Mesterséges intelligencia és kulturális örökség

Palkó Gábor, Indig Balázs

Digitális Örökség Nemzeti Laboratórium
ELTE BTK TI Digitális Bölcsészet Tanszék

2021. november 11.

A laboratórium célja

- *N*agy tömegű, *magyar szöveget* is tartalmazó anyag
  - sajtóanyagok
  - médiatermékek
  - web 2.0-es források (blog, fórum, chat, stb.)
  - határon innen és túl
- bármilyen jellegű kutatás, illetve oktatás számára
  - bölcsészeti
  - társadalomtudományi
  - piaci
- elérhető, értelmezhető legyen
  - széles körben
  - szemantikus mélységben

A laboratórium célja (folyt.)

- A rendelkezésre álló szolgáltatásokat, algoritmusokat fejleszteni kell
  - nagyobb és jobb tanítóanyaggal
  - magasabb színvonalú programozási architektúrával
- jó minőségű be- és kimenetek előállításával
  - szabványos
  - nemzetközi projektekben is használható

Probléma

- A magyar nyelvi források feldolgozása senkinek nem érdeke
  - Sok kicsi, diverz forrás fog eltűnni hamarosan (a méretgazdálkodás miatt)
  - Az eredmények rövid távon nem hasznosíthatóak
- Ez még a jelenleginél is nagyobb versenyhátrányt jelent
  - A nemzetközi módszerek nem vehetők át kritika nélkül (a *one-size-fits-all* nem mindig áll)
  - Ha nincs anyag, nincs min szakembert képezni
- A magyar, különösen a határon túli digitális örökség összefüggésében
- A magyar nyelv (és szorosan utána a többi kutatás) ellehetetlenül

AI to the rescue!

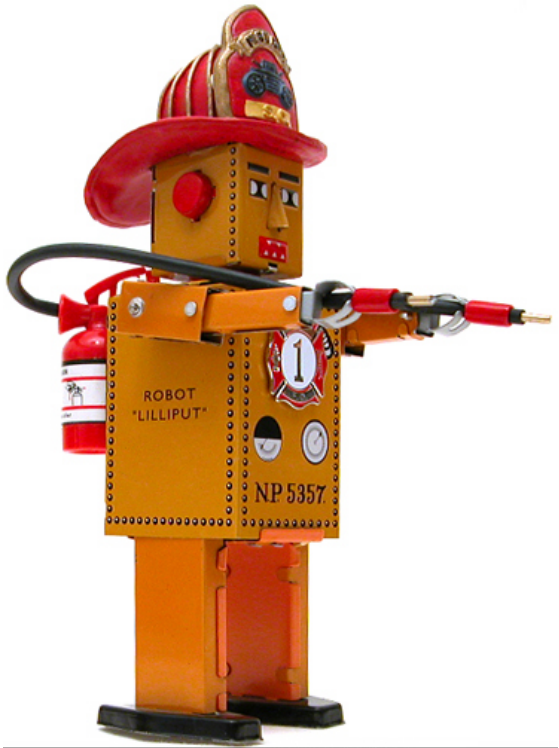

Photo copyright © ToyTent

„Ha egy probléma nincs megoldva, akkor ahhoz mesterséges intelligencia kell, ha meg van oldva, akkor azt csak egy okos algoritmus hajtja.”

A munkamenet

1. Felkutatjuk a memóriaintézményeket, feltárjuk a releváns forrásokat
2. Jogilag tisztázzuk a felhasználhatóságukat, megállapodásokat kötünk
3. Jó minőségben digitalizáljuk őket
   - Az eleve digitális forrásoknál (*born digital*) ez nem szükséges, de a többi lépés azonos
4. Számítógéppel és ahol kell, mesterséges intelligenciával feldolgozzuk
5. Szabványos, a hosszútávú megőrzést elősegítő formátumba (TEI XML) konvertáljuk
   - Szabványos metaadatokkal (Schema.org)
   - Reprodukálhatóan
   - A teljesség igényével
6. Szemantikus adatgazdagítással kereshetővé tesszük az anyagot
7. Repozitáljuk a létrejött tartalmakat állandó azonosítókkal ellátva (pl. DOI, Handle)
8. Elérhetővé tesszük az anyagokat szolgáltatások formájában

Létrejött haszon

- Az általános digitalizáció elősegítése
- A könyvtárak új funkciót kapnak: a kutatási adatok kezelése, rendszerezése
- A kutatók az új anyagokkal dolgozhatnak (a hangsúly áttevődik az MI-re)
  - Ellenőrizhetik a meglévő eredményeket
  - Azokra építve (!) új, ellenőrizhető eredményeket hozhatnak létre
- A közemberek számára pontosabb hozzáférés eddig elérhetetlen anyagokhoz
- Az ipar számára létfontosságú összehasonlítások, modellek készíthetők
  - Későbbi kutatások és ipari projektek hátországa
- Egyetemi képzés megújítása, kiművelt emberfők képzése gyakorlati projektek által
- A nemzetközileg alkalmazható mintaprojekt, monetizálható know-how

Főbb projektek (korpuszok)

- Folyamatosan bővülő vers-, regény-, dráma- és online hírportálkorpusz keresővel
  - Kereső: `https://verskorpusz.elte-dh.hu/`
    - Ismertető: `https://elte-dh.hu/verskorpusz/`
    - 13 063 db vers több mint 40 szerzőtől (Horváth 2020)
  - Kereső: `https://regenykorpusz.elte-dh.hu/`
    - Ismertető: `https://elte-dh.hu/regenykorpusz/`
    - 100 regény, 80 szerzőtől 7 000 000 szó terjedelemben
  - Kereső: `https://dramakorpusz.elte-dh.hu/`
    - Ismertető: `https://elte-dh.hu/dramakorpusz/`
    - Jelenleg 40 dráma, hamarosan a `https://dracor.org/` -on
  - Kereső: https://cikk-kereso.elte-dh.hu/
    - Ismertető: `https://elte-dh.hu/cikk-kereso/`
    - Jelenleg 1 303 864 cikk (Indig, Knap és tsai. 2020)
- Ezek számtalan újszerű bölcsészeti kutatás katalizátorai

Főbb projektek (nyelvtechnológia/nyelvészet)

- A korpuszok feldolgozása jó minőségű NLP eszközlánccal (Indig, Sass, Simon és tsai. 2019)
  - Klasszikus és mélytanulásos módszerek ötvözésével (Simon és tsai. 2020)
  - Szabványos be-kimenettel, API-val (Indig, Sass és Mittelholcz 2020)
  - Az ígéretes egyedi modulok is megvizsgálhatók
- Szabadon elérhető referenciakorpusz fejlesztése
  - szófaji egyértelműsítés, szótövesítés
  - szentiment elemzés
  - névelemfelismerés
  - stb.
- A belépési küszöb csökkentése
  - Egy szót automatikusan morfémákra bontani még nem volt ilyen könnyű! `http://emmorph.herokuapp.com/`
  - Publikus korpuszkereső a *Magyar Nemzeti Szövegtár* (Oravecz, Váradi és Sass 2014) után elsőként Magyarországon `https://sketchengine.elte-dh.hu/`

Kutatási kérdések

„Ismeritek azt a gyötrő és butító állapotot, mikor az ember kínlódva elismétel magában egy szót, mondjuk »leves«, és a végén már nem tudja, miről van szó, és csak ezt tudja: »leves«, és nem lát semmit, és végre már nem biztos abban se, hogy mondják: »leves« vagy »levés«.”

(Karinthy Frigyes)

- Visszatérő kérdések:
  - Ha a szerkezet elemezhető, akkor pontos a modellünk, de nem generalizálunk túl?
  - Meg kell nézni az adat sötét oldalát is!
  - A korpuszlekérdező pontosságot, míg a vektorterek fedést mutatnak
  - Pipeline hatás (0,8 x 0,8 = 0,64 $\rightarrow$ Az egyharmad rossz!)

Széleskörű együttműködések az MI jegyében

- SOTE: Covid és kapcsolatos terminológiák/neologizmusok
- MOME: A modern és a klasszikus építészet internetes megítélése
- OSZK MIA: A webaratás finomítása filológiai szempontok mentén
- ELTE:
  - BTK: A személyjelölés konstrukcióinak korpuszalapú, kognitív poétikai vizsgálata (OTKA)
  - IK: Webarchívumok kiberbiztonsági kérdései (Lendák, Indig és Palkó 2021)
  - TáTK: Depresszió és gyűlöletbeszéd vizsgálata az online szövegekben (OTKA)
- stb.

Köszönöm a figyelmet!

https://dh-lab.hu/
https://elte-dh.hu/
https://github.com/elte-dh
https://zenodo.org/communities/elte-dh

Hivatkozások I

- Horváth Péter, „Az ELTE Verskorpusz automatikus annotációs eljárásai révén nyerhető kvantitatív adattípusok”, *Nyelvtan, diskurzus, megismerés*, 2020, 313–332. old.
- Indig Balázs, Árpád Knap és tsai., „The ELTE.DH Pilot Corpus – Creating a Handcrafted Gigaword Web Corpus with Metadata”, English, *Proceedings of the 12th Web as Corpus Workshop*, Marseille, France: European Language Resources Association, 2020. máj., 33–41. old., isbn: 979-10-95546-68-9, url: https://www.aclweb.org/anthology/2020.wac-1.5.
- Indig Balázs, Bálint Sass és Iván Mittelholcz, „The xtsv Framework and the Twelve Virtues of Pipelines”, English, *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 2020. máj., 7044–7052. old., isbn: 979-10-95546-34-4, url: https://aclanthology.org/2020.lrec-1.871.

Hivatkozások II

- Indig Balázs, Bálint Sass, Eszter Simon és tsai., „One format to rule them all – The emtsv pipeline for Hungarian”, *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Italy: Association for Computational Linguistics, 2019. aug., 155–165. old., doi: 10.18653/v1/W19-4018, url: https://aclanthology.org/W19-4018.
- Lendák Imre, Balázs Indig és Gábor Palkó, „WARChain: Blockchain-Based Validation of Web Archives”, *Socio-Technical Aspects in Security and Trust*, 2021, 121–134. old., doi: 10.1007/978-3-030-79318-0_7.
- Oravecz Csaba, Tamás Váradi és Bálint Sass, „The Hungarian Gigaword Corpus”, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. máj., 1719–1723. old., url: http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf.
- Simon Eszter és tsai., „Újabb fejlemények az e-magyar háza táján”, *XVI. Magyar Számítógépes Nyelvészeti Konferencia*, 2020, 29–42. old.