Automatikus korpuszépítés tulajdonnév-felismerés céljára

Nemeskey Dávid Márk[1], Simon Eszter[2]

[1] MTA SZTAKI
1111 Budapest, Lágymányosi utca 11., e-mail:nemeskey.david@sztaki.mta.hu

[2] MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr u. 33., e-mail: simon.eszter@nytud.mta.hu

Kivonat A felügyelt gépi tanulási módszerek alkalmazásához nagyméretű annotált korpuszokra van szükség, amelyek előállítása rendkívül emberierőforrás-igényes. Több lehetőség van az annotációs költségek csökkentésére, ezek közül az egyik az automatikus annotálás. Cikkünkben egy nyelvfüggetlen módszert mutatunk be, mellyel bármely Wikipédiával rendelkező nyelvre előállítható tulajdonnévi címkéket tartalmazó korpusz. Az automatikus annotálás során a DBpedia ontológiai kategóriáit képeztük le CoNLL-névosztályokra. Cikkünkben a magyar korpusz részletes hibaelemzését és kiértékelését adjuk.

Kulcsszavak: tulajdonnév-felismerés, korpuszépítés, automatikus annotáció, Wikipédia

1. Bevezetés

Az automatikus tulajdonnév-felismerés (Named Entity Recognition, NER) a természetes nyelv feldolgozását célzó alkalmazások közül az egyik legnépszerűbb, mivel hatékonyan automatizálható, és eredménye hasznos bemenete különböző magasabb szintű információkinyerő és -feldolgozó rendszereknek. A feladat során strukturálatlan szövegben kell azonosítani és az előre definiált osztályok valamelyikébe besorolni a neveket. A tulajdonnév-felismerés feladata a 6. Message Understanding Conference (MUC) egyik versenykiírásában jelent meg először 1995-ben [1]. Itt három alfeladatot különítettek el: tulajdonneveket, temporális és különböző numerikus kifejezéseket kellett felismerni. A NER-közösségen belül a temporális és a numerikus kifejezések annotálása is elfogadott, de a leginkább vizsgált típusok a személy-, földrajzi és intézménynevek. Ezek mellé vezettek be a CoNLL-versenyeken [2,3] egy negyedik típust, amely az előző háromba nem tartozó egyéb tulajdonneveket foglalja magában. Az azóta eltelt időben ezek az annotációs sémák váltak nemzetközileg elfogadottá.

A versenyekre épített és aztán közzétett tulajdonnév-annotált korpuszok képezik azokat a sztenderdeket, amelyek összemérhetővé teszik az egyes névfelismerő rendszereket. Ezek a korpuszok meglehetősen korlátozott méretűek és témaspecifikusak. Kellően robusztus tulajdonnév-felismerő rendszerek építéséhez

viszont nagyméretű, a téma tekintetében heterogén korpuszokra van szükség. A kézi annotálás rendkívül idő-, erőforrás- és szakértelemigényes feladat, ezért az elmúlt időkben különösen nagy hangsúly került az annotált erőforrások automatikus előállítására. Ennek egy módja, ha már rendelkezésre álló korpuszokat dolgozunk össze; ekkor a különböző annotációs sémák és címkekészletek összeillesztése állít elénk problémákat. Egy másik lehetőség az olyan webes közösségi tartalmak felhasználása korpuszépítéshez, mint például a Wikipédia, a Wiktionary vagy a DBpedia. Megint másik megközelítés az annotáció automatizálása, ami az esetek nagy részében egy már rendelkezésre álló adathalmazon tanított rendszer új szövegen való futtatását jelenti.

Cikkünkben egy olyan megközelítést mutatunk be, mely ezen lehetőségeket kombinálja: automatikus eszközökkel tulajdonnév-annotált korpuszokat építettünk Wikipédia szócikkekből. Munkánk során új módszert alkalmaztunk: a DBpedia ontológiai kategóriáit képeztük le CoNLL-névosztályokra. A módszert egyelőre a magyar és az angol Wikipédiára alkalmaztuk.

A cikk a következőképpen épül fel. A 2. fejezetben bemutatjuk a Wikipédia eddigi felhasználási módjait a tulajdonnév-felismerés területén. A 3. fejezetben leírjuk a korpuszépítési módszert, elsősorban a magyar nyelvű adatokra koncentrálva. Az alkalmazott módszer részletes hibaelemzését a 4., a korpuszok leírását a 5., míg a kiértékelést és az eredményeket a 6. fejezet adja. Cikkünket az elért eredmények rövid összefoglalása zárja (7. fejezet).

2. Wikipédia és tulajdonnév-felismerés

A Wikipédia egy többnyelvű, nyílt tartalmú, az internetes közösség által fejlesztett webes világenciklopédia[3]. Több mint 22 millió szócikkével kincsesbánya a különböző természetesnyelv-feldolgozó fejlesztések számára; használták már többek között jelentésegyértelműsítére, ontológia- és tezauruszépítésre, valamint kérdésmegválaszoló rendszerekhez (további alkalmazási lehetőségekért lásd [4]). Mivel a Wikipédia címszavainak jelentős része tulajdonnév, adja magát a lehetőség, hogy a tulajdonnév-felismeréshez is használjuk.

A Wikipédia legkézenfekvőbb alkalmazási módja nagyméretű névlisták előállítása, melyek javítják az általános célú névfelismerők hatékonyságát, felügyelt és felügyelet nélküli módszerek esetében is (pl. [5] és [6]), továbbá a név-egyértelműsítésben is fontos szerepet játszanak (pl. [7]). A Wikipédiában található tudás jegyek formájában is beépíthető tulajdonnév-felismerő rendszerekbe: például Kazama és Torisawa [8] kísérlete azt bizonyítja, hogy a Wikipédia kategóriacímkéinek automatikus kinyerése növeli egy felügyelt névfelismerő rendszer pontosságát.

A Wikipédia alkalmazására a névfelismerés területén egy másik lehetőség magának a szövegbázisnak a felhasználása. Richman és Schone [9] kevés erőforrással rendelkező nyelvek Wikipédia-cikkeiből épített korpuszokat, amelyekben a Wikipédia inherens kategóriastruktúráját használták fel a tulajdonnevek annotálásához. Nothman et al. [10] a szócikkek első mondatából kiindulva címkézte fel

[3] http://wikipedia.org

a szövegben belinkelt neveket, így építve automatikusan tulajdonnév-annotált korpuszt.

Az általunk alkalmazott módszer az említettektől annyiban tér el, hogy mi a DBpedia ontológiai osztályait képeztük le a sztenderd CoNLL-névosztályokra, majd ezeket Wikipédia-entitásokhoz kötöttük. A Wikipédiából, tekintve a szócikkek nagy számát, kevés erőforrással rendelkező nyelvekre is tudunk kellően nagy méretű korpuszokat építeni, amelyek bemenetül szolgálhatnak névfelismerő rendszerek tanításához és teszteléséhez. Tudomásunk szerint az általunk létrehozott korpusz az első magyar nyelvű automatikusan tulajdonnév-annotált korpusz, amely szabadon elérhető és felhasználható további kutatásokhoz.

3. Korpuszépítés

A korpuszépítő algoritmus a Wikipédia két feltételezett tulajdonságán alapszik, miszerint a szócikkek többsége megnevezett entitásokról szól, valamint a folyó szövegben előforduló entitásokat a kereszthivatkozások azonosítják. Algoritmusunk e két feltételezés következményeit használja ki. Hasonlóan a Nothman et al. [10] által leírtakhoz, a korpuszépítés a következő lépésekből áll:

1. a Wikipédia-cikkeket entitásosztályokba soroljuk;
2. a cikkeket mondatokra bontjuk;
3. felcímkézzük a tulajdonneveket a szövegben;
4. kiszűrjük a rossz minőségű mondatokat.

Az algoritmus alapjaiban nyelvfüggetlen: bármely nyelvre alkalmazható, amely rendelkezik megfelelő méretű Wikipédiával. Egyedül a harmadik lépés az, ahol figyelembe kell venni a nyelv, illetve a használt annotációs séma sajátosságait. Ennek oka, hogy az egyes nyelvek, illetve sémák eltérnek abban, hogy mit tekintenek annotálandó elemnek: pl. a *római* szó a magyarban nem számít névnek, míg angol megfelelője, a *Roman* a CoNLL-séma szerint `Misc` címkét kapna.

Mivel célunk egy minél tisztább, vagyis a gold standard színvonalat közelítő korpusz előállítása volt, ha választanunk kellett egy-egy lépésnél a pontosság (precision) vagy a teljesség (recall) között, mindig az előbbi mellett döntöttünk. A Wikipédia mérete lehetővé teszi, hogy szigorú szűrések mellett is korábban nem látott méretű korpuszt állítsunk elő.

A továbbiakban röviden ismertetjük a fenti lépéseket, kizárólag a magyar nyelvű korpuszra koncentrálva. Részletes leírás, illetve az angol nyelvű korpuszban felmerülő problémák kifejtése [11]-ben található.

3.1. Wikipédia-cikkek mint entitások

Ahhoz, hogy a folyó szövegben lévő kereszthivatkozásokat felhasználhassuk a tulajdonnevek azonosítására, a hivatkozott Wikipédia-cikkeket névosztályokba kell sorolnunk. A Wikipédia saját kategóriarendszere a kategóriák nagy száma, illetve

a rendszerezettség teljes hiánya miatt nem alkalmas erre a célra. Egyes szerzők, mint Kazama és Torisawa [8] vagy Nothman et al. [10] felügyelt tanulással oldották meg ezt a feladatot. Mi azonban a klasszifikációs hibák elkerülése céljából a DBpedia [12] típushierarchiájának felhasználása mellett döntöttünk.

A DBpediában[4] a típusok egy OWL[5] ontológia részei. A tudásbázis a Wikipédia-entitások egy részhalmazát tartalmazza, és mindegyik entitáshoz hozzárendeli – többet között – azon ontológiaosztályokat, amelyekbe az tartozik. Mivel a DBpediának nincs magyar változata, a magyar entitáslista csak olyan angol oldalak adatait tartalmazza, melyeknek létezik megfelelője a magyar Wikipédiában. Ezáltal a feldolgozás köréből kiesnek kifejezetten magyarspecifikus oldalak, de így is 58.337 oldal anyagát dolgoztuk fel.

A magyar nevek kategorizálásához a Szeged NER [13] korpuszban alkalmazott névosztályokat vettük alapul, ezekre képeztük le a DBpedia egyes ontológiaosztályait. Ezután minden entitáshoz hozzárendeltük az őt tartalmazó legszűkebb osztály címkéjét, vagy `O`-t, ha az osztály nem minősül annotálandónak a Szeged NER korpusz sémája szerint. Így egy 46.461 elemű, felcímkézett, angol nyelvű tulajdonnévlistát kaptunk. Utolsó lépésként meghatároztuk ezen oldalak magyar és egyéb nyelvű megfelelőit, és ezeket is felvettük a listára. Az idegen nyelvű oldalak meghagyásának oka, hogy a magyar Wikipédia időnként ezekre is hivatkozik.

3.2. A cikkek feldolgozása

A korpusz építéséhez a magyar Wikipédia 2012. március 9-i állapotát vettük alapul. Az XML-fájlokból az mwlib[6] könyvtár segítségével nyertük ki a nyers szöveget. A tokenizáláshoz egy házon belül előállított statisztikai eszközt használtunk, amelyet a Szeged korpuszon [14] tanítottunk. A lemmatizálást és a morfológiai elemzést a HunMorph [15] és az erre épülő egyértelműsítő, a HunDisambig segítségével végeztük.

3.3. Tulajdonnevek címkézése

A címkézés két feladatot foglal magába: egyrészt azonosítani kell az entitásokat a folyó szövegben, másrészt besorolni őket a megfelelő névosztályba. A szövegben előforduló kereszthivatkozásokat potenciális neveknek tekintjük, és minden linknél megnézzük, hogy az oldal, amire mutat, szerepel-e a korábban előállított listában (lásd a 3.1. fejezetben). Ha igen, felcímkézzük a megfelelő címkével; ha nem, típusa `Unk` lesz. Végül minden mondatot, amelyben `Unk`-ként címkézett elem szerepel, eldobunk.

Kiinduló alapfeltételezésünk, miszerint az egyes Wikipédia-cikkek megnevezett entitásokról szólnak, és a szövegben előforduló entitásokat kereszthivatkozások azonosítják, nem minden esetben állja meg a helyét. Algoritmusunk ezért több helyen finomításra szorul.

[4] A korpusz elkészítéséhez az akkor aktuális 3.7-es verziót használtuk.
[5] http://www.w3.org/TR/owl-ref/
[6] http://code.pediapress.com

Először, a Wikipédia nem minden szócikke szól megnevezett entitásokról: egyes köznevek, dátumok és egyéb, nem tulajdonnévi elemek is kaptak saját oldalt. Annak érdekében, hogy az algoritmus ne kezelje ezeket `Unk` típusú entitásként, majd dobja el a mondatokat, amikben szerepelnek, szigorítottuk az entitásfelismerés szabályait: csak olyan kereszthivatkozást tekintünk potenciális névre való utalásnak, melynek szavai nagybetűvel kezdődnek. Itt és később is kihasználtuk a magyar nyelv azon tulajdonságát, hogy a tulajdonneveket, és csak azokat kezdjük nagybetűvel. Kivételt képeznek ez alól természetesen a mondatkezdő pozícióban szereplő szavak. Szigorúan véve, ez a módszer minden mondatot eldobna, így annyit módosítottunk rajta, hogy a mondatkezdő szót csak akkor tekintjük potenciális annotálandó elemnek, ha a szófaja főnév. (A magyarban a morfológiai címkékre csak korlátozottan támaszkodhattunk ebben a feladatban, hiszen a KR-kódolás nem különbözteti meg a tulajdonneveket és a közneveket.)

Másodszor, nem minden tulajdonnéven találunk kereszthivatkozást. Ennek oka kettős: ha egy Wikipédia-cikkben adott entitásra többször is névvel utalnak, csak az első alkalommal linkelik a saját oldalához. De előfordulhat az is, hogy az entitás nem rendelkezik saját szócikkel. Az ilyen esetek kezelésére minden szócikknél fenntartunk egy listát, ahol az abban felismert entitásokat, illetve egyéb nagybetűs említéseit, mint például a rájuk mutató átirányító és egyértelműsítő lapok címeit gyűjtjük. Ha ezután egy olyan mondattal találkozunk, amelyben nagybetűs szócsoport szerepel, ellenőrizzük, hogy a csoport egyezike a listában szereplő egyik entitás nevével. Amennyiben igen, felcímkézzük; ha nem, a nagybetűs szavakat ismeretlen entitásnak tekintjük.

3.4. Szűrés

Ahogy már említettük, az azonosítatlan entitást tartalmazó mondatokat kidobtuk a korpuszból, hogy növeljük az annotálás pontosságát.

A hagyományos névfelismerő rendszerek tiszta, viszonylag pontosan annotált korpuszokból tanulják meg a megfelelő paramétereiket, és a tesztelésükhöz is hasonló adathalmazra van szükség. Ezért a rossz minőségű, töredékes mondatokat, amelyek nem nagybetűvel kezdődnek és nincs mondatzáró írásjel a végükön, szintén kiszűrtük. (Az így maradt 19 milliós adathalmazzal dolgoztunk tovább; a következőkben erre referálunk teljes korpuszként.) De a minőség javítása érdekében további szűrő lépéseket is lehet tenni, így például eldobhatjuk az olyan mondatokat is, amelyek nem tartalmaznak ragozott igét. Ezzel ugyan azok is kiesnek, amelyek szabályos jelen idejű, kijelentő módú, létigét (nem) tartalmazó mondatok, de az automatikus tokenizálás és mondatra bontás hibáiból származó töredékes mondatokat eltávolíthatjuk, és így az ebből fakadó annotációs hibákat is kiszűrhetjük (lásd a 4. fejezetet).

Ha viszont közösségi tartalmak (User Generated Content) feldolgozására kívánja valaki használni a korpuszokat, amelyek köztudottan sokkal zajosabbak, mint a kézzel annotált adatok, és sok töredékes mondatot tartalmaznak, hasznos lehet minél kevesebb szűrő használata. Ezért a rossznak minősített mondatokat nem dobtuk el végleg, hanem ezt az adathalmazt is elérhetővé tettük.

4. Problémás esetek, hibaelemzés

A magyar Wikipédia korpusz gépi annotálását kézzel ellenőriztük a korpusz egy kis részén. A 19 milliós teljes korpuszt vettük alapul, amelyből véletlenszerű mondatválogatással csináltunk egy 18.830 tokent tartalmazó mintakorpuszt. Ezt kézzel felannotáltuk, és összehasonlítottuk a kézi és a gépi annotálás eredményét, amely az 1. táblázatban látható. Ha a gépi módszert egy annotátornak tekintjük, akkor az F-mérték az annotátorok közötti egyetértést mutatja.

1. táblázat. A gépi és a kézi annotálás közötti egyetértést mutató eredmények a mintakorpuszon.

	Pontosság(%)	Fedés(%)	$F_{\beta=1}$(%)	Entitások száma
LOC	98.72	95.65	97.16	161
MISC	95.24	76.92	85.11	26
ORG	89.66	89.66	89.66	29
PER	88.30	89.25	88.77	93
Összesítve	94.33%	91.59%	92.94	309

A négy kategória igazságmátrixa (2. táblázat) jól mutatja, hogy a típustévesztés aránya elhanyagolható. Ezekből az értékekből kiszámítottuk az annotátorok közötti egyetértést mérő Cohen kappát is. A 0,967-es összesített érték Landis és Koch [16] skáláján elhelyezve arról tanúskodik, hogy a korpusz annotációja megfelel a gold standard színvonalnak.

2. táblázat. A manuálisan annotált mintakorpusz igazságmátrixa.

Auto↓ / Gold→	PER	ORG	LOC	MISC
PER	83	1		2
ORG		26	1	1
LOC		1	154	
MISC			1	20

A típustévesztés alapvetően két okra vezethető vissza. Az első esetben a DBpediában lévő típusinformáció helytelen, például az *MTA* DBpedia-beli osztálya `WorldHeritageSite`, ami miatt `Loc` címkét kap `Org` helyett. Hasonló eset, amikor egy több referenciával rendelkező névnek csak egyik referense szerepel a DBpediában, így használattól függetlenül mindig ugyanazt a címkét kapja.

A típustévesztések másik részét a Wikipédia rossz kereszthivatkozásai okozzák. Például az egyik cikk szerkesztője a *Walt Disney Co.* cégnévnek csak egy

részét linkelte be, mégpedig a személyről szóló oldalra. Ezért ebben a cikkben ennek a cégnévnek a különböző változatai (*Disney*, *Walt Disney* stb.) is mind személynévként lettek jelölve.

3. táblázat. A névfelismerés további hibái.

	PER	ORG	LOC	MISC
Hibásan névként felismert szavak (álpozitív)	1	0	1	0
Fel nem ismert nevek (álnegatív)	3	0	5	4
Részlegesen felismert nevek	7	1	0	0

Jóval gyakoribbak a névfelismerés további hibái, vagyis bizonyos szavak hibásan névként való azonosítása és egyes nevek felismerésének elmulasztása, illetve az entitáshatárok pontatlan felismerése. Az egyes névtípusokra lebontott hibák számát a 3. táblázat mutatja.

Az entitáshatárok pontatlan felismerésének oka leggyakrabban az, hogy a név környezetében lévő értelmező szerepű nyelvi egység is a Wikipédia-címszó része. Emberekre utaló linkek esetében ezek a szavak nagyrészt rangjelölők, pl. *Szent István király*, *I. Benedek pápa*. Ezeket egy tematikus tiltólistával a későbbiekben ki lehet szűrni.

A Wikipédia-címszavakban szereplő értelmező szavak alkalmanként a teljes entitás felismerését is megakadályozzák. Ez akkor fordul elő, ha a hivatkozott oldal teljes címe nem tulajdonnév, viszont tartalmaz egy általunk ismert tulajdonnevet: pl. *ókori Róma*, *magyar Wikipédia*.

A fel nem ismert `Misc` típusú tulajdonnevek mindegyike írók műveit felsoroló oldalakon fordul elő. Ezen műveknek nincs saját szócikkük, ezért címkézésük nehéz. Mivel egy műcímben bármilyen nyelvi elem előfordulhat, az általunk alkalmazott szűrők együttese sem képes kiszűrni ezeket. Megoldást jelenthetne az, ha ezeket a kizárólag műcímeket felsoroló oldalakat külön kezelnénk, és egy komplex rendszert építenénk a feldolgozásukra. Mivel ez egy külön nyelvfeldolgozási feladat, jelen fejlesztésen belül nem vállalkozunk rá; a jövőben az ilyen oldalakat kihagyjuk a korpuszból.

A hibák egy további részét az alkalmazott nyelvfeldolgozó eszközök tévesztései okozzák. Az automatikus tokenizálás és mondatra bontás hibás működésére példa, amikor a rövidítést tartalmazó név (pl. *Warner Bros.*) utolsó eleme, vagyis a pont mondatvégi írásjelként értelmeződik, így nem annotálódik a névvel együtt. Máskor a mondatrabontás során a szövegben levő link szétszakad, így a maradék elem nem kapja meg a megfelelő címkét. Mivel a mondat első szavát csak akkor tekintjük potenciális entitásnak, ha főnév, a szófajmeghatározás hibája folytán előfordul, hogy átsiklunk egy mondatkezdő néven. Például a *Hél visszaengedte volna* mondatban a *Hél* szót igeként azonosította a HunDisambig, így nem tekintettük tulajdonnév-jelöltnek. E problémákra esetleg megoldást jelenthet az alkalmazott eszközök teljesítményének javítása, vagy más eszközök használata.

Tipikus jelenség a magyarban, amikor egy név összetétel eleme lesz, pl. *Bizánc-ellenes*. Ilyen esetekben a köznév az összetétel alaptagja, vagyis az határozza meg a referenciát. A referenciaváltozást természetesen a címkézés változásának kell követnie, ami viszont nem, vagy nehezen kezelhető automatikusan, mivel nagyjából bármilyen köznév kapcsolódhat névhez. A helyzetet bonyolítja az is, hogy a mozaikszavakhoz, rövidítésekhez, valamint nem ejtett magánhangzóra végződő idegen nevekhez is kötőjellel kapcsoljuk a toldalékokat, amelynek a felszíni szerkezete nagyon hasonlít az összetételekéhez. Ennek a problémának a megoldása még további vizsgálatokat követel.

A felsoroltak mellett természetesen implementációs hibákra is fény derült, amelyek azonban összességében csak néhány esetben okoztak rossz címkézést. Ezeket a jövőben javítani fogjuk.

5. A korpuszok leírása

A korpuszokat Creative Commons Attribution-Sharealike 3.0 licensz alatt publikáljuk, vagyis ugyanolyan feltételekkel adjuk tovább, ahogy a Wikipédiából letöltöttük. Szabadon elérhetőek a `http://hlt.sztaki.hu` oldalon keresztül, valamint a META-SHARE tárhelyről (http://www.meta-net.eu/). A META-SHARE egy nyílt rendszer, amely lehetővé teszi a nyelvi erőforrások megosztását. Létrehozója a META-NET, az Európai Bizottság által alapított nyelvtechnológiai hálózat.

A fájlok ún. *multitag* formátumban vannak, amelyben a tartalmas sorokat tabulátor választja el. Az első oszlop tartalmazza magukat a szövegszavakat, az egyes oszlopokban pedig a különböző szintű annotációk találhatók. A mondathatárokat üres sorok jelölik. A névcímkéken kívül minden token mellett szerepel a töve és a hozzá tartozó teljes morfológiai elemzése KR-kódokkal. Két további oszlopban közöljük, hogy sima szöveg vagy kereszthivatkozás-e az adott token, és ha utóbbi, akkor melyik szócikkre utal.

6. Kiértékelés

Kiértékelésünkben megmutatjuk, hogy a létrehozott magyar nyelvű korpusz kiválóan használható a tulajdonnév-felismerés teljesítményének növelésére több módon is.

A kiértékeléshez a Hunner [17] tulajdonnév-felismerő rendszert használtuk. A csak az egyes korpuszokra jellemző jegyeket (pl. főnévi csoportok jelölése, Wikipédia-linkek) kidobtuk, hogy növeljük a korpuszok összehasonlíthatóságát. Így a következő jegykészlettel dolgoztunk: mondatkezdő és -vég pozíciók, szóalakon alapuló jegyek, morfológiai információ és listajegyek.

Az eredmények kiszámításához a sztenderd CoNLL-módszert alkalmaztuk, vagyis az annotációt csak akkor vettük helyesnek, ha a kezdő- és végpozíció is stimmelt, és a rendszer által kibocsátott címke megegyezett a gold standard címkével. Ezen alapulva a szokásos pontosságot, fedést és F-mértéket számoltuk.

6.1. Az adatok

A korpusz a fent leírt szűrő eljárások után maradt mondatokat tartalmazza, így azokat is, amelyekben nincs egy név sem. Ezeket azért tartottuk meg, hogy amennyire lehetséges, megőrizzük a nevek eredeti, Wikipédia-beli eloszlását. Viszont amikor megvizsgáltuk az egyes korpuszok telítettségét a nevek szempontjából, arra jutottunk, hogy a gold standard adathalmazzal való összevetéskor inkább sűrítjük a szöveget, vagyis kivesszük azokat a mondatokat, amelyekben nincs név. A 4. táblázat mutatja a magyar korpuszokra vonatkozó számszerű adatokat, melyekből jól látható, hogy a Wikipédiából generált korpusz telítettsége meglehetősen alacsony. A szövegnek ez a hígsága valószínűleg annak köszönhető, hogy a módszerünk szigorú, vagyis inkább minden olyan mondatot eltávolítottunk, amelyben nem lehetett beazonosítani a nevet, minthogy rosszul annotált nevek maradjanak benne.

4. táblázat. A magyar Wikipédia és a Szeged NER korpusz mérete és telítettsége.

	huwiki	**sűrített huwiki**	**Szeged NER**
token	19.108.027	3.512.249	225.963
NE	456.281	456.281	25.896
telítettség (%)	2,38	12,99	11,46

6.2. Kísérletek és eredmények

Jelen cikkben csak a magyar korpuszon elért eredményeket közöljük, az angolra vonatkozó részletes adatokért lásd korábbi cikkünket [11]. A korpusz kétféleképpen lett kiértékelve: először saját magán, aztán egy választott gold standard adathalmazon. A nevet nem tartalmazó mondatok kiszűrése után maradt 3,5 millió tokenes korpuszt 90-10%-os arányban tanító és kiértékelő halmazra osztottuk.

Mivel a névkategóriák leképezésénél a Szeged NER korpusz címkekészletét használtuk, ezért adta magát, hogy a korpusz kiértékeléséhez is ugyanezt alkalmazzuk. Többek által (pl. [10] és [18]) bizonyított tény, hogy a korpuszok közötti kiértékelés sokkal rosszabb eredményt ad, mint a saját kiértékelő halmazon való mérés. Különböző típusú szövegek esetén a különbség 20-30% is lehet. A helyzet a mi esetünkben is nagyon hasonló (lásd az 5. táblázatot az eredményekért): a Wikipédián tanított rendszer teljesítménye közel sem olyan jó a gold standard korpusz kiértékelő halmazán mérve, mint a saját kiértékelő halmazán.

Az általunk épített korpuszt további módokon is használhatjuk a tulajdonnév-felismerés teljesítményének növelése érdekében. Egy kézenfekvő megoldás nagyméretű névlisták kinyerése a Wikipédiából, és azok hozzáadása gazetteer listaként a tanításhoz. Ez a módszer több mint 1%-kal növelte az F-mértéket.

5. táblázat. Eredmények a magyar Wikipédia korpuszon.

tanítás	teszt	Pontosság (%)	Fedés (%)	F-mérték (%)
Szeged	Szeged	94,50	94,35	94,43
huwiki	huwiki	90,64	88,91	89,76
huwiki	Szeged	63,08	70,46	66,57
Szeged_wikilisták	Szeged	95,48	95,48	95,48
Szeged_wikitag	Szeged	95,38	94,92	95,15

Egy másik kísérletünkben a rendszert a Wikipédia korpuszon tanítottuk, majd az általa kibocsátott címkéket jegyként hozzáadtuk a gold standard korpuszon való tanításhoz és teszteléshez. Ezzel a módszerrel is sikerült javítani a rendszer teljesítményét.

A kiértékelés legfontosabb eredményének a saját teszthalmazon elért 89,76%-os F-mértéket tartjuk. A kézi hibaelemzés tanulságaival együtt ez arról tanúskodik, hogy az általunk épített korpusz akár önálló gold standard adathalmazként, akár kiegészítő erőforrásként jól használható automatikus névfelismerő rendszerek építéséhez.

7. Összegzés

Cikkünkben egy új módszert mutattunk be, amellyel létrehoztunk egy magyar nyelvű, automatikusan tulajdonnév-annotált korpuszt a Wikipédiából. Az eddig alkalmazottakkal ellentétben a mi metódusunk egy leképezést valósít meg a DBpedia ontológiai osztályairól a hagyományos címkekészletekre. Az így generált címkéket aztán a rendszer hozzárendeli a Wikipédiában szereplő entitásokhoz.

Módszerünk nyilvánvaló előnyei, hogy nagyban csökkenti az annotálás költségeit, valamint hogy sokkal nagyobb adathalmazokat állíthatunk elő általa, mint kézi annotációval. Egy másik előnye, hogy bármely Wikipédiával rendelkező nyelvre alkalmazható, így kevés erőforrással rendelkező nyelvekre is előállíthatunk a gold standard minőséget közelítő korpuszokat. A létrehozott korpuszok a továbbiakban számos módon alkalmazhatók a tulajdonnév-felismerő rendszerek hatékonyságának növelésére. Amennyiben kellően tiszta a korpusz, vagy az adott nyelvre nem létezik gold standard tisztaságú adathalmaz, felügyelt gépi tanulási rendszerekhez használható tanításhoz és kiértékeléshez. Továbbá erőforrásokkal bővebben ellátott nyelvek esetében is hasznosítható a klasszikus sajtó stílustól eltérő szövegek tulajdonnév-annotálásához.

További, újdonságnak számító eredményünk, hogy az általunk előállított korpuszok szabadon elérhetőek és felhasználhatóak. Tudomásunk szerint ez az első magyar nyelvű automatikusan előállított tulajdonnév-annotált korpusz. Az angol erőforrások tekintetében is hasonló a helyzet: tudomásunk szerint a Semantically Annotated Snapshot of English Wikipedia [19] mellett az itt publikált korpusz az egyetlen szabadon felhasználható tulajdonnév-annotált korpusz.

Jelen cikkünkben a DBpedia ontológiai kategóriáit a sztenderd tulajdonnév-címkékre képeztük le, de a módszerben benne rejlik a lehetőség finomabbra hangolt tulajdonnév-hierarchiák támogatására is. Az internetes közösség által létrehozott tartalmak, mint a Wikipédia és a DBpedia, folyamatosan növekszenek, ezáltal egyre több információ felhasználását teszik lehetővé. A módszer frissítésével egyre nagyobb és finomabban annotált korpuszokat tudunk létrehozni a jövőben.

Köszönetnyilvánítás

A fejlesztés az OTKA 82333. számú projektjén belül valósult meg. A fejlesztést támogatta továbbá a CESAR projekt (No. 271022). A szerzők ezúton fejezik ki köszönetüket Zséder Attilának a Wikipédia-szövegek feldolgozásában végzett munkájáért, és Kornai Andrásnak támogatásáért.

Hivatkozások

1. Sundheim, B.: MUC-6 Named Entity Task Definition (v2.1). In: Proceedings of the Sixth Message Understanding Conference (MUC6). (1995)
2. Sang, T.K., F., E.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2002, Taipei, Taiwan (2002) 155–158
3. Sang, T.K., F., E., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada (2003)
4. Medelyan, O., Milne, D., Legg, C., Witten, I.H.: Mining meaning from Wikipedia. Int. J. Hum.-Comput. Stud. **67**(9) (2009) 716–754
5. Toral, A., Munoz, R.: A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In: EACL 2006. (2006)
6. Nadeau, D., Turney, P., Matwin, S.: Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. Advances in Artificial Intelligence (2006) 266–277
7. Bunescu, R., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. (2006) 9–16
8. Kazama, J., Torisawa, K.: Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. (2007) 698–707
9. Richman, A.E., Schone, P.: Mining Wiki Resources for Multilingual Named Entity Recognition. In: Proceedings of ACL-08: HLT, Columbus, Ohio, Association for Computational Linguistics (2008) 1–9
10. Nothman, J., Curran, J.R., Murphy, T.: Transforming Wikipedia into named entity training data. In: In Proceedings of the Australasian Language Technology Association Workshop 2008. (2008) 124–132
11. Simon, E., Nemeskey, D.M.: Automatically generated NE tagged corpora for English and Hungarian. In: Proceedings of the 4th Named Entity Workshop (NEWS) 2012, Jeju, Korea, Association for Computational Linguistics (2012) 38–46

12. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – A crystallization point for the Web of Data. Web Semantics **7**(3) (2009) 154–165
13. Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., Csirik, J.: A highly accurate Named Entity corpus for Hungarian. In: Electronic Proceedings of the 5th International Conference on Language Resources and Evaluation. (2006)
14. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus. A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In Hansen-Schirra, S., Oepen, S., Uszkoreit, H., eds.: COLING 2004 5th International Workshop on Linguistically Interpreted Corpora, Geneva, Switzerland, COLING (2004) 19–22
15. Trón, V., Gyepesi, Gy., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software. (2005)
16. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1) (1977) 159–174
17. Varga, D., Simon, E.: Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica **18** (2007) 293–301
18. Ciaramita, M., Altun, Y.: Named-entity recognition in novel domains with external lexical knowledge. In: Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing. (2005)
19. Atserias, J., Zaragoza, H., Ciaramita, M., Attardi, G.: Semantically Annotated Snapshot of the English Wikipedia. In: Proceedings of LREC 2008. (2008)