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.2022.45

Szemantikus névelem-azonosítás magyar nyelvű szövegeken (a HuWikifier bemutatása)

Nemeskey Dávid Márk – Palkó Gábor
Eötvös Loránd Tudományegyetem Bölcsészettudományi Kar Történeti Intézet, Digitális Bölcsészet Tanszék, Digitális Örökség Nemzeti Laboratórium
nemeskey.david@btk.elte.hu, palko.gabor@btk.elte.hu

Bevezetés

A közgyűjteményi és webes forrású anyagok szemantikus címkézését céljául kitűző fejlesztés a Digitális Örökség Nemzeti Laboratórium keretei között valósul meg. A Nemzeti Laboratórium program az ITM kezdeményezésére 2020-ban indult, jelenleg összesen 18 Laboratórium működik párhuzamosan. Érdemes kiemelni, hogy bölcsész- és társadalomtudományi területen csak két laboratórium tevékenykedik. A társadalmi innovációt kutató és előmozdító TinLab szintén az ELTE vezetésével kezdte meg működését, akár a konzorciumi formában létrejött Digitális Örökség Nemzeti Laboratórium. Az összes többi laboratórium a sokkal könnyebben piacosítható „Hard Science" területén jött létre. A laboratóriumok ötéves futamidővel működnek, többnyire konzorciumi formában. A DH-LAB együttműködő partnere a Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézete, melynek keretei között a DigiPhil projekt működik, és amelynek technikai megújítása[1] a DH-LAB támogatásával valósul meg. A DH-LAB további partnere a Magyar Nemzeti Levéltár, amely elsősorban a gépi kézírásfelismeréssel kapcsolatos tevékenységekben, illetve a digitálisan született anyagok kezelése (born digital curation) területén aktív, ahogyan további konzorciumi partnerünk, a Miskolci Egyetem is.

A digitális örökség Nemzeti Laboratórium létrehozásának fő célja az volt, hogy kidolgozza a nemzeti kulturális örökség mesterséges intelligencia alapú feldolgozásának, kutatásának és oktatásának, valamint a lehető legszélesebb körű közzétételének a módszertanát, és mindezt úgy, hogy a kifejlesztett módszertanok és eszközök, valamint a szakértői kompetencia piaci hasznosítására is gyakorlatokat alakítson ki.

Ez a konferencián bemutatott mindkét eszközre, a szemantikus címkézőre és a gépi kézírásfelismerőre egyaránt érvényes.

A digitális kulturális örökség kutathatóvá tételével szélesebb felhasználói kör számára nyújthatunk célzottabb intelligens hozzáférést, méghozzá szemantikus mélységben. A digitalizálás folyamatának gyorsítása és javítása MI eszközök révén, valamint a tömeges digitalizálás eredményeinek gépi tárgyszavazása, a kéziratok automatikus felismertetése többszörösére emelheti a széles körben, intelligens módon felhasználható kulturális tartalmak mennyiségét.

De térjünk át az előadás voltaképpeni tárgyára, a szemantikus címkézőre. Adódik a kérdés: tulajdonképpen miért építünk tematikus címkéző eszközt? Amellett, hogy a digitális örökség minden területén hasznos, a DH-LAB három fejlesztési iránya is épít az ezen eszköz nyújtotta lehetőségekre.

Az első tevékenység, amit kiemelnék, a webaratás alprojekt keretei között folyik.[2] A fejlesztés középpontjában egy nyílt hozzáférésű szoftver, a Web Article Curator[3] áll. Ennek a szoftvernek

1 Lásd jelen kötetben

2 https://keptar.oszk.hu/kereses/reszletes.phtml?id=78376

3 The ELTE.DH Pilot Corpus – Creating a Handcrafted Gigaword Web Corpus with Metadata Balázs Indig, Árpád Knap, Zsófia Sárközi-Lindner, Mária Timári, Gábor Palkó In the Proceedings of the 12th Web as Corpus Workshop (WAC XII), pages 33-41 Marseille, France 2020, https://doi.org/10.5281/zenodo.3755323

a segítségével webes forrású cikkek aratása történik, a weboldalak, elsősorban hírportálok cikkei részletes és szabványos formátumú metaadatokkal együtt kerülnek be a DH-LAB fejlesztette keresőszolgáltatásokba. Ezek jelenleg mintegy 3 millió cikket tartalmaznak. Mivel ezek többnyire szerzői jogvédett tartalmak, ezért a szolgáltatásban a nagyközönség csak a metaadatok között kereshet. Kutatási együttműködés keretében ugyanakkor a webes forrású cikk-korpuszt bármely érdeklődő kutatóval megosztjuk.

A célunk az, hogy ezeket a cikkeket a szemantikus címkéző segítségével a teljes szövegű szabad szavas keresésénél mélyebben tegyük kereshetővé.

Egy másik projektünk, ahol a szemantikus indexelésnek nagy a jelentősége, DH-LAB készülő repozitóriumi metakeresője, amely a magyarországi repozitóriumok dokumentumainak metaadataiban és azok szövegében egyszerre képes kereséseket végrehajtani. A SZTAKI és az MTA KIK közös keresője hasonló céllal jött létre, ám jelenleg csak a metaadatokat gyűjti össze, a dokumentumok szövegét nem kezeli.

Célunk az, hogy a szolgáltatás a dokumentumok teljes szövegében is lehetővé tegye a keresést, méghozzá szemantikus módon is. Ennek egyik fő akadálya jelenleg, hogy a gyűjteményi anyagok a repozitóriumokban többnyire változó minőségű PDF formátumú fájlokban vannak jelen. Célunk, hogy ezekből a dokumentumokból gépi feldolgozásra alkalmas formátumban nyerjük ki az adatokat, lehetőleg szabványos XML formátumban. Ennek különféle módszerei lehetségesek, azon dolgozunk, hogy a szolgáltatás automatikusan a legmegfelelőbb munkamenetet válassza ki a lehetséges eszközláncok közül, hogy a gépi feldolgozás számára a legalkalmasabbat válasszuk ki.

De miért is van szükség arra, hogy dokumentumokat wikifikáljunk vagy más módon szemantikus címkével lássuk el a szövegek elemeit?

Idézzünk fel egy példát. A képernyőképen az Ontotext cég által létrehozott szolgáltatás egy oldala látható. Az Ontotex hírportálok szövegét gyűjti össze és azokat szemantikusan címkézi. A képen egy a digitális kulturális örökség szempontjából releváns, a mesterséges intelligencia és a múzeumok kapcsolatáról szóló egészen friss cikk látható. Az Ontotext rendszere nem pusztán a cikkek szövegében keres, de azok elemeit különféle kategóriákba sorolt címkékkel látja el (pl. személyek, intézmények, földrajzi helyek, témák), és az ezek alkotta szemantikus hálózatot is bejárhatjuk.

Amikor Tim Berners-Lee a szemantikus web fogalmát bevezette, éppen ilyen szolgáltatás lebeghetett a szeme előtt. Olyan eszközök létrehozását sürgette, amelyek kizárják a többértelműségből származó tévedéseket (példájában a Carlsberg karaktersor szerepelt, ami egyszerre jelölhet egy sörmárkát, egy futballcsapatot és egy földrajzi entitást).

A szemantikus címkézés célja pontosan ez: entitások és entitástípusok azonosítása különféle szövegbeli megfogalmazásokat összekötve (pl. Budapest, Magyarország fővárosa), illetve azonos szövegbeli megfogalmazások értelmi differenciálása (pl. arany mint anyagnév és Arany [János] mint személy).

Az első olyan projekt, amely a Huwikifier szemantikus címkéit kutatási célú mintázatfelismerésre használja, helyneveket azonosít. A kísérlet során Kiss Tamás digitális bölcsész szakértővel együttműködve a COST Action *Distant reading for European Literary History* című projekt keretei között az ELTE Digitális Bölcsészet tanszékén készült regénykorpusz regényeit elemeztük. A Huwikifier segítségével mintegy 14 ezer helyszínt azonosítottunk a 100 magyar nyelvű regényben, majd a Notegoat szolgáltatás segítségével térképes vizualizáció formájában jelenítettük meg az eredményeket. Egy ilyen elemzés képes lehet kulturális és/vagy irodalomtörténeti trendek felismerésére, illetve ilyen érvelések megerősítésére (vagy éppen cáfolatára).

Implementáció

Az eddigiekben egy szélesebb áttekintést adtunk a feladatról és a környezetről, amiben a rendszert használjuk. Ebben a fejezetben a „színfalak mögé lépünk” és egy részletesebb képet nyújtunk a Huwikifierről és a wikifikáció folyamatáról.

Wikifikáció

Definíciószerűleg a wikifikálás a szemantikus annotáció egy olyan változata, ahol a szöveg fontosabb szavait a Wikipédia[4] entitásaival (oldalai URL-jével) címkézzük fel. Ennek a hagyományos – egyedi ontológián alapuló – szemantikus annotációval szemben számos előnye van. Egyrészt lehetővé teszi az annotációt saját ontológia készítése nélkül, hiszen adott egy nagy, közösségileg épített, több nyelven is elérhető entitás-adatbázis. A Wikipédia-entitások ráadásul bárki számára ismerős és érthető fogalmakat takarnak. Másodrészt az entitásokhoz tartozó cikkek, illetve az oldalak közötti hivatkozások lehetőséget biztosítanak az entitások automatikus felismerésére.

Természetesen a Wikipédiának hátrányai is vannak: a cikkek minősége hullámzó lehet, illetve a mennyiségük is változó az egyes nyelveken belül. A cikk írásának idején a legnagyobb Wikipédia, az angol körülbelül 6,5 millió cikket tartalmaz, míg a magyar kicsit többet, mint fél milliót. Ez utóbbi jelentősen korlátozza egy, a magyar Wikipédiára épülő rendszer felidézését (*recall*).

Algoritmusok

Az itt használt „wikifikáció” szó az első rendszer, a *Wikify!* nevéből származik (Mihalcea és Csomai, 2007). Az évek során több hasonló algoritmust is kidolgoztak, amik mind az alábbi három lépést valósítják meg különböző módszerekkel:

1. Entitás-jelöltek felkutatása a wikifikálandó szövegben.
2. A jelöltek azonosítása Wikipédia-entitásokkal. Mivel egy jelölt több különböző entitásra is utalhat (pl. *Washington* város, állam, vagy személy?).
3. Az entitások szűrése aszerint, hogy mennyire relevánsak a szöveg tematikája szempontjából.

A Wikify! kulcsszókinyeréssel állítja fel a jelölteket, majd a jelölt környezete és a célentitás(ok) cikkei közötti szöveges hasonlóság alapján egyértelműsít. Milne és Witten egyrészt a szöveges helyett egy, a bejövő hivatkozások halmazhasonlóságán alapuló *szemantikus hasonlóság (semantic relatedness, SR)* metrikát vezet be (2008a), másrészt gépi tanuló algoritmusokat alkalmaz a jelöltek megtalálására és egyértelműsítésére (2008b).

Hoffart és társai (2011) alkalmaznak először gráf alapú hasonlósági módszert jelöltazonosításra. A jelöltek és a hozzájuk csatolható entitások egy *jelölt-entitás gráfot* feszítenek ki. Minden jelöltnek és lehetséges entitásnak egy-egy csomópont felel meg. Egy jelölt és egy entitás csomópont között akkor húzódik él, ha a jelölt utalhat az adott entitásra; két entitás között pedig akkor, ha azok valamilyen értelemben hasonlóak. Az entitások hasonlósági metrikája algoritmusfüggő, és az élek a metrika értékével súlyozottak.

Hoffart és társai módszere ebben a gráfban keres olyan sűrű algráfokat, amik minden jelölthöz egy entitást rendelnek. Brank és társai (2017) ezt az ötletet vitték tovább Wikifier[5] nevű rendszerükben, és a Google PageRank (Brin és Page, 1998) algoritmusával azonosítják

4 https://www.wikipedia.org/

5 https://wikifier.org

a legfontosabb entitásokat. Ehhez a személyre szabott (*personalized*) PageRank algoritmust használják, ahol a sztochasztikus szörföző a jelölt csomópontokból indul, és legmagasabb PageRankkel rendelkező entitásokkal annotál.

Megfigyelhető, hogy a legtöbb wikifikációs módszer a fenti három lépésből főleg az elsőre és a másodikra koncentrál; a relevanciaszűrés kevésbé kutatott terület. A Wikify! A kulcsszavak, míg a Wikifier a PageRank pontszámok alapján sorrendezi az entitásokat, és egy bizonyos küszöb alattiakat eldobja. A küszöb azonban mindkét esetben az algoritmus hiperparamétere, aminek helyes megválasztása jelentősen befolyásolja a rendszer teljesítményét.

Huwikifier

A mi rendszerünk – ahogy neve is utal rá – a Wikifier újraimplementálásán alapul. Annak ellenére döntöttünk az újraimplementálás mellett, hogy a Wikifier algoritmusa nyelvfüggetlen, és az oldalon (illetve az API-n) keresztül magyarul is használható. Az egyik indok épp a nyelvfüggetlenség: a Wikifier nem veszi figyelembe a magyar nyelv jellegzetességeit, ami jelentősen rontja az eredményt. A másik ok, hogy olyan funkciókat is implementálni tudjunk, amik nem szerepelnek az eredeti, zárt rendszerben.

Jelen (béta) állapotában a Huwikifier az eredeti Wikifier algoritmusának minden lépését megvalósítja. A rendszernek két végpontja van: egy REST API, ami a Wikifier API nekünk szükséges részhalmazát támogatja, és egy HTML tesztinterfész, ami egy megadott dokumentumon lefuttatja az annotációt, és különféle tesztelési információkat jelenít meg. A Wikifierrel való API kompatibilitás fontos szempont volt, hogy az azt használó szemantikus keresőt könnyen, a végpont átírásával egyszerűen át tudjuk állítani a Huwikifier használatára.

A tesztüzem alatt a rendszer nem elérhető a nyilvánosság számára, de a kész rendszert nyilvánosságra tervezzük hozni.

Wikifikálás magyarul

Az eredeti Wikifiert magyar mondatokkal tesztelve, két fő hibaosztályt ismertünk fel. Ezek közül az első a morfológia figyelmen kívül hagyása. A Wikifier a jelöltek megtalálásához a szövegben az egyes entitások valamilyen ismert, Wikipédiában látott említését keresi. A keresés azonban szigorúan a felszíni formára korlátozódik. Ez a módszer jól működik angolra, ahol egy lemmához kevés lehetséges felszíni alak tartozik; azonban egy, a magyarhoz hasonlóan gazdag morfológiájú nyelvben az *adatritkaság* problémákat okozhat. Egy valós példa erre Székesfehérvár, aminek említései között előfordulnak a „*Fehérvár*”, „*Fehérvárról*” alakok, de pl. a „*Fehérvárnál*” nem, ezért ha az annotálandó szöveg ezt tartalmazza, az algoritmus nem ismeri fel azt jelöltként.

A másik hiba a szófajok figyelmen kívül hagyása. A „*Tüdejét is kiköpi, ahogy szív egy kis levegőt.*” mondatban a Wikifier mind a „*tüdő”-t*, mint a „*szív”*-et annotálja a megfelelő testrészek oldalaival, függetlenül attól, hogy az utóbbi ebben mondatban egy ige.

A Huwikifier ezeket a problémákat az emtsv[6] (Indig és társai, 2019) segítségével oldja meg. Egyrészt mind a szöveget, mind az említéseket lemmatizálja, így az első példában mind a *„Fehérvárról”*, mind a *„Fehérvárnál” „Fehérvár”*-ra egyszerűsödik. Ez nemcsak abban segít, hogy az adatritkaságot megszüntesse, hanem az említésadatbázis méretét is jelentősen lecsökkenti, hiszen elég a lemmákat eltárolni a felszíni alakok helyett. Másrészt az emtsv a szöveg minden szavához visszaadja a szófajt is, amit felhasználunk a jelöltek szűréséhez: minden jelölt, ami nem főnévre végződik (vagyis nem főnévi csoport) törlésre kerül.

6 https://github.com/nytud/emtsv

A Huwikifier az emtsv-t REST API-n keresztül éri. A Huwikifier teljesítményét szemlélteti, hogy az emtsv felel egy átlagos online média cikk feldolgozási idejének 2/3–3/4-ért.

Új funkciók

A Huwikifiert, bár még béta állapotban van, elkezdtük bővíteni az eredeti rendszeren túlmutató funkciókkal. Ezek egyike a TEI XML[7] kezelése.

A digitális bölcsészet nagy hangsúlyt fektet a dokumentumokhoz tartozó metaadatokra, emiatt nyers szöveg helyett a repozitóriumok tipikusan TEI XML formátumú dokumentumokat tartalmaznak. Ennek megfelelően e formátum támogatása alapvető cél volt. A Huwikifier képes fogadni TEI XML formátumú bementet az API-n keresztül, és az annotációkat a szavak XML ID-jaival indexelni. Ezt utána a kliens mint külön *standoff annotation* mezőt adhatja hozzá a dokumentumhoz (hasonlóan egyéb, nem tokenszintű annotációkkal, mint pl. névelemek).

Mivel a TEI XML-eink már eleve tartalmazzák a legalapvetőbb nyelvi annotációkat (lemma, szófaj), ezért ebben az esetben a Huwikifier maga már nem hívja meg az emtsv-t, így egy TEI XML dokumentum feldolgozása gyorsabb, mint a nyers szövegé lenne.

A másik új funkciót a fenti COST Action projekt inspirálta. A Huwikifierbe felvettük a WikiData[8] ontológia IS-A kapcsolatait. Amennyiben ez az adat elérhető, a rendszer nem csak az egyes entitások WikiData azonosítóját adja vissza, hanem azok közvetlen taxonómiai szülőkategóriáit is[9]. Ezen kívül a kliens kikötheti, hogy csak bizonyos kategóriába eső entitásokra kíváncsi; ez esetben a többi entitást töröljük a kimenetből. Bár a fenti kutatás idején ez a funkció még nem volt elérhető, a további, hasonló projektek végrehajtását jelentősen meg fogja könnyíteni.

Terveink közé tartozik ennek a funkciónak a bővítése az IS-A-n kívüli, tetszőleges relációk támogatásával.

Távlati tervek

A Huwikifier már jelenlegi állapotában is használható, és alkalmazható digitális bölcsészeti kutatások támogatására. Mielőtt azonban késznek nyilváníthatnánk, fontos, hogy az algoritmus robusztusabb legyen: kevesebb irreleváns vagy téves entitást annotáljon. Jelenleg ennek tesztelése és a talált hibák javítása folyik.

Hosszabb távon a rendszert kétféleképp tervezzük továbbfejleszteni. Ahogy láttuk, a különféle wikifikáló algoritmusok más és más módszerekkel oldották meg a főbb lépéseket. Tervezzük a Huwikifier algoritmusát más rendszerekben alkalmazott megoldásokkal, úgymint kulcsszó-kereséssel és gépi tanulási módszerekkel kiegészíteni. Természetesen utóbbiak ma neurális, ún. mélytanulásos modellek lennének. Ehhez szükségünk lehet megfelelő tanító és tesztadat létrehozására is.

A másik irány a Wikipédián kívüli, további tudásbázisok bevonása. Olyan adatokra gondolunk itt, mint más wiki oldalak, enciklopédiák, ember- (ki-kicsoda) vagy cégnyilvántartások, stb. Mivel ezekben nincs feltétlenül meg a Wikipédia interkonnektivitása, vagy az entitások szöveges körülírása, automatikus annotációra való alkalmazhatóságuk külön kutatás tárgya lesz.

---

7 https://tei-c.org/

8 https://www.wikidata.org/

9 Hasonló funkció létezik az eredeti Wikifierben is, de az DBPedia kategóriákat ad vissza.

Bibliográfia

Brank, Janez, Gregor Leban, and Marko Grobelnik. „Annotating documents with relevant wikipedia concepts." *Proceedings of SiKDD* 472 (2017).

Sergey Brin and Larry Page (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems. 30 (1–7): 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, et al. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK.. Association for Computational Linguistics.

Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai: One format to rule them all – The emtsv pipeline for Hungarian. In: Proceedings of the 13th Linguistic Annotation Workshop. Association for Computational Linguistics, 2019, 155-165. https://doi.org/10.18653/v1/W19-4018

David Milne and Ian H. Witten. 2008a. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links (AAAI 2008). Association for the Advancement of Artificial Intelligence. Chicago, IL, USA, 25-30.

David Milne and Ian H. Witten. 2008a. Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management (CIKM ,08). Association for Computing Machinery, New York, NY, USA, 509–518. https://doi.org/10.1145/1458082.1458150

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM ,07). Association for Computing Machinery, New York, NY, USA, 233–242. https://doi.org/10.1145/1321440.1321475