Publikációgyűjtemény tudásbázisának építése természetes nyelven

Hornyák Zsuzsanna, Mészáros Tamás[1]

Bevezetés

A publikációk és tudományos eredmények elektronikus formában hatalmas mennyiségben hozzáférhetőek lettek az elmúlt évtizedekben. A fejlődő információs technológiáknak köszönhetően a publikációkat összegyűjtő digitális könyvtárak azonnali és könnyű hozzáférést biztosítanak a keresett adatokhoz. Ebben az új, digitális tudományos társadalomban a könyvtárosok szerepe sem tűnt el, hanem átalakult. Szükség van szerkesztőkre és szakemberekre, akik a digitális könyvtárak elképesztő információ mennyiséget rendezik, katalogizálják és elérésükhöz egyre pontosabb, korszerűbb technológiákat fejlesztenek ki.

A digitális könyvtárak mögött futó informatikai rendszerek hatékony működéséhez a keresőalgoritmusoknak részletes információkra van szükségük a tárolt publikációkról. Ma már széles körben használnak különböző meta adatokat a publikációk fontosabb adatainak rögzítésére - pl. cím, szerző, témakör, publikálás éve, kulcsszavak stb. A célirányos kereséseket nagy mértékben segítik ezek az információk, de nem árulnak el eleget a publikációk tényleges tartalmáról, a bennük szereplő állításokról illetve eredményekről.

Az interneten elérhető tartalom gépek által értelmezhető formában történő tárolása a szemantikus web egyik fő célkitűzése. Ennek keretében az elmúlt évtizedben több új technológia is fejlesztésre került, amelyek segítségével a szabadszöveges dokumentumok tartalma formálisabban leírható, tudásbázisok létrehozhatóak. Ezen formalizmusok azonban lassan, vagy alig terjednek el, mivel bonyolultak és az átlagfelhasználó számára nem kézenfekvő elkészítésük. Komoly, szervezett munkának köszönhetően bizonyos kutatási területeken létrejöttek a tárgyterület fogalmait lefedő és rendszerező ontológiák[1], illetve bizonyos gyakori elemek (pl. kémiai struktúrák[2]) leírására szolgáló formalizmusok, ám ezek csak részben járulnak hozzá a publikációk szemantikus tartalmának kinyeréséhez.

Az előadásunkban egy olyan megoldást mutatunk be, melynek segítségével a publikációk szemantikus reprezentációit könnyen, természetes nyelven lehessen elkészíteni. Az így létrehozott szemantikus absztraktok célja a publikációk tartalmának összefoglalása a felhasználók számára kényelmes módon, de olyan kontrollált formában, amely alapján

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, Méréstechnika és Információs Rendszerek Tanszék

[2] http://www.xml-cml.org/

egyértelműen legenerálható a cikkek formális szemantikai reprezentációja, így létrehozva egy intelligensen kereshető tudásbázist publikációk gyűjteményeihez.

Szemantikus publikáció

A szemantikus publikáció alatt az olyan megoldásokat értjük, amelyek a publikációk szövege mellett a gépek által értelmezhető információkat is elérhetővé tesznek a digitális könyvtárakban. Ezen információk előállítása a szerzők feladata lenne, de többnyire a gyűjtemények kezelőire hárul. A legelterjedtebb megoldások a tárgyterület ontológiája vagy taxonómiája alapján kiemelik (és megjelölik) a fontos fogalmakat a szabad szövegben, azokat összekapcsolva máshol elérhető információkkal (pl. ZooKeys folyóirat[3]). Ezek különböző szövegbányászati technológiákra épülnek, amelyek eredményei nem teljesen megbízhatóak, mindig felülvizsgálásra szorulnak. Noha ezek a dinamikus, online publikációk már intelligensebben jelenítik meg a cikk adatait, a tényleges tartalmi állítások szerinti keresést még mindig nem segítik elő.

Összetettebb szemantikus állítások létrehozását célozta meg azonban a *strukturált digitális absztrakt* (SDA) kezdeményezés[2], amelynek keretében valamilyen strukturált formában összefoglalják a cikk állításait, a hagyományos absztrakthoz hasonlóan. A FEBS folyóiratban bevezetett rendszer egy egyszerű táblázat kitöltését igényelte a szerzőktől. Itt az adott struktúrának megfelelően rögzítették, milyen fehérje-interakciók szerepelnek az adott cikkükben. A táblázatot publikálás előtt a szerkesztők felülvizsgálták, majd az adatai alapján létrehozták a publikáció strukturált absztraktját, amely már gépek által értelmezhető módon elérhetővé tett néhány állítást a cikkből.

Az SDA ötlete egy komoly lépés összetettebb állítások formalizálásához, de elkészítésük körülményes és kötött struktúra nehezen kiterjeszthető. Mind a szerzőknek és a szerkesztőknek sok plusz munkát igényelt elkészítésük. A *kontrollált nyelvű szemantikus absztraktok* ötlete az SDA továbbfejlesztésének tekinthető. A cél még mindig a fontos tartalmi állítások kinyerése, de a szemantikus absztraktokban ez természetes nyelven történik, a megszokott kivonatokhoz hasonlóan, nem pedig egy hatalmas táblázat kitöltésével. Az absztraktok strukturáltságát a kontrollált nyelvek biztosítják, amelyek a szabadszöveges szemantikus absztrakt könnyű gépi feldolgozását teszik lehetővé.

Kontrollált természetes nyelv

A kontrollált természetes nyelvek a hétköznapi nyelv valamely részhalmazát alkotják. Nyelvtani és lexikális szabályok határozzák meg a kontrollált nyelvben elfogadott mondatszerkezeteket, így létrehozva a természetes nyelvű szövegeknek egy, gépek által egyértelműen értelmezhető halmazát[3]. Kontrollált nyelveket sok helyen alkalmaznak egy adott problémára specifikusan kialakítva, például felhasználói útmutatók vagy lekérdező felületek készítésekor.

[3] http://www.pensoft.net/journals/zookeys/

A szemantikus absztraktok alapját is egy-egy, tárgyterület specifikus kontrollált nyelv (illetve az azt meghatározó nyelvtan) alkotja. Ezek előnye, a táblázatos struktúrákkal szemben, hogy könnyen bővíthetőek, akár dinamikusan futási időben is. Egyszerűen adódik a témához kapcsolódó ontológiák bekapcsolása is, a lexikális szabályok segítségével lehetőség van közvetlenül az ontológia fogalmakra hivatkozni, és az ezekkel kapcsolatos információ a gépi fordításkor automatikusan rendelkezésre áll.

A kontrollált nyelvű állítások noha a felhasználóhoz közel álló természetes nyelven készítendőek el, a szabályok miatti megkötésekhez alkalmazkodniuk kell a szemantikus absztrakt íróinak. Ahhoz, hogy ez gördülékenyen történhessen, érdemes egy olyan szerkesztői felületet biztosítani a felhasználóknak, amely támpontokat nyújt szerkesztés közben. Inkrementális, prediktív nyelvtani elemzőknek hála lehetséges mondatok alkotása közben a felhasználónak folyamatosan visszajelzéseket adni arról, milyen szavakkal tudják folytatni az adott félkész szöveget. Ilyen típusú megoldások már több helyen létrejöttek ontológiák szerkesztésére (pl. GINO[4]).

Digitális könyvtárak kiterjesztése

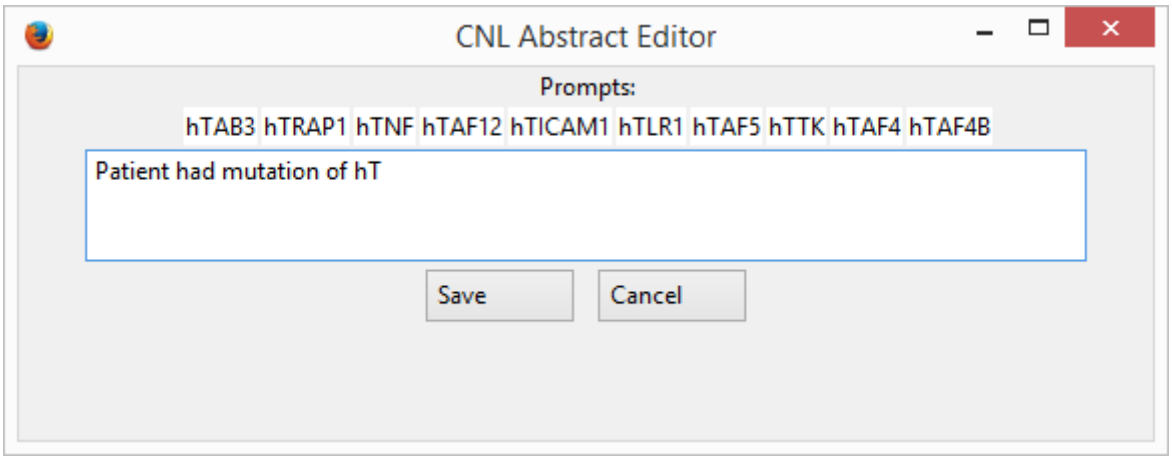
Az előbbiekben felvázoltuk a szemantikus absztraktok ötletét. Ezeknek elsődleges célja egy kézenfekvő médium biztosítása publikációk szemantikus reprezentációjának létrehozásához. A segítségükkel létrehozható egy tudásbázis, amely lehetővé teheti a publikációkkal kapcsolatos tartalmi kérdések megválaszolását, azaz a relevancia-alapú keresést kiegészíthetjük valamilyen tudáslekérdező rendszerrel. A tudásbázis felépítéséhez szükséges lépések összefoglalása:

- Adott publikációgyűjtemény tárgyterületének meghatározása, ontológiák, illetve taxonómiák felkutatása vagy készítése.
- Elkészítendő tudásbázissal szembeni elvárások felvázolása (mire irányul a tartalmi keresés).
- A tudásbázishoz szükséges tudás reprezentációs nyelv vagy formalizmus kiválasztása (pl. XML, RDF, Prolog), a gép által értelmezett információk struktúrájának létrehozása.
- Az elvárások és az ismert publikációk alapján egy kontrollált nyelvtan készítése. Ennek két fő része a beviteli állítások mondatszerkezetének definiálása, és a mondatszerkezetekhez tartozó (egyértelmű) formális reprezentáció meghatározása.
- Kényelmes beviteli eszköz biztosítása, amelynek segítségével a publikáció beküldői vagy a gyűjtemény szerkesztői elkészítik a kontrollált nyelvű szemantikus absztraktot.
- A szemantikus absztraktok gépi értelmezése, segítségükkel tudásbázis építése a publikációk tartalmáról.
- Intelligens keresés biztosítása a tudásbázis állításain a gyűjtemény publikációinak hatékonyabb eléréséhez.

Példa implementáció

A konferencia előadás során bemutatásra kerül egy prototípus implementáció, mely szemlélteti a bemutatott elképzelések egy lehetséges megvalósítását. Az elkészített szoftver a népszerű publikáció rendező alkalmazás, a Zotero[4] kiegészítésével jött létre.

A Zotero egy nyílt forráskódú projekt, amely könnyen kiegészíthető az elérhető Javascript, illetve Szerver API segítségével. A prototípus megoldáshoz a Zotero Firefox-ba ágyazott verziójához készítettem egy plugint, amely maga is egy Firefox kiegészítő. A felhasználói felületet a Mozilla XML alapú leírónyelvével (XUL) készítettük el, a háttérben futó logikát pedig Javascript-ben. Az API-kon keresztül lehetőség van a Zotero-ban elmentett publikációk meta adatainak kinyeréséhez és azok kiegészítéséhez az elkészített szemantikus absztrakttal.



1. ábra - Kontrollált nyelvű absztrakt szerkesztő felület

A Zotero menüjének kiegészítésével minden publikációhoz elérhetővé válik egy "kontrollált nyelvű absztrakt szerkesztő" ablak, amelynek jelenlegi változata az 1. ábrán szerepel. Első látásra nagyon egyszerűnek tűnik, de minden lényegi funkcionalitást tartalmaz: szabadszöveges gépelési lehetőség, aktuális lehetséges folytatáslista (rész-szó szerint szűrve) és az absztrakt elmentése.

A szerkesztői ablak folyamatosan kommunikál a távoli szerveralkalmazással, ahol az ablak megnyitásakor létrejön egy, a felhasználó és az adott publikáció számára dedikált nyelvtani elemző a megfelelő nyelvtani profillal. Utóbbi magában foglalj a nyelvtani szabályokat és a kapcsolódó ontológiákat is. A nyelvtani elemző minden begépelt karakter után kap frissítést, és inkrementális elemzéssel eldönti, hogy az adott ponton milyen folytatások elfogadottak (ha még érvényes az aktuálisan szerkesztett mondat), majd ezek listáját elküldi a kliensnek.

A prototípus megvalósításához egy orvostudományi mintanyelvtant hoztunk létre, amely genetikai rák kutatásokkal kapcsolatos állítások megfogalmazását teszi lehetővé. A nyelvtanba bekapcsoltuk a PRO[5] ontológiát, amely a megfelelő gén taxonómiát biztosítja.

---

[4] http://www.zotero.org
[5] http://pir.georgetown.edu/pro/pro.shtml

Az ontológia betöltéséhez az Apache Jena[6] keretrendszert használva SPARQL lekérdezésekkel szabtuk meg, mely fogalmakra van szükség. Ez persze dinamikusan módosítható, a kinyert ontológia információt az aktuális alkalmazáshoz igazítva.

Az alábbi mintakód mutatja be Prolog DCG leíró nyelven a "Patient had mutation of BRCA1"[7] alakú mondatoknak megfelelő nyelvtant.

```
sentence => [patient], [had], genechange, [of], $gene.
$change => [overexpression].
$change => [mutation].
genechange => $change.
genechange => $change, [and], $change.
```

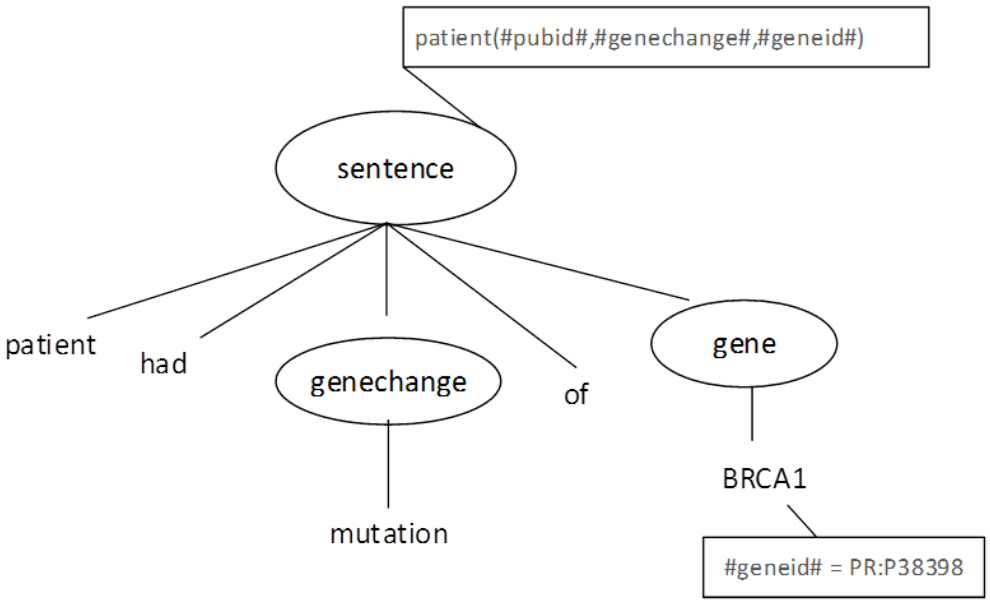
A `$gene` szimbólum jelöli a mondatszerkezetben az ontológiára történő hivatkozást. Ennek a szimbólumnak a feloldása dinamikusan történik, naprakész ontológia információt használva - tehát a nyelvtan készítőjének nem kell nyilvántartani a területhez tartozó ontológiát, csupán egyértelmű hivatkozást kell biztosítani rá (a nyelvtani profil segítségével).

A példában nem látszik, de a nyelvtan fontos részét képezi az egyes struktúrákhoz tartozó formális reprezentációk. Az elkészített szemantikus absztrakt mondatainak elemzése során az illeszkedő nyelvtani szabályok alapján úgynevezett elemzési fák épülnek fel. Az elemzési fából kinyerhetőek a szabályok és a szimbólumok, amelyeknek járulékos információi alapján generálhatjuk le az adott mondathoz tartozó formális reprezentációt.

A 2. ábrán látható a mintamondat elemzésekor elkészülő elemzési fa. A példához egy egyszerű Prolog szintaktikájú reprezentáció jön létre:

[6] https://jena.apache.org/
[7] "A páciensnek BRCA1 mutációja volt."

```
patient(245689,"mutation",PR:P38398).
```



2. ábra - Mondat elemzési fája

A formális reprezentáció készítése mindig a gyökér szimbólumból indul. Ez tartalmazza az alapvető struktúrát, kettős kereszttel jelölve bizonyos paramétereket, melyek a fa későbbi bejárásával kerülnek feloldásra (pl. `genechange`). A kapott formális Prolog mondat tartalmazza az absztrakthoz kapcsolódó publikáció egyedi azonosítóját (`pubid`) és a hivatkozott ontológia fogalom referenciáját (`geneid`) is.

A szemantikus absztrakt elemzése után felépül az állításokból egy formális tudásbázis. Ezen a tudásbázison már nem csak kulcsszavakra, hanem valamilyen összetett logikai állításra is lehet keresni, pl. "mely publikációkban foglalkoznak olyan páciensekkel, akiknek konkrétan BRCA1 mutációjuk volt de nem volt BRCA2 mutációjuk". A Zotero kiegészíthető egy új kereső szolgáltatással, amely a szerver tudásbázisával kommunikálva komplex lekérdezéseket is meg tud valósítani a tárolt gyűjteményen.

Összefoglalás

A digitális könyvtáraknak fontos része hatékony keresőszolgáltatások biztosítása a felhasználók részére. Ezen szolgáltatások fejlesztéséhez szükséges van intelligens, szemantikus adattárolásra a publikációk tartalmáról. A cikkben bemutattuka kontrollált nyelvű szemantikus absztraktok ötletét. Ezen absztraktok természetes nyelven, könnyen elkészíthetőek a szerkesztők által, kiegészítve a publikációkról tárolt eddigi meta adatokat valamilyen formális logikai reprezentációval is. A prototípus implementáció példát ad egy lehetséges alkalmazási területre, de a felvázolt ötletek tetszőleges információs rendszer kiterjesztéséhez hasznosíthatóak.

Irodalomjegyzék

[1] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat. Biotechnol.*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007.

[2] M. Gerstein, M. Seringhaus, and S. Fields, "Structured digital abstract makes text mining easy," *Nature*, vol. 447, no. 7141, pp. 142–142, Oct. 2007.

[3] T. Kuhn, "A Survey and Classificiation of Controlled Natural Languages," 2012. [Online]. Available: http://attempto.ifi.uzh.ch/site/pubs/papers/kuhn2013cl.pdf. [Accessed: 10-Oct-2013].

[4] A. Bernstein and E. Kaufmann, "GINO – A Guided Input Natural Language Ontology Editor," in *The Semantic Web - ISWC 2006*, vol. 4273, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, Eds. Springer Berlin / Heidelberg, 2006, pp. 144–157.