Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre

Farkas Richárd[1], Szarvas György[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.
{rfarkas, szarvas}@inf.u-szeged.hu

Kivonat: Cikkünkben bemutatunk egy, számos alkalmazásban kiemelkedő pontosságot elérő statisztikai tulajdonnév-felismerő rendszert. A modell, elsősorban az összegyűjtött nagyméretű tulajdonsághalmaz, illetve az abban rejlő lehetőségek hatékony kiaknázásának köszönhetően több összehasonlításban is versenyképesnek bizonyult a hasonló problémákra ismert legjobb módszerekkel.. Részletesen bemutatjuk a tulajdonnevek azonosításában és kategorizálásában elért eredményeinket magyar nyelvű gazdasági híreken, valamint angol nyelvű újságcikkeken és orvosi kórlapok szövegein. Az eredetileg magyar nyelvre kifejlesztett statisztikai modell apró módosításokkal, minden eddig publikáltnál jobb eredményt ért el a standard angol nyelvű tulajdonnév adatbázison, valamint első helyen végzett egy orvosi kórlapok anonimizálására kiírt nyílt nemzetközi versenyen.

1 Bevezetés

A tulajdonnevek azonosítása (és kategorizálása) folyó szövegben meghatározó fontosságú számos számítógépes nyelvfeldolgozó alkalmazás során. Példaként tekinthetjük a különböző információkinyerő rendszereket, ahol a tulajdonnevek általában jelentős, információt hordozó szerepet töltenek be a szövegben, vagy a gépi fordítási alkalmazásokat, ahol értelemszerűen más módon kell kezelni emberek, szervezetek neveit, mint a szöveg többi részét.

E cikkben bemutatjuk statisztikai tulajdonnév-felismerő rendszerünket, amelynek hatékonyságát három különböző feladaton is vizsgáltuk. Az első tesztfeladat magyar nyelvű gazdasági szövegek feldolgozása volt. Méréseinkhez a Szeged Treebank [3] gazdasági rövidhíreit használtuk. Ugyanazt a rendszert alkalmaztuk angol nyelvű újsághírekben (sport, politikai, gazdasági témákból) szereplő tulajdonnevek felismerésére, melyhez a CoNLL-2003 konferencia adatbázisát használtuk [15], illetve angol nyelvű orvosi zárójelentések anonimizálására. Anonimizálás alatt páciensek, doktorok, kórházak stb. neveinek felismerését és véletlenszerű azonosítókkal való lecserélését értjük [11].

A következő fejezetben bemutatjuk a tulajdonnév-felismerési feladatot, ezután a harmadik fejezetben kerül részletes ismertetésre az általunk készített rendszer és

annak főbb építőelemei. A negyedik részben bemutatjuk a három speciális problémát, amelyen rendszerünket teszteltük, majd elemezzük az elért eredményeket. Végül az utolsó fejezetben összefoglaljuk, értékeljük munkánkat.

2 Tulajdonnév-felismerés

Cikkünkben a *tulajdonnév* kifejezést az angol *named entity* (névkifejezés) magyar megfelelőjeként fogjuk használni. A *named entity* megnevezés magában foglal olyan kategóriákat is, amelyek nem tulajdonnevek (mint például telefonszámok, mennyiségek stb.), de a felismerés elsődleges célpontjai mégis a tulajdonnevek, a többi osztály általában egyszerű szabályok segítségével azonosítható. Hasonló módon elképzelhető az is, hogy bizonyos tulajdonneveket az adott problémánál nem áll szándékunkban jelölni (például az orvosi alkalmazásoknál), mi mégis – az egyszerűség kedvéért – a magyar *tulajdonnév* elnevezés mellett döntöttünk.

2.1 A tulajdonnév-felismerési feladat

A megoldandó feladat kétszintű: egyrészt fel kell ismerni a szöveg(ek)ben az előre definiált kategóriákba tartozó tokensorozatokat, másrészt be kell azokat sorolni a megfelelő osztályokba. Az osztályozás során meg kell különböztetni a tulajdonnevek kezdő tokenjeit, és a tulajdonnév részét képező belső elemeket. Ennek elsősorban akkor van jelentősége, amikor a szövegben egymást követően több azonos kategóriába tartozó tulajdonnév található, mert ilyenkor ezek segítségével állapítható meg az elemek kezdőpozíciója.

Az adott problémákhoz rendelkezésre áll egy-egy tanító adathalmaz, ahol a tulajdonneveket kézzel bejelölték. A cél ennek felhasználásával egy olyan modell tanítása, amely ismeretlen szövegen is hatékonyan felismeri az adott kategóriákat (induktív tanulás). Fontos megjegyeznünk, hogy a tanult modell csak a tanító halmazhoz hasonló jellemzőkkel rendelkező szövegeken működik pontosan.

2.2 A tulajdonnév-felismerés története

A tulajdonnév-felismerés problémájával a 90-es évek közepétől foglalkoznak. A Message Understanding Conference (MUC) [2] sorozat angol nyelvű újsághírek automatikus feldolgozását tűzte ki célul. A MUC-7 során a tulajdonnevek azonosítása és a *személynév, földrajzi név, szervezet, egyéb* kategóriákba sorolása, valamint egyéb, időt, mennyiséget stb. leíró kifejezések felismerése volt a feladat. Az utóbbi években további nyelvekre fókuszált a kutatás, mint például a spanyol, a német, a kínai. A Conference on Computational Natural Language Learning (CoNNL) által meghirdetett nyílt versenysorozaton 2003-ban a tulajdonnevek felismerése volt a feladat egyazon modellel angol és német nyelvű szövegekben [15].

Ebbe a trendbe jól illeszkednek a magyar nyelvvel kapcsolatos kutatásaink: létrehoztuk az első releváns méretű (200.000 szó) magyar nyelvű tulajdonnévi korpuszt [12], valamint implementáltuk az első statisztikai alapú magyar tulajdonnév-

felismerő modellt, melynek eredményei az angolra publikált eredményekkel összevethetőek.

A szakterület egy másik folyamatosan bővülő iránya a felismerő rendszerek alkalmazása különböző domainekre. Az első rendszerek (MUC, CoNNL) általános újságcikkekre koncentráltak, napjainkra azonban előtérbe került például a bioinformatikai vagy orvosi szövegek feldolgozása is [16][11]. Annak érdekében, hogy megvizsgáljuk, hogy az újsághírekre kifejlesztett rendszerünk hogyan viselkedik más domainen, idén részt vettünk egy orvosi kórlapok anonimizálását célul kitűző versenyen [14]. A versenyen első helyezést értünk el, ami bizonyítja, hogy sikerült a tulajdonnév-felismerés problémájára általánosan felhasználható eszközt építenünk.

3 A tulajdonnév-felismerő rendszer felépítése

Rendszerünk három főbb összetevőből áll: a szavakhoz tartozó tulajdonságvektorok kinyeréséből, több statisztikai modell betanításából, majd azok összekombinálásából az ismeretlen szöveg bejelölésekor.

3.1 Megközelítésünk

Az általunk alkalmazott gépi tanulási módszer eltér a problémára leggyakrabban alkalmazott, legsikeresebbnek tartott technikáktól. Szekvenciák tanulása helyett (mint például Conditional Random Fields, Maximum Entropy Models stb. [1]) szóalapú osztályozásként kezeljük a problémát. Ilyen típusú modelleket korábban is alkalmaztak tulajdonnevek felismerésére – leggyakrabban Support Vector Machine osztályozót [8][11] –, de a szekvenciális tanulók az elmúlt években sokkal „divatosabbá" váltak.

Az általunk választott döntésifa-alapú megközelítés gyors tanítást és kiértékelést biztosít, ami lehetővé tette hatalmas jellemzőkészlet használatát. Természetesen – a szóalapú osztályozás ellenére – a környezetre, kontextusra vonatkozó információkat mi sem hagyjuk figyelmen kívül: jellemzőként beépülnek a modellbe a megelőző és rákövetkező szavak főbb tulajdonságai, valamint a megelőző szavakra a modell által javasolt tulajdonnévi címkék.

3.2 Nagyméretű tulajdonsághalmaz

Egy igen bő tulajdonsághalmazt gyűjtöttünk össze, amely leírja az egyes szavakat, illetve azok rögzített szélességű környezetét. A következő kategóriákba sorolhatjuk ezeket a jellemzőket:

- **Felszíni jellemzők**: kis/nagy kezdőbetű, szóhossz, tartalmaz-e számot, van-e nagybetű a szó belsejében, arab/római szám-e stb., illetve legyűjtöttük a tanuló halmaz legjellegzetesebb két-, hárombetűs szórészleteit.
- **Frekvenciainformációk**: token előfordulási gyakorisága (webről gyüjtött frekvenciaszótárban), kis- és nagybetűs előfordulások aránya, mondat eleji előfordulások és nagybetűs előfordulások aránya.

- **Környezeti jellemzők**: mondatbeli pozíció, megelőző szavakra modell által javasolt tulajdonnévi címke (online kiértékelés), zárójelben, idézőjelek közt van-e; a tanító halmazból legyűjtöttük, hogy a megelőző/rákövetkező szavakból melyek azok, amelyek az egyes osztályokat implikálhatják (szűrésüket a szóalakok egyes osztályok közti entrópiája alapján végeztük el).
- **Egyértelmű tulajdonnevek listája**: Felvettük egy-egy listába azokat a szavakat és többszavas kifejezéseket, amelyek a tanító halmazon legalább ötször előfordultak, és az esetek legalább 90 százalékában ugyanabba az osztályba tartoztak.
- **Tulajdonnév szótárak**: magyar és angol keresztnevek, vállalattípusok (mint pl. *kft.*, *rt.*), nagyvárosok és országok, stb. Összesen nyolc angol és négy magyar listát alkalmaztunk, amelyeket mind az Internetről töltöttünk le.

Az egyes feladatoknál felhasználtunk még néhány problémaspecifikus jellemzőt, ezeket a következő fejezetben a problémák tárgyalásánál mutatjuk be.

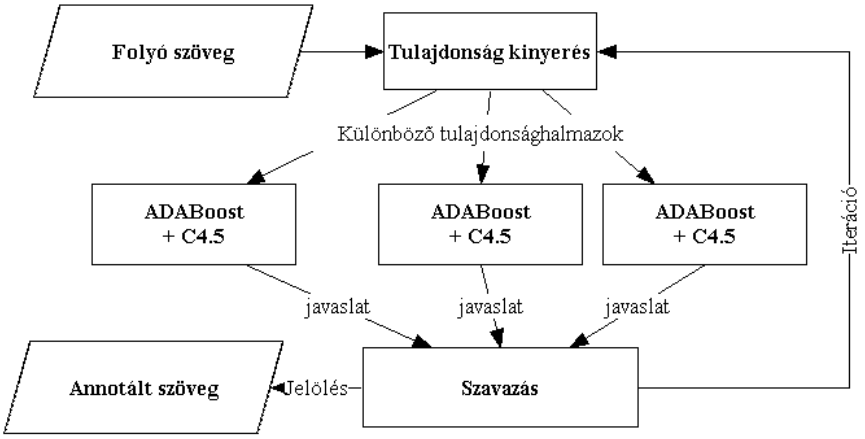
Fontos megemlítenünk, hogy mind a magyar, mind az angol nyelvű feladatoknál próbálkoztunk nyelvtani információk felhasználásával (POS kódok és chunk kódok), de ez egyik esetben sem javított értékelhetőt a modell pontosságán, sőt néhol összekavarta a rendszert, ezért teljes mellőzésük mellett döntöttünk.

3.3 A tulajdonsághalmaz felosztása és újrakombinálása

Az egyes tulajdonságok és azok halmazai más és más szemszögből közelítik/magyarázzák a problémát. Úgy gondoltuk, az általunk összegyűjtött jellemzők nagy száma lehetővé teszi, hogy részekre bontsuk a teljes halmazt úgy, hogy a részhalmazokon külön-külön tanítva is erős modellek legyenek építhetők. Ezeknek a modelleknek a kombinálása később jobb eredményre vezethet, mint az eredeti teljes halmazon tanított egyedi modell.

Hipotézisünk helyességének vizsgálatára a következő újszerű eljárást – amelyet a az 1. ábra szemléltet – dolgoztuk ki: a jellemzőket csoportosítottuk úgy, hogy egy csoportba hasonló jellegű tulajdonságok kerüljenek. Az így kialakított csoportokat – bonthatatlan egységnek tekintve – azok összes lehetséges kombinációját kiértékeltük a tanító halmazon (futásigény miatt egyszerű döntési fákat tanítottunk csak) és az öt legjobban teljesítő kombinációt tartottuk csak meg (tehát az eredeti teljes tulajdonsághalmaz öt nem diszjunkt részhalmazát).

Ezekre külön-külön tanítottunk egy-egy teljes modellt (boostingolt döntési fákat) majd a tesztelési fázisban a modellek jelölési javaslatait szavaztattuk a következő egyszerű döntési szabály segítségével: *ha van három megegyező javaslat, akkor fogadjuk azt el, ellenkező esetben adjunk nem-tulajdonnévi címkét*. Tehát ha a modellek nem képesek „egyezségre jutni” akkor inkább nem jelölünk tulajdonnevet. Ez a stratégia statisztikailag jobb eredményt ad, mintha találomra választanánk az öt javaslat közül, ugyanis a hibásan jelölt tulajdonnév két büntetőpontot von maga után (a pontosság és a fedés is sérül), míg a nem jelölt tulajdonnév csak egyet (a fedés csökken).



1. Ábra A tulajdonnév-felismerő rendszerünk sematikus váza

3.4 Iteratív tanulás

Az orvosi kórlapokon szerepelnek a – folyó szövegrészeken túl – rekordba rendezett információk is. Ezekben a tulajdonnevek felismerése és kategorizálása lényegesen egyszerűbb feladat, mint a folyó szövegekben. Ezért ennél a problémánál a következő iteratív tanítást hajtottuk végre: az első modell célja csak a rekordokban szereplő információk kinyerése volt. Az itt megtalált neveket, azok részeit (pl. személyneveknél hasznos a családi és keresztnév különválasztása) felhasználtuk egy következő tanítási fázisban, ahol már folyó szöveget is elemeztünk, de csak a biztos helyeken jelöltünk tulajdonneveket. Ezeket ismét hozzáadtuk az „ismert" egyedek listájához, amivel egy végső tanítási lépésben már az összes tulajdonnév-felismerése volt a cél.

A fent leírt módszert általánosíthatjuk: minden iterációban csak a biztos egyedeket jelöljük, hogy ezekből tulajdonságokat nyerünk ki, amelyeket a következő iterációbeli tanítás folyamán felhasználunk. Tehát egy nagyon pontos (de lefedésben gyenge) modellből kiindulva, iterálva jutunk el egy végső modellig, aminek a pontossága és fedése a kívánt szintű. Ennek az általánosított módszernek az empirikus vizsgálatát a jövőben fogjuk elvégezni.

3.5 Gépi tanuló algoritmusok

Számos klasszifikációs technikát kipróbáltunk a probléma megoldása folyamán (Neurális hálók, Support Vector Machines stb), de legjobbnak a döntésifa-alapú technikák bizonyultak [4]. Ez elsősorban annak köszönhető, hogy az ID3 algoritmus [9] diszkrét tulajdonságok – és a modellünkben túlnyomórészt diszkrétek a tulajdonságok – kezelésére lett kidolgozva (a numerikus tanulókkal ellentétben), másodsorban a fák tanításának és predikciójának sebessége kezelhetővé tette a hatalmas tulajdonsághalmazt.

Az egyszerű döntési fák által elért pontosság javítására Shapire AdaBoostM1 [10] algoritmusát használtuk. A módszer több iteráción keresztül javít egy tetszőleges, gyenge tanulót (mint például esetünkben a döntési fát) úgy, hogy minden iterációban a tanító halmazon eltévesztett minták súlyát növeli, a helyesen jelöltekét csökkenti a következő modell mintaválasztásához.

A kísérletekhez a WEKA keretrendszer [7] egy kiegészített változatát használtuk. A döntési fánál mindig az alapértelmezett paramétereket, a boostingnál 30 iterációt használtunk. A paraméterek finomhangolásával minden bizonnyal talán modelleink további javulását érhetnénk el.

4 Eredmények

A három tulajdonnév-felismerési problémát és a kapcsolódó eredményeinket (mindenhol frázisszintű $F_{\beta=1}$ metrikát használva[11]) kronológiai sorrendben mutatjuk be.

4.1 Magyar nyelvű gazdasági rövidhírek elemzése

A Szeged Korpusz 200 ezer szóból álló, gazdasági rövidhíreket tartalmazó szegmensében (NewsML) bejelöltük a *szervezet, személy, hely* és *egyéb* kategóriákba tartozó tulajdonneveket (SzegedNE korpusz). Az annotátorok közti végső egyetértési szint 99,89% lett [12]. Ez a korpusz az első magyar nyelvű tulajdonnévi korpusz[12].

A korpusz jellegzetessége a szervezetek túlsúlya a tulajdonneveken belül. Ráadásul ezek felismerése általában egyszerűbb a többi osztályhoz képest a cégformára utaló frázisvégződések (*kft., rt*. stb.) miatt. Ez magyarázza az angolra publikált eredményeknél lényegesen jobb eredményeinket.

1. Táblázat: Eredmények a SzegedNE korpuszon

	$F_{\beta=1}$
Szervezet	95,84%
Személy	94,67%
Hely	95,07%
Egyéb	85,96%
Mindösszesen	94,77%

Az előző fejezetben bemutatott alaprendszer – a korábban még nem látott szövegrészleteken mért – osztályszintű eredményeit az 1. Táblázat tartalmazza [13]. Sajnos ezen eredményeinket nem tudjuk más rendszerekkel összehasonlítani, ez idáig csak szabályalapú rendszerek készültek magyar nyelvű tulajdonnevek felismerésére [6], a mi rendszerünk az első statisztikai modell.

[11] A kiértékelést a CoNNL versenyhez kiadott szkripttel végeztük, ami letölthető a http://www.cnts.ua.ac.be/conll2003/ner/ oldalról

[12] A korpusz kutatási célokra ingyenesen hozzáférhető a http://www.inf.u-szeged.hu/~hlt/index.html oldalon

4.2 Angol nyelvű újságcikkek elemzése

A fent említett négy tulajdonnévi kategória felismerése volt a célpontja a CoNLL által kiírt nyílt versenynek 2003-ban [15]. A tanító adatbázis Reuters híreket[13] tartalmazott 1996-ból, amelyek felöleltek sport, politikai és gazdasági témákat egyaránt.

A korpuszon a *szervezet* kategória pontossága szignifikánsan rosszabb, mint a magyar szövegen volt (mint ahogy az a 2. Táblázatból is leolvasható). Ez annak köszönhető, hogy a sporthírekben a csapatnevek (amik természetesen *szervezet*ek) illetve a városok nevei nagyon nehezen különböztethetőek meg (pl. *Los Angeles beat Boston*).

A feladat specialitásait kihasználandó az alap tulajdonsághalmazt két további tulajdonsággal bővítettük: először is a legtöbb hír három jól elkülönülő részre volt bontható (cím/rövid összefoglaló a cikk elején; riporter, helyszín, dátum; maga az újsághír), másrészt a Reuters témakódok alá sorolja a híreket, ezeket a kódokat is felvettük külön jellemzőként. Ez utóbbitól reméltük a fent említett város-csapat többértelműség feloldását. Sajnos ebben csalódnunk kellett.

2. Táblázat: Eredmények a CoNLL-2003 adatbázison

	egyéni	hibrid
Szervezet	84,53%	88,32%
Személy	93,55%	96,27%
Hely	92,90%	93,43%
Egyéb	79,67%	82,29%
Mindösszesen	89,02%	91,41%

A versenyen győztes egyéni modellhez [5] képest rendszerünk 2,3%-kal kisebb relatív hibával működik a kiértékelési adatbázison, de igazi haszna hibrid rendszerekben mutatkozik meg: mivel modellünk más megközelítésen alapszik, mint a versenyen induló modellek, eredményeinket kombinálva az ott szereplő legjobb rendszerekkel lényegi javulás érhető el. A versenyen szereplő három legjobb modell többségi szavazásos kombinációja 90,3%-os eredményt hozott. Ha a mi rendszerünk lép a győztes modell helyébe a szavazás utáni eredmény 91,41%, ami már számottevő, 11,44%-os relatív hibacsökkenést jelent [13].

4.3 Orvosi kórlapok anonimizálása

Az orvosi szakszövegek adatbányászati célú felhasználásához elengedhetetlen az abban szereplő személyes adatok védelmének biztosítása. Ezért mielőtt publikussá válik egy orvosi szövegekből álló adatbázis, az abban előforduló személyek neveit (orvos, páciens), telefonszámát, lakhelyét, a kórház nevét stb. anonimizálni kell. A feladat tehát itt is tulajdonnevek bizonyos jól körülhatárolt osztályainak felismerése és kategóriákba sorolása.

Erre az információkinyerési feladatra az MIT Computer Science and Artificial Intelligence Laboratory, Informatics for Integrating Biology and the Bedside (i2b2) kutatóintézet nyílt versenyt írt ki idén nyáron [11], amelyre beneveztük rendszerünket

[13] http://www.reuters.com/researchandstandards/

is. A szervezők biztosítottak egy 200 ezer szavas annotált tanító adathalmazt, majd egy körülbelül 50 ezer szavas halmazon értékelték ki a rendszereket.

Az orvosi kórlapok – melyek szerkezete ugyan nem kötött, de – tartalmaznak strukturált, rekordokba rendezett egységeket. A rekordok határait és azok belső szerkezetét egyszerű szabályokkal azonítottuk, és a fejléceket új tulajdonságként hozzátettük az alap jellemzőkészlethez (ezeket a fejléceket használtuk fel a 3.4 fejezetben bemutatott iteratív tanulás első iterációjában is). A másik tulajdonság, amivel a halmazt bővítettük az a könnyebben felismerhető osztályokra (*dátum, életkor, telefonszám, azonosítók*) felírt reguláris kifejezések voltak.

Az egyik legfontosabb jellemzőnek a szó környezetének – implikáló – szóalakjai bizonyultak. Arra, hogy a környezetet pontosan hogyan használjuk fel, három különböző módszert dolgoztunk ki. Ennél a feladatnál az idő rövidsége miatt nem hajtottuk végre a teljes tulajdonsághalmaz-felbontást és újrakombinálást (amit részletesen a 3.3. fejezetben mutattunk be), helyette az erre a három módszerre épített modelleket kombináltuk többségi szavazással.

3. Táblázat: Eredmények az i2b2 adatbázison

	$F_{\beta=1}$
Kórház	92,69%
Doktor	95,88%
Páciens	96,21%
Hely	63,79%
Életkor	100,00%
Dátum	99,25%
Azonosító	99,33%
Telefonszám	98,31%
Mindösszesen	97,41%

A 3. Táblázat tartalmazza a teszthalmazon, az egyes osztályokon elért eredményeinket. Ahogy azt vártuk, a négy egyszerűen felismerhető osztály pontossága látványosan jobb, mint a klasszikus tulajdonnévosztályoké. A *hely* kategória erősen alulreprezentált, a tanító halmazban mindössze 150 hely kifejezés szerepelt, ez magyarázza az igen gyenge felismerési eredményét. A másik három „érdekes” osztályon elért eredményeink összehasonlíthatók a korábbi, újsághírekkel foglalkozó feladatok hasonló osztályain (*személy*, *szervezet*) elért pontosságokkal [14].

A versenyen 16 rendszer vett részt. A testre szabott általános tulajdonnév-felismerő rendszerünkkel elért 97,41%-os pontossággal a versenyen első helyezést értünk el.

4.3 Az egyes feladatok modelljei közti eltérések

E fejezetben bemutattuk, hogyan alkalmaztuk tulajdonnév-felismerő rendszerünket három különböző probléma megoldására. A magyar nyelvű gazdasági hírekre fejlesztett modell angolra adaptálásakor csak a szótárakat cseréltük le azok angol nyelvű megfelelőire, valamint felhasználtuk az újsághírek speciális tulajdonságait (témakód, dokumentumon belüli rész).

A kórlapok anonimizálásánál szintén kihasználtuk a dokumentumok szerkezetét (rekord fejlécei és reguláris kifejezések), azon felül csak a cégvégződés listát (*ltd.* stb.) cseréltük le kórháznévvégződés-listára (mint pl. *Hospital*). Minden más tekintetben ugyanazokat a tulajdonságokat, ugyanazokat a tanulókat, paraméterbeállításokat és technikákat használtunk mindhárom statisztikai modell építése folyamán.

5 Diszkusszió

Az előző fejezetekben bemutattuk statisztikai tulajdonnév-felismerő rendszerünket, amelyet sikeresen alkalmaztunk magyar nyelvű gazdasági rövidhírekben található tulajdonnevek felismerésére és kategorizálására. A kisebb változtatásokon keresztülment modellel minden korábbinál jobb eredményt értünk el a standard angol nyelvű tulajdonnévfelismerési adatbázison (CoNLL), és megnyertük az i2b2 orvosi kórlapok anonimizálására kiírt nemzetközi nyílt versenyt is.

Rendszerünk sikerét elsősorban az összegyűjtött nagyméretű tulajdonsághalmaznak és abban rejlő potenciálok hatékony kiaknázásának (tulajdonságmegosztás, majd rekombináció, jól megválasztott tanuló modell) köszönheti. Szeretnénk még egyszer kiemelni, hogy az elemzéshez csak felszíni jegyeket, illetve a tanító adatbázisból kinyerhető statisztikai jellemzőket használtunk fel. A rendszer nem függ semmilyen külső modelltől – mint például POS-tagger – és nincs szüksége semmilyen nyelv-, illetve domainfüggő szakértői tudásra (az i2b2 versenyen például számos más rendszer használta az orvosi *Medical Subject Headings* kódokat).

Természetesen – mint minden induktív tanulási modell – rendszerünk csak akkor alkalmazható, ha rendelkezésre áll megfelelő méretű tanító adatbázis (mindhárom esetben a körülbelül 200 ezer szavas halmaz kielégítőnek bizonyult), és a jelölendő szöveg főbb jegyeiben megegyezik a tanító halmazzal.

A jövőben ezért szeretnénk megvizsgálni, hogy milyen lehetőségeink vannak, ha nem áll rendelkezésre elégséges méretű előre bejelölt példákat tartalmazó adatbázis (ugyanis annak előállítása általában igen költséges). Ezen az úton meg is tettük az első lépéseket, folyamatban vannak olyan kísérletek, amelyekben azt vizsgáljuk, hogy jelöletlen szövegek, illetve internetes keresőmotorok hogyan segíthetik a jelölt szövegeken tanuló modelleket.

Bibliográfia

1. Hai L. Chieu and Hwee T. Ng.: Named Entity Recognition with a Maximum Entropy Approach. Proceedings of CoNLL-2003 (2003)
2. Nancy Chinchor.: MUC-7 Named Entity Task Definition. Proceedings of Seventh Message Understanding Conference (1998)
3. Dóra Csendes, János Csirik and Tibor Gyimóthy: The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus. Proceedings of TSD 2004, vol. 3206 (2004)
4. Richárd Farkas, György Szarvas, András Kocsor: Named Entity Recognition for Hungarian using various Machine Learning Algorithms. Acta Cybernetica (2006)

5. Radu Florian, Abe Ittycheriah, Hongyan Jing and Tong Zhang: Named Entity Recognition through Classifier Combination. Proceedings of CoNLL-2003 (2003)
6. Kata Gábor, Enikő Héja, Ágnes Mészáros, Bálint Sass: Nyílt tokenosztályok reprezentációjának technológiája. IKTA-00037/2002, Budapest, Hungary (2002)
7. S. Garner. Weka: The waikato environment for knowledge analysis. (1995)
8. C. Lee, W.-J. Hou and Chen, H.-H. Annotating multiple types of biomedical entities: A single word classification approach. In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004).
9. Ross Quinlan: C4.5: Programs for machine learning, Morgan Kaufmann (1993)
10. Rob E. Shapire: The Strength of Weak Learnability. Machine Learnings, Vol. 5 197-227 (1990)
11. Tawanda Sibanda, Ozlem Uzuner, and Ozlem Uzuner. Role of local context in automatic deidentification of ungrammatical, fragmented text. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, pp 65-73, New York City, USA, June (2006)
12. György Szarvas, Richárd Farkas, László Felföldi, András Kocsor, János Csirik: A highly accurate Named Entity corpus for Hungarian. Proceedings of International Conference on Language Resources and Evaluation (2006)
13. György Szarvas, Richárd Farkas, and András Kocsor: A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. DS2006, LNAI 4265, pp. 267-278 (2006)
14. György Szarvas, Richárd Farkas, Szilárd Iván, András Kocsor, Róbert Busa-Fekete: An Iterative Method for the De-identifcation of Structured Medical Text. In Proceedings of American Medical Informatics Association, (2006)
15. Erik F. Tjong Kim Sang, and Fien De Meulder: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Proceedings of CoNLL-2003 (2003)
16. Y. Tsuruoka J-D. Kim, T. Ohta and Y. Tateisi. Introduction to the bio-entity recognition task at jnlpba. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA),Geneva, Switzerland, 2004.