalizált vagy eleve digitálisan keletkezett adatok kezelési technikáinak magas szintű elsajátítását követően képesek ezek egyéni, illetve csoportmunkában történő, a jogi keretek adta mozgástéren belüli, akár projektszerű kezelésére. Mindebbe beleértendő az adatok biztonságának és hitelességének megőrzése, a „fair use" körülményeinek megteremtése, valamint a kezelt adatok másodlagos felhasználási célokra történő kijelölése/felhasználása, az erre alkalmas rendszerek adatokkal történő feltöltése, üzemeltetése.

Szemantikus technológiák szolgáltatássá fejlesztése

A fent vázolt képzés elengedhetetlen, hogy a kutatók és a memória-intézmények szakemberei tudják, a XXI. században hogyan kell és hogyan lehet az adatokkal bánni, de mindehhez természetesen szükség van dedikált infrastruktúrákra. Számos adatrepozitórium-fejlesztési projekt indult el az elmúlt években. A SZTAKI és az ELKH például a Dataverse szoftverre építi a Concorda repozitóriumot, míg számos helyen a szemantikus web elveinek megfelelő Invenio RDM-et tesztelik. A KIFÜ és a DH-LAB mellett az ELTE-n is indul egy adatrepozitórium pilot. Az adatrepozitóriumi funkciókat messze meghaladja az a fejlesztés, amely elnyerte az NKFIH-ITI megtisztelő TOP50 kutatási infrastruktúra elismerést.

A Laboratórium a Monguz Információtechnológiai Kft.-vel együttműködve egy olyan hardver- és szoftver-infrastruktúrát fejleszt, amely lehetővé teszi a nemzeti kulturális örökség mesterséges intelligencia alapú feldolgozását, kutatását, oktatását és közzétételét saját fejlesztésű, magyar nyelvre optimalizált nyelvfeldolgozó alkalmazások segítségével.[9] A Nemzeti Kutatási és Fejlesztési és Innovációs Hivatal ezt az úttörő fejlesztést díjazta.

Egyetlen mondatban összefoglalva a felsoroltakat: a mesterséges intelligencia mint kultúrtechnika nemcsak a kulturális örökség, de a kutatási adatok kezelésének minden szintjén forradalmi változások kezdeményezője, ám az ezekben rejlő potenciált csak a fent vázolt intézményi együttműködések formájában leszünk képesek kihasználni.

9 L. a konferencián: Kiss Tamás-Palkó Gábor: *Magyarország Ígéretes Kutatási Infrastruktúrája: A Digitális Örökség Nemzeti Laboratórium (DH-LAB) és a Qulto Kutatási Infrastruktúra (Qulto KI)*