Szótáralapú névelem-felismerés szóhatárainak javítása gépi tanulási módszerrel

Móra György[1], Farkas Richárd[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.,
gymora@inf.u-szeged.hu

[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport,
Szeged, Tisza Lajos krt. 103. III. lépcsoház
rfarkas@inf.u-szeged.hu

Kivonat: Cikkünkben angol biológiai és magyar nyelvű névelemeket felismerő rendszert mutatunk be. Megközelítésünk a szótáralapú és a gépi tanuló módszerek előnyeit ötvözi. A szótáralapú névelem-felismerők egy adott adatbázis alapján jelölik a szövegbeli előfordulásokat, így a névelemek előfordulásaihoz hozzárendelhetőek azok egyedi azonosítói. Az illesztett névelemek határainak korrekcióját, valamint a hibásan illesztett kifejezések kiszűrését a feltételes véletlen mezők módszerén alapuló statisztikai rendszerrel végeztük el. Módszerünk összehasonlítva más megközelítésekkel a magyar tulajdonnevek felismerésében közel azonos, a biológiai névelemek felismerésében pedig jobb eredményt ért el, mint a klasszikus névelem-felismerő módszerek.

1 Bevezetés

Jelen munkánkban egy hibrid névelem-felismerő rendszert mutatunk be, amely ötvözi a szótáralapú névelem-azonosítás előnyeit a gépi tanulási módszerek rugalmasságával. A névelem-felismerés során a szövegben megtalálható olyan elemeket azonosítjuk, amelyek valamilyen egyedi névvel rendelkező objektumot jelölnek. Ilyenek a tulajdonnevek, amelyek személyt, földrajzi helyet vagy szervezetet jelölnek. Ezeken kívül a különböző tudományterületeknek – mint például a biológiának – rendszerint saját névelemtípusai vannak.

A biológiai szövegek feldolgozásában különösen fontos a névelem-felismerés alkalmazása. A szövegben megtalálható gének, fehérjék és egyéb biológiai névelemek közötti relációk, valamint a gazdagabb információtartalommal bíró biológiai események kinyerésének alapja a névelemek azonosítása, amely a jelenlegi rendszerekben a szótárakban található entitásnevek segítségével történik.

Mind a biológiai névelemekhez, mind a magyar tulajdonnevekhez rendelkezésre állnak szótárak, amelyek a kifejezések előfordulásainak jelentős hányadát lefedik. A kizárólag szótárakat alkalmazó névelem-felismerő rendszerek jó fedést biztosítanak, ám a pontosság a szótár méretével csökken. Ennek kiküszöbölésére rendszerint szakértők által alkotott utófeldolgozási szabályokat alkalmaznak.

Bemutatunk egy módszert, amely alapvetően szótárillesztést hajt végre, de gépi tanuló rendszerek alkalmazásával a névelemek környezetének figyelembevételével kiszűrhetik a hibásan jelölt névelemeket, illetve korrigálhatjuk a névelemek határait, anélkül, hogy elveszítenénk azok pontos azonosításának lehetőségét.

Kísérleteink folyamán magyar nyelvű hírekben és angol nyelvű biológiai szövegekben jelöltünk névelemeket szótárillesztő módszer segítségével. A rendelkezésünkre álló tanító adatbázisok segítségével a névelemjelöltek és a környezetükben található szavakból jellemzőket nyertünk ki, amelyeket felhasználva a feltételes véletlen mezők módszer segítségével gépi tanuló modellt építettünk

2 Kapcsolódó munkák

A biológiai névelemek felismerésére alkalmazott szótáralapú módszerek általában valamilyen szabadon hozzáférhető adatbázist használnak a névelemek azonosításához. Ezek az adatbázisok akár több millió entitás adatait tartalmazhatják, amelyeket folyamatosan bővítenek és frissítenek. Az adatbázisok hivatkozásokkal kapcsolódnak egymáshoz. Az egyes biológiai entitásokra valamely adatbázisbeli azonosítójával egyértelműen lehet hivatkozni. Ahhoz, hogy a szövegben a névelemek kisebb írásmódbeli eltérései ne befolyásolják azok felismerését, a szavak normalizált alakjait illesztik a szótárban található szinonimákhoz, ezzel sok olyan változatát is meg lehet találni a névelemeknek, amelyek a szinonimák között pontosan nem szerepelnek.

Több szótár, illetve ontológia alkalmazásával a rendszerek fedése nő, azonban a hibás jelölések növekvő száma problémát jelent. Ennek kiküszöbölésére többnyire szabályalapú vagy gépi tanulást alkalmazó utófeldolgozást használnak [4].

A szótáralapú rendszerekkel szemben az általánosan alkalmazott statisztikai névelem-felismerő módszerek rendszerint valamilyen szekvenciaalapú gépi tanuló algoritmust alkalmaznak [1, 2]. Ezek lényege, hogy egy előzetesen kézzel jelölt szöveg mondatait tokenekre bontják, és ennek a tokenláncnak a címkéit tanulják valamilyen szekvenciatanuló algoritmus segítségével. Az egyik legelterjedtebb ilyen algoritmus a feltételes véletlen mezők módszere [2]. Más megközelítések a potenciálisan névelemeket tartalmazó mondatrészek, illetve szószerkezetek azonosítását követően hozzárendelik a megjelölt kifejezésekhez legjobban hasonlító szótárelemeket [5].

3 Szótáralapú névelem-felismerés

A szótáralapú módszerek legfontosabb előnye, hogy az illesztett névelemek szótárbeli bejegyzésük alapján hozzárendelhetőek a névelem által jelölt entitásokhoz, függetlenül attól, hogy az entitás mely szinonimája fordult elő a szövegben.

Az így azonosított egyedekről további információk nyerhetőek ki más – az adott objektumról információkat tartalmazó – adatbázisokból. Földrajzi helyek esetében ilyen adatok lehetnek a hely pontos koordinátái vagy a közelben található nevezetességek, személyek esetén az életrajzi adatok, szervezetek esetében pedig azok földrajzi

helye, illetve kapcsolatai különböző személyekkel. A lehetőségeknek csak a rendelkezésre álló adatbázisok által tárolt információk mennyisége szab határt.

A szótáralapú névelem-felismerés egy, a névelemeket és azok szinonimáit tartalmazó lista segítségével történik. A névelemek lehetséges alakjait keressük a szövegben, rendszerint valamilyen normalizációt alkalmazva. A normalizáció segítségével a ragozott vagy más írásmódú alakok is felismerhetőek.

3.1 A szótárillesztő algoritmus

A biológiai névelemeket tartalmazó adatbázisok sokszor több tízmillió szinonimát tartalmaznak. Ennyi névelem normalizált szövegre illesztése igen időigényes, ezért az illesztéshez a Lucene[1] Java-alapú kereső és indexelő keretrendszer felhasználásával fejlesztettünk normalizált szótárillesztő rendszert.

A szótárban található szinonimák normalizált formáira indexet építettünk, majd a szöveg szavainak normálalakjait ebben kerestük. A keresés mondatonként történt. A mondat minden szavához hozzárendeltük azokat a lehetséges névelemeket, amelyek normálalakjában az adott token szerepelt. A következő lépésben a névelemek hossza alapján kiszűrtük azokat a tokensorozatokat, amelyek nem elég hosszúak az adott névelemhez. Ezzel jelentősen csökkent a lehetséges jelölések száma. A megmaradt, immár kezelhető mennyiségű jelöléssorozatot illesztettük a szövegre.

3.2 Angol biológiai névelemek

A biológiai névelemek egyértelmű, egyedi azonosítása elengedhetetlen a különböző bioinformatikai alkalmazások számára. Kiterjedt adatbázisok állnak rendelkezésre, amelyek tartalmazzák az ismert gének, fehérjék, fajok és egyéb biológiai entitások neveit, valamint kapcsolatait [8]. Az Entrez Gene egy géneket és azok szinonimáit tartalmazó adatbázis, ennek az elemeit használtuk jelen munkában is névelemek azonosítására [7]. Az általunk használt adatbázis közel 6 millió gén 7,9 millió szinonimáját tartalmazza.

A biológiai entitások neveinek normalizált formáit az LVG2010 programcsomagban [7] található szövegnormalizáló segítségével határoztuk meg. A biológiai entitások listáját a szövegben jelölt névelemekkel egészítettük ki annak érdekében, hogy egy kellően nagy méretű, a génnevek mellett fehérjeneveket is tartalmazó szótárt szimuláljunk.

A statisztikai módszerek gépi tanulásához és a kiértékeléshez a BioCreative II génnév-felismerési feladatának [6] teszt- és tanító dokumentumait használtuk. Az illesztés során a szöveg minden tokenjéhez hozzárendeltük azokat a névelemeket, amelyeknek a normált alakjai az adott szövegbeli tokent normalizálva tartalmazzák. Ha több egymás utáni tokent is megjelölt az illesztés egy adott névelemmel, akkor annak az összes token részsorozatát megvizsgáltuk, hogy az entitás teljes normált alakjához illeszkedik-e a tokensorozat. Az egyező sorozatokat megjelöltük.

[1] http://lucene.apache.org

3.3 Magyar névelemek

Az általunk használt magyar nyelvű névelemszótár *243 497* elemet tartalmazott. A névelemszótárat kiegészítettük a szövegben jelölt névelemekkel. Ezek illesztéséhez a névelemeket kisbetűssé alakítottuk, a szavakról a magyarlanc szófaji egyértelműsítő [9] segítségével eltávolítottuk az esetleges jeleket és ragokat, így előálltak a névelemek normalizált alakjai. A normalizáció segítségével a szótáralapú módszer például illeszteni tudta a *magukénak tudhatják a Magyarországon évente szétosztott* szövegrészben a *Magyarország* névelemet.

A mérésekhez a HVG cikkeit tartalmazó *144 507* token méretű dokumentumhalmazt használtuk[2]. A kiértékelő halmazt a dokumentumok 30%-a képezte, a fennmaradó részt a statisztikai rendszerek tanítására használtuk. A szövegeket a biológia névelemeknél alkalmazott szótárillesztő módszer segítségével jelöltük.

4 Statisztikai névelem-felismerés

A gépi tanulást alkalmazó rendszerek teljesítménye erősen függ a tanító adatbázis jellemzőitől, és általában gyengébb eredménnyel alkalmazhatóak más stílusú vagy más részterületet lefedő szövegeken. Előnyük, hogy olyan névelemeket is felismerhetnek, amelyek a rendelkezésre álló szótárakban nem találhatók meg. A szótáralapú módszer jelöléseit jellemzőként felhasználva a klasszikus statisztikai névelem-felismerők teljesítménye javul, de az így előálló annotációk már nem rendelkeznek a szótáralapú módszerek előnyeivel. Célunk olyan rendszer megalkotása volt, amely felveszi a versenyt a szótárjellemzőket használó klasszikus névelem-felismerőkkel.

Az általunk alkalmazott statisztikai névelem-felismerő rendszer a magyar, valamint az angol nyelvű névelem-felismerésben általánosan használt felszíni és nyelvi jellemzőket használta (részletesen l. [1]). A Mallet nevű programcsomag feltételes véletlen mezők módszerét használtuk a szekvenciák tanulására és predikálására [3].

A szótáralapú és statisztikai módszerek előnyeinek ötvözéséhez egy speciális szekvenciajelölési feladatot fogalmaztunk meg. A tanítóhalmaz dokumentumait a szótáralapú módszer segítségével jelöltük, majd az így keletkezett szótárjelölések három token sugarú környezetét véve statisztikai modellt tanítottunk a véletlen mezők módszerének használatával. Minden, a szótáralapú módszerrel megjelölt kifejezés környezete egy külön szekvenciát alkotott. A szekvenciák tartalmazhatták a szomszédos szótárjelölést is, amelyekről az adott láncban nem kellett döntést hozni. Ezeket eltérő címkével jelöltük meg. A névelemet jelentő címkesorozatot a statisztikai rendszer módosíthatta, így megváltoztatva a jelölést vagy annak határait.

A szekvenciajelölési feladatnak a szótárjelölések környezetére való korlátozásával a névelemtokenek aránya nagyobb az egyes tanító példányokban, mintha az egész mondat tokenláncán tanítanánk a statisztikai névelem-felismerőt. Az így felépített modell a szótárillesztés hibáit tanulja meg kiküszöbölni, és nem tanulja meg feleslegesen az olyan mondatrészek *nem névelem*ként való címkézését, amelyek nem tartalmaznak névelemet jelölő szavakat.

[2] http://www.inf.u-szeged.hu/rgai/corpus_ne

A tanító és a kiértékelő halmaz dokumentumaiban a szótárak minden előfordulását normalizáció alkalmazásával illesztettük. A normalizációhoz a szótárak leírásánál használt magyar és angol nyelvű módszereket használtuk.

A normalizált illesztés a szavak sorrendjét nem veszi figyelembe, ennek a biológiai névelemek illesztésénél van jelentősége, ahol a több tokenből álló névelemekben a szavak sorrendje gyakran változó, például a *G-protein coupled receptor family C group 5 member D* fehérje megnevezése lehet *5 member of G-protein coupled receptor family C group* is. Mivel a szótárak nem tartalmazzák az összes lehetséges írásmódot, a tokensorozat normalizálásakor a névelem tagjait rendszerint sorba rendezik, illetve bizonyos stopszavakat nem vesznek figyelembe az illesztésnél.

5 Eredmények

A különböző névelem-felismerő rendszerek által elért eredményeket a sztenderd F-mérték metrika alkalmazásával adjuk meg frázis- és tokenszinten. A frázisszintű kiértékeléskor csak a névelem minden tokenjének egyezése számított jó jelölésnek, míg a tokenszintű kiértékelés esetén minden tokenre megvizsgáltuk, hogy az automatikus jelölés egyezik-e a kézi címkével.

5.1 Biológiai névelemek felismerése

A szótáralapú módszer fedése a szótár kibővítésének köszönhetően majdnem teljes volt, de a pontosság csak a 0,25-ös értéket érte el. A szótár által illesztett szavak és azok határainak statisztikus javítása a fedést 0,15-del csökkentette ugyan, de a pontosság 0,86-ra nőtt így összességében a névelem-felismerés pontossága 0,40-ről 0,85-re nőtt. A kiértékelést frázis- és tokenszinten is elvégeztük, az eredményeket az 1. táblázat tartalmazza. Az általunk fejlesztett névelem-felismerő rendszer eredményei a szótár+CRF oszlopban találhatóak.

1. táblázat: A szótáralapú módszer és a statisztikai javítást alkalmazó biológiai névelem-felismerő összehasonlítása.

		szótár+CRF	**szótár**
FRÁZIS	Pontosság	0,860	0,252
	Fedés	0,838	0,979
	F-mérték	0,849	0,401
TOKEN	Pontosság	0,805	0,345
	Fedés	0,835	0,969
	F-mérték	0,819	0,509

Módszerünket klasszikus statisztikai névelem-felismerő módszerekkel hasonlítottuk össze. A 2. és 4. táblázatban *statisztikai szótárral* elnevezésű mérések során a statisztikai névelem-felismerő a szótáralapú illesztés jelöléseit jellemzőként használta. Azt tapasztaltuk, hogy a szótárjelölések alkalmazása egyaránt pozitívan befolyásolta a

statisztikai jelölés pontosságát, valamint fedését a szótár jelöléseit nem alkalmazó változathoz képest.

2. táblázat: A statisztikai módszerrel javított szótárillesztést alkalmazó és a hagyományos statisztikai biológiai névelem-felismerők eredményeinek összehasonlítása.

		szótár+CRF	statisztikai	statisztikai szótárral
FRÁZIS	Pontosság	0,860	0,715	0,745
	Fedés	0,838	0,643	0,678
	F-mérték	0,849	0,677	0,710
TOKEN	Pontosság	0,805	0,636	0,662
	Fedés	0,835	0,598	0,631
	F-mérték	0,819	0,617	0,646

A szótár+CRF alapú módszer a klasszikus megközelítésnél minden tekintetben jobban teljesített, így a névelem-felismerés F-mértéke 15 százalékponttal meghaladta a szótárat használó, illetve 17 százalékponttal a szótár nélküli klasszikus statisztikai megközelítés teljesítményét.

5.2 Magyar tulajdonnevek felismerése

A biológiai névelemekhez hasonlóan a magyar tulajdonnevek esetében is magas fedést tapasztaltunk a szótáralapú névelem-felismerő használatakor, azonban itt a módszer pontossága nem volt annyira alacsony, mint a biológiai névelemek esetén. A statisztikai javítás azonban a magyar tulajdonnevek esetében is jelentősen, 17 százalékponttal javította a felismerés F-mérték szerinti teljesítményét. A frázis- és tokenszintű kiértékelés eredményeit a 3. táblázat tartalmazza.

3. táblázat: A szótáralapú és a statisztikai javítás módszerét használó magyar tulajdonnév-felismerő összehasonlítása.

		szótár+CRF	szótár
FRÁZIS	Pontosság	0,981	0,695
	Fedés	0,952	0,957
	F-mérték	0,967	0,805
TOKEN	Pontosság	0,980	0,755
	Fedés	0,908	0,907
	F-mérték	0,943	0,824

Az angol biológiai névelemektől eltérően a magyar nyelvű szövegekben található tulajdonnevek felismerésénél a statisztikai módszerrel javított szótárillesztés nem ért el jobb eredményt a szótárt mint jellemzőt alkalmazó klasszikus statisztikai módszerhez képest. Ennek oka valószínűleg az, hogy míg a biológiai szövegekben a szótárak-

ban szereplő névelemek nagy része nem névelemként is előfordul, a cikkekben szereplő magyar tulajdonnevek általában egyértelműbbek voltak, így egyszerű jellemzőként felhasználva a klasszikus feltételes véletlen mezőket alkalmazó módszer is eredményesen tudta alkalmazni, anélkül, hogy a szótárillesztés miatti fedéscsökkenés negatívan hatott volna a teljesítményre.

A 4. táblázat eredményeiből látszik, hogy bár a pontosság tekintetében 10 százalékponttal felülmúlta a klasszikus megközelítés által elért eredményt a módszerünk, F-mérték szerinti teljesítménye fél százalékponttal kisebb volt, mint a klasszikus névelem-felismerőé.

4. táblázat: A statisztikai módszerrel javított szótárillesztést alkalmazó és a hagyományos statisztikai tulajdonnév-felismerők eredményeinek összehasonlítása.

		szótár+CRF	**statisztikai**	**statisztikai szótárral**
FRÁZIS	Pontosság	0,981	0,902	0,971
	Fedés	0,952	0,865	0,973
	F-mérték	0,967	0,883	0,972
TOKEN	Pontosság	0,980	0,895	0,960
	Fedés	0,908	0,834	0,937
	F-mérték	0,943	0,863	0,949

A szótárat használó és nem használó klasszikus statisztikai módszerek teljesítménye között 9 százalékpont a különbség.

6 Konklúzió

A biológiai névelemek esetében az általunk kifejlesztett hibrid megközelítés által elért eredmények alátámasztják, hogy egy kellően nagy fedésű szótár segítségével eredményesen és kellő pontossággal ismerhetőek fel biológiai entitások nevei, anélkül, hogy lemondanánk a szótáralapú módszerek előnyeiről.

A magyar névelemek esetében a klasszikus módszeren nem javít megközelítésünk aminek oka valószínűleg az, hogy itt a tulajdonnév/köznév többértelműség elenyészően kicsi.

A jövőben további biológiai adatbázisok bevonásával és a normalizációs módszerek javításával olyan hibrid névelem-felismerőt kívánunk fejleszteni, amely egyszerre több névelemtípus jelölését is el tudja végezni.

Köszönetnyilvánítás

A kutatást – részben – a BAROSS_DA07-DA_TECH_07-2008-0028 projekt támogatta.

Hivatkozások

1. Farkas R., Szarvas Gy.: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. In:IV. Magyar Számítógépes Nyelvészeti Konferencia (2006)
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning (2001) 282–287
3. McCallum, A. K.: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (2002)
4. Corbett, P., Milward, D.: Annotating Biomedical Entities with I2E using Multiple Ontologies. In: Proceedings of the First CALBC Workshop (2010)
5. Ando, R. K.: BioCreative II Gene Mention Tagging System at IBMWatson
6. Smith, L., Tanabe, L., Ando, R., Kuo, C. J., Chung, F. I., Hsu, C. N., Lin, Y. S., Klinger, R., Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C., Povinelli, R., Vlachos, A., Baumgartner, W., Hunter, L., Carpenter, B., Tsai, R., Dai, H. J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Lopez, M. M., Mata, J. Wilbur, J. W.: Overview of BioCreative II gene mention recognition. Genome Biology Vol. 9 Suppl. 2. (2008)
7. The NCBI handbook. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2002) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books.
8. Torii, M., Hu, Z., Wu, C. H., Liu, H.: BioTagger-GM: A Gene/Protein Name Recognition System. Journal of the American Medical Informatics Association Vol. 16 No. 2. (2009) 247–255
9. Zsibrita J., Nagy I., Farkas R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (2009)