Magyar nyelvű orvosi szakcikkek hivatkozásainak automatikus feldolgozása

Farkas Richárd[1], Kojedzinszky Tamás[1], Sliz-Nagy Alex[1],
Tímár György[2], Zsibrita János[1]

[1] Szegedi Tudományegyetem, Informatikai tanszékcsoport
Szeged, Árpád tér 2.
rfarkas@inf.u-szeged.hu
[2] Comfit kft.
gyorgy.timar@comfit.hu

Kivonat: Cikkünkben bemutatunk egy szakirodalmi hivatkozások feldolgozására kidolgozott nyelvtechnológiai rendszert. A rendszer két legfontosabb modulja egy közelítő illesztésen alapuló visszakereső modul és egy szekvenciajelölő modul, ami a hivatkozások egyes elemeit azonosítja. Ez utóbbi megoldásnál egy újszerű kétlépcsős újrarangsoroló technikát is ismertetünk.

1. Bevezető

A magyar nyelvű orvosi szaklapok egy zárt rendszert alkotnak, a publikációk többsége nem érhető el publikusan, így azokat nem indexelik a sztenderd citációs adatbázisok (mint például Web of Science, Scopus vagy Google Scholar).

A Comfit kft. és a Szegedi Tudományegyetem Informatikai Tanszékcsoportjának közös projektjében azt a célt tűztük ki, hogy a cég magyar nyelvű orvosi szaklap adatbázisában (körülbelül 70 000 publikáció) szereplő hivatkozásokat automatikus eszközökkel feldolgozzuk, hogy az később alkalmas legyen tudománymetrikai mutatók számítására.

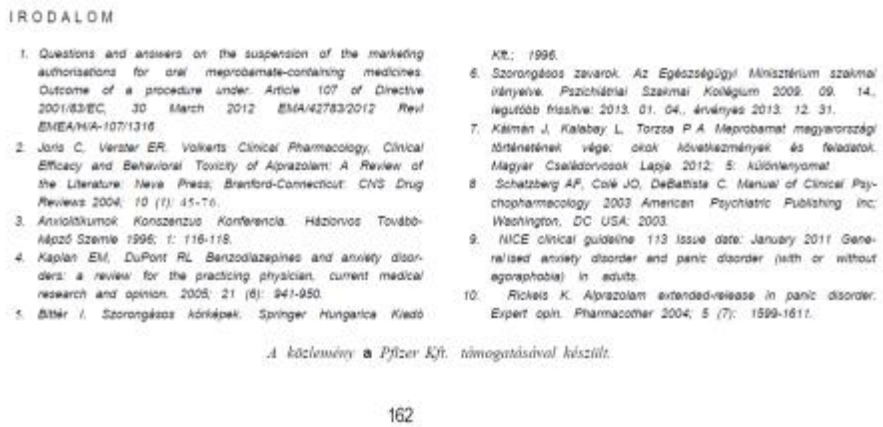
A feldolgozásra egy háromlépéses rendszert dolgoztunk ki. Először azonosítani kell a hivatkozásblokkokat a publikációkban és az egyes hivatkozásrekordokat szegmentálni kell. Ezután egy modul bejárja a rekordokat és megvizsgálja, hogy ismert publikációs adatbázisokban szerepel-e a hivatkozás. Végül a nem illesztett hivatkozásrekordokat elemezzük. Ehhez egy osztályozó eldönti, hogy újságcikkről van-e szó és ha igen, akkor a hivatkozás elemeit (szerzők nevei, cím, újság, évszám stb.) azonosítjuk. Ennek segítségével a meglévő publikációs adatbázis egy kézi jóváhagyás után gyorsan bővíthető.

2. Hivatkozásrekordok azonosítása

Első lépésben a pdf formátumban lévő újságcikkekből kellett a szöveges tartalmakat kinyerni. Itt komoly gondot okozott a többhasábos szerkesztés és a grafikonok, hirdetések nagy száma. A rendszer formázott szöveges bemenetét végül egy manuális szabályrendszer segítségével állítottuk elő. Ez a bemenet folyószöveges részeket tartalmazott. Az előfeldolgozás főbb lépései a hasábok azonosítása és azok összekötése valamint a sorvégek detektálása, sorvégi elválasztások helyreállítása.

A hivatkozás blokkok felismerésére ezután egy Conditional Random Fields alapú szekvenciajelölő módszert [1] fejlesztettünk ki, ami a szöveg egyes soraihoz REFERENCIA vagy NEMREFERENCIA címkét rendel. Ennek tanításához 150 cikkben kézzel bejelöltük a hivatkozásblokkokat. A rendszer egyes sorokat leíró jellemzőkészlete tartalmi (pl. az „irodalom" vagy „referencia" sztringeket tartalmazza), formai (pl. milyen hosszú, digitek és egyéb karakterek aránya), valamint környezeti (pl. a következő és megelőző 5 sorban hány évszám szerepel) jegyeket tartalmazott.

A hivatkozásblokkokat végül reguláris kifejezések és egyéb szabályok segítségével bontjuk rekordokra. Ez a rekordra bontás kiaknázza a hivatkozásblokkok azon tulajdonságát, hogy valamilyen módon sorszámozva vannak azok. Gyakran előfordul ugyanis, hogy egy sorszámmal kezdődik, de az csak a hivatkozásrekord része (pl. oldalszám), és nem egy új rekord sorszáma (lásd például a 1. ábrán).



IRODALOM

1. Questions and answers on the suspension of the marketing authorisations for oral meprobamate-containing medicines. Outcome of a procedure under Article 107 of Directive 2001/83/EC, 30 March 2012 EMA/42783/2012 Rev1 EMEA/H/A-107/1316
2. Joris C, Verster ER. Volkerts Clinical Pharmacology, Clinical Efficacy and Behavioral Toxicity of Alprazolam: A Review of the Literature. Neva Press; Branford-Connecticut: CNS Drug Reviews 2004; 10 (1): 45-76.
3. Anxiolitikumok Konszenzus Konferencia. Háziorvos Továbbképző Szemle 1996; 1: 116-119.
4. Kaplan EM, DuPont RL. Benzodiazepines and anxiety disorders: a review for the practicing physician. current medical research and opinion. 2005; 21 (6): 941-950.
5. Bitter I. Szorongásos kórképek. Springer Hungarica Kiadó Kft.; 1996.
6. Szorongásos zavarok. Az Egészségügyi Minisztérium szakmai irányelve. Pszichiátriai Szakmai Kollégium 2009. 09. 14., legutóbb frissítve: 2013. 01. 04., érvényes 2013. 12. 31.
7. Kálmán J, Kalabay L, Torzsa P A. Meprobamat magyarországi történetének vége: okok következmények és feladatok. Magyar Családorvosok Lapja 2012; 5: különlenyomat
8. Schatzberg AF, Cole JO, DeBattista C. Manual of Clinical Psychopharmacology 2003 American Psychiatric Publishing Inc; Washington, DC USA: 2003.
9. NICE clinical guideline 113 Issue date: January 2011 Generalised anxiety disorder and panic disorder (with or without agoraphobia) in adults.
10. Rickels K. Alprazolam extended-release in panic disorder. Expert opin. Pharmacother 2004; 5 (7): 1599-1611.

A közlemény a Pfizer Kft. támogatásával készült.

162

1. Ábra. Hivatkozási blokk a Ferencz Cs.: Hogyan tovább? Meprobamat után… Háziorvosi Továbbképző Szemle. 2013. 18 pp 160-162 cikkből.

3. Közelítő illesztések adatbázisban

Rendelkezésünkre állt a Medline adatbázis[1] 26 millió nemzetközi orvostudományi publikációs adatbázisa és a Comfit kft. magyar publikációs adatbázisa 70 ezer elemmel. Ezek az adatbázisok strukturált formában tartalmazzák a publikációk metaadatait (szerzők, cím, újság stb). Az adatbázisban történő pontos kereséshez számos közelítő heurisztika implementálására volt szükség, ugyanis a hivatkozások hemzsegnek az elgépelésektől, rövidítésektől és hibáktól.

A keresést a SolR rendszerben[2] implementáltuk. Ez hatékony közelítő keresést biztosított több tízmillió rekord felett is. Az illesztés elfogadására egy küszöbértéket határoztunk meg, ami a tokenszintű TF-IDF-el súlyozott koszinusz távolság és egy karakteralapú szerkesztési távolságból képzett aggregált hasonlósági mértékre vonatkozott. A két megközelítés együttes alkalmazására azért volt szükség, mert a tokenalapú metrika képes az egyes szavak (tipikusan a hivatkozás elemeinek) sorrendbeli különbségének kezelésére, míg a karakteralapú szerkesztési távolság képes kezelni az elírásokat, de az egyes tokenek sorrendjének felcserélését nem.

4. Kétlépcsős módszer tulajdonnév-felismerésre

Azokat a hivatkozásokat, amelyeket nem sikerült az adatbázisokban azonosítani, osztályoztuk újságcikk/könyv/könyvfejezet/URL/egyéb kategóriákba. A feladatra egy kézi szabályrendszert dolgoztunk ki, ami különböző reguláris kifejezéseken, valamint a leghasonlóbb adatbázisrekordból a szövegrészletre visszailleszthető mezők számosságán alapul.

Végül kidolgoztunk egy szekvenciajelölő algoritmust, ami az ismeretlen újságcikk hivatkozásrekordok egyes elemeit azonosítja (szerzők, cím, év, újság, oldalszámok). Ez a módszer lehetőséget biztosít az adatbázisokban nem szereplő, új hivatkozások összegyűjtésére és az adatbázis bővítésére. Ennek tanítására az ún. távoli felügyelet módszerét követtük, a sikeresen illesztett adatbázisrekordok egyes mezőit visszajelöltük az eredeti szövegrészletekre. Ez egy elég zajos, de nagyméretű (52 ezer rekord) tanító adatbázist eredményezett.

Maga a szekvenciajelölő módszer egy újszerű kétlépcsős megközelítés. Itt először egy tanított Maximum Entrópia Markov-modell [1] megadja a 100 legvalószínűbb szekvenciát, majd egy második felügyelt tanuláson alapuló újrarangsoroló lépés, frázis- és szekvenciaszintű jellemzők kiaknázásával, kiválasztja a legjobb szekvenciát. Ennek motivációja az, hogy a sztenderd szekvenciajelölők (MEMM, CRF stb.) tipikusan csak a lokális környezet leírására alkalmas jellemzőkkel dolgoznak, mint például a megelőző és a rákövetkező 3-4 token. De a jellemzőkészlet nem kódol az egész szekvencia jelölésére vonatkozó *nem lokális* információkat. A legegyszerűbb ilyen információ az lehet, hogy az egyes címkékből hány összefüggő címkesorozat predikálódott.

[1] *www.ncbi.nlm.nih.gov/pubmed*
[2] *http://lucene.apache.org/solr/*

A referenciaelemek azonosításánál például triviális megkötés, hogy legfeljebb *egy cikkcím* és egy *újságnév* kerülhet jelölésre. Egy ilyen jellegű megkötést nem lehet az egyszerű szekvencia jelölőkbe bevezetni, az csak speciális dekóder esetén lenne lehetséges, ami a keresési tér robbanásához vezetni.

Számos struktúrapredikciós probléma esetén bevált megoldás [2], hogy egy egyszerű(bb) és gyors első fázisa a rendszernek *n* darab lehetséges jó megoldást ad, majd egy második lépésben a lehetséges megoldásokat leírjuk nem lokális jellemzőkkel hiszen itt már rendelkezésre állnak az egész jelölésszekvenciára mint megoldásra vonatkozó információk is. Ezen jellemzők alapján újrarangsorolhatjuk az első fázis által megadott jelölteket. Az újrarangsorolás történhet a felügyelt tanulási paradigma keretein belül. Ekkor a tanító adatbázist lehetséges jelöltek és a legelőre rangsorolandó elem vagy elemek alkotják. Jelen rendszerben a 100 legvalószínűbb címkesorozatot írtuk le jellemzőkkel, majd a *max P(y|Y)* célfügvényre optimalizáló újrarangsoroló implementációt alkalmaztunk [3].

5. Eredmények

A rendszerrel a 2012 és 2013 alatt megjelent 13367 db magyar nyelvű orvosi szakcikkeket dolgoztunk fel. Ezek közel egynegyedében azonosítottunk hivatkozásblokkot, ami 66766 hivatkozásrekordot tartalmazott. Ezek közül 52353 rekordot tudtunk azonosítani a rendelkezésre álló adatbázisban. A fennmaradó rekordok közül az osztályozónk szerint 5621db elem van, amely az adatbázisban nem szereplő újságcikkre hivatkozik.

Az 52 ezer illesztett rekordon tanítottuk és kiértékeltük (80-20% arányban megbontva azt tanító és kiértékelő adatbázisra) a hivatkozáselem felismerő, kétlépcsős szekvenciajelölőnket. Ennek eredményeit az 1. táblázat foglalja össze. Megjegyezzük, hogy mivel a hivatkozásokban nagyon ritkán fordulnak elő egyik címkéhez sem tartozó tokenek, ezért az *O* címke szerepeltetése a kiértékelési metrikában életszerű.

1. táblázat. Címkékkénti eredmények újrarangsorolással.

	Pontosság	Fedés	F-mérték
Szerző	94.2933	99.0349	96.6059
Cím	95.0974	97.5388	96.3027
Újság	96.2294	95.1563	95.6898
Év	97.1175	99.5097	98.2991
Oldalszám_mettől	95.3211	99.2425	97.2423
Oldalszám:_meddig	92.9677	98.9209	95.8520
O	97.3818	94.3603	95.8473

6. Összegzés

Poszterünkön bemutattuk tudományos folyóiratok hivatkozásblokkjainak feldolgozására kialakított rendszerünket. A rendszer több modulból épül fel, amelyek rendre a számítógépes nyelvészet vívmányait aknázzák ki. Ez a rendszer akkor tud helyesen működni, ha rendelkezésre áll egy nagyméretű strukturált citációs adatbázis. Ennek felhasználásával – az ún. távoli felügyelet módszerét követve – építhetünk automatikusan annotált tanító adatbázist a gépi tanulási eljárásoknak.

Köszönetnyilvánítás

Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Bibliográfia

1. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields. Foundations and Trends in Machine Learning (4) (2012)
2. Farkas, R., Schmid, H.: Forest Reranking through Subtree Ranking. In: Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP-2012 (2012) 1038-1047
3. Charniak, E., Johnson, M. Coarse-tofine n-best parsing and maxent discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL ’05 (2005) 173–180