Részben felügyelt tanulási módszerek a tulajdonnév felismerésben

Farkas Richárd[1]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
rfarkas@inf.u-szeged.hu

Kivonat: Az általános gépi tanulás egyik paradigmája a részben felügyelt tanulás az elmúlt években ismét előtérbe került. A módszer célja a jelöletlen adatban rejlő összefüggések kihasználásával javítani a pusztán jelölt adatokat használó tanuló algoritmusokon. Más szemszögből ezen technikák alkalmazásával kevesebb annotált adattal, így kevesebb emberi élőmunka felhasználásával ugyanolyan (vagy közel ugyanolyan) pontosságú modellek építhetőek, mint a nagyobb jelölt adatbázist használó modellekkel. A számítógépes nyelvészet statisztikai megközelítéseiben a megfelelő méretű és minőségű annotált tanítókorpuszok megléte alapfeltétel. Cikkünkkel arra szeretnénk felhívni a figyelmet, hogy a részben felügyelt technikák alkalmazása mellett jelentősen kisebb méretű korpuszok kézi annotálása is elégséges. Ezt az állítást empírikusan is alátámasztjuk magyar és angol tulajdonnév-felismerési korpuszok felhasználásával.

1 Bevezetés

Korábbi munkánkban [5] bemutattuk korpusz alapú tulajdonnév-felismerő modulunkat amelyet három különböző adatbázison (amelyek mind nyelvét, mind témáját tekintve különböztek) értékeltünk ki. Akkori munkákból azt a következtetést vontuk le, hogy ha rendelkezésre áll megfelelő méretű, manuálisan jelölt tanító adatbázis akkor gépi tanulási módszerek felhasználásával igen jó pontosságú automatikus taggelő rendszert lehet építeni.

Azonban minden egyes új tématerület új, kézileg jelölt adatbázis építését követeli meg, ami igen költséges és időigényes feladat. Ezzel szemben az esetek többségében rendelkezésre áll nagy mennyiségű, de címkézetlen szöveg, gondoljunk csak az Internetre. A felügyelt és felügyelet nélküli gépi tanulási paradigmák közt félúton helyezkednek el a részben felügyelt tanulási technikák (semi-supervised learning) [24] amelyeknek a célja a jelöletlen adatban rejlő információ felhasználása a jelölt adatok alapján történő modell-építés folyamán.

Részben felügyelt technikákkal számos általános gépi tanulási feladatban sikerült a felügyelt modellen jelentékenyen javítani (például [14]), valamint nemzetközi szinten megtörtént ezek adaptációja a számítógépes nyelvészeti problémákra [16][18]. Cikkünk elsődleges küldetése, hogy felhívja a figyelmet ezen technikák hasznosságára. Munkánk folyamán magyar illetve angol tulajdonnév korpuszokon teszteltünk néhány, a gépi tanulásban sztenderdnek számító technikát illetve tárgyaljuk a számító-

gépes nyelvészeti problémák sajátosságait kiaknázó technikák lehetőségét, majd ezek közül empirikusan is tesztelünk néhányat.

2 Részben felügyelt tanulási megközelítések

Az alábbiakban ismertetjük a főbb általános részben felügyelt tanulási megközelítéseket, majd feszegetjük azt a kérdést, hogy milyen számítógépes nyelvészet specifikus módszereket lehet érdemes használni.

2.1 Részben felügyelt tanulási technikák

A részben felügyelt tanulásnál rendelkezésre áll egy kis méretű jelölt adathalmaz és egy nagy méretű de jelöletlen adathalmaz. A cél jelöletlen adatokból nyerhető mintázatok hasznosítása a címkézett adatbázison történő modell-építés folyamán, azaz jobb modell építése. Az általános gépi tanulásban használt részben felügyelt technikákat három nagy csoportba sorolhatjuk. Ezt a három megközelítést nagyon röviden bemutatjuk ebben a fejezetben, részletesebb tárgyalás található például [24]-ben.

A legkorábbi megközelítések az ún. **generatív módszerek**kel kapcsolatosan születtek [1]. A generatív modellek (pl. Hidden Markov Models [11]) a címkézés feltételes valószínűségét közvetlenül próbálják meg leírni feltéve az inputot (esetünkben szavakat és jellemzőiket). A 'közvetlenül' azt jelenti, hogy feltételezünk valamilyen eloszlást, ami általában több elemi eloszlás kombinációja (mixture model), és annak paramétereit a tanító adatbázisból becsüljük. A nagy mennyiségű jelöletlen adat segítségével az összetett eloszlás komponensei meghatározhatóak. A generatív modellek közé sorolhatjuk a klaszterezés alapú módszereket is (ahol először a jelöletlen adatot klaszterezzük, az egyes klaszterekhez címkét rendelünk és ezzel bővítjük az eredeti adatbázist) hiszen minden klaszterező algoritmus csak akkor működhet helyesen, ha felfedezi a mintát generáló eloszlást. A generatív modellek gyakorlati használhatóságát éppen az a tény teszi nehézkessé, hogy ismernünk kell a input adat eloszlását [3].

Az ún. **bootstrapping** megközelítések nem feltételezik speciális generatív módszer alkalmazását, esetükben tetszőleges felügyelt gépi tanulási algoritmus alkalmazható. Itt a tanító adatbázist automatikusan címkézett egyedekkel (iteratívan) bővítjük. Az ön-tanulás (self-training) [23] folyamán egy gépi tanulási módszer a címkézett adatbázis alapján épített modell segítségével felcímkézi a jelöletlen adathalmazt, majd a legmegbízhatóbb, automatikusan jelölt adatokkal bővíti a tanító adatbázist és megismétli a modellépítési műveletet. Az együtt-tanulásban (co-training) két (vagy több) osztályozó címkézi fel a jelöletlen halmazt és a megbízható, automatikusan címkézett egyedekkel egymás címkézett adatbázisát bővítik (ezen keresztül „tanítják egymást"). Az osztályozók származhatnak különböző algoritmus osztályokból [7], de használhatjuk ugyanazt a módszert is különböző jellemző-készlettel [13].

A legfiatalabb részben felügyelt tanulási módszerek a **vágás az alacsony sűrűségű területeken** (low density separation) irányelvet követik. Ezeknél a módszereknél a kiértékelő adatbázist is felhasználjuk, mint jelöletlen adat. A cél tulajdonképpen ezeknek az adatoknak a jelölése (transzduktív megközelítés) és nem az új, ismeretlen

példákon predikáló modell fejlesztése (induktív tanulás). A legismertebb ilyen módszer a Transductive Support Vector Machine (TSVM) [19] – az SVM egy kiterjesztése - estében a jelöletlen pontokat felhasználjuk a kernel térbeli maximális margójú vágás megtalálására (elkerüljük a vágással azokat a régiókat ahol a jelöletlen pontok sűrűsége nagy).

A gráf-alapú megközelítések az utóbbi években kerültek előtérbe [2]. Itt a címkézett és címkézetlen egyedeket (beleértve a kiértékelési adatbázist is) egy gráf csúcsainak képzeljük el, ahol két csúcs közti él súlya a két egyed hasonlóságával arányos (a gyakorlatban csak a legközelebbi szomszédokat kötjük össze éllel). Egy megfelelő hasonlósági metrika és a gráf vizsgálatával könnyen számolhatunk lokális sűrűségi értékeket a gráfban. A gráfban ezek után az alacsony sűrűségű helyeken (élek mentén) kell a vágást elvégeznünk. A vágás után kapott klaszterekben a kiértékelendő egyedeket a klaszterben szereplő címkézett csúcsoknak megfelelően jelöljük.

Ezek az alacsony sűrűségű régiókban vágó módszerek elméletileg jól alátámasztottak, azonban a gyakorlatban csak kis adatbázisokra alkalmazhatóak (mind tár, mind időigényük igen magas). Még azok a letölthető megoldások is amelyek magukat nagy adatbázisokon működőnek írják le (large scale solutions) nem adnak megoldást 100 jellemző mellett 30 ezer egyedre egy héten belül[1], pedig a tulajdonnév felismerési problémában 300 jellemzővel és 3 millió egyeddel kell dolgoznunk.

2.2 Részben felügyelt tanulás a számítógépes nyelvészetben

Ha számítógépes nyelvészeti problémákra fókuszálunk lehetőség nyílik azok specifikumainak kiaknázására. Úgy gondoljuk, hogy ez a terület még nincs kielégítően körüljárva. Itt mindössze a számítógépes nyelvészeti problémák két speciális tulajdonságát tárgyaljuk röviden, melyek kiaknázásas nem lehetséges a sztenderd részben felügyelt technikákkal.

A természetes nyelv szekvenciális tulajdonsága lehetővé teszi, hogy összetettebb statisztikákat (szabályosságokat) fedezzünk fel a jelöletlen szövegekben. Ilyen statisztikák a szó és karakter n-grammok, szógyakoriságok (ahol megkülönbözethetünk kis és nagy kezdőbetűs vagy mondat eleji előfordulásokat is [8]) valamint a nyelvmodellek, amibe beleértünk minden olyan modellt ami a nyelv szabályosságait valószínűségi alapon próbálja meg modellezni (tehát nem csak a szűk értelemben vett $P(w_t|w_{t-1})$ feltételes valószínűséggel leírható nyelvi modellt). Az ilyen jellegű statisztikákat felhasználhatjuk a felügyelt tanulási modell jellemző-terének konstrukciójának folyamán.

Egy másik érdekes tulajdonsága az emberi szövegekkel kapcsolatos problémáknak, hogy az Interneten szinte korlátlan mennyiségben fordul elő folyó szöveges információ. A World Wide Web-nek, mint jelöletlen korpuszt azonban nem kezelhetjük ugyanúgy, mint az egyéb offline korpuszokat (nem tudjuk például egy szó összes előfordulásán végigiterálni), azt csak a kereső-motorok (pl. Google, Yahoo) segítsé-

[1] Két ilyen programcsomagot töltöttünk le és teszteltünk:
http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html és
http://www.learning-from-data.com/te-ming/semil.htm

gével tehetjük meg effektíven. Ehhez meg kell fogalmaznunk kéréseket (query) majd a találatul kapott oldalakat letölthetjük (és feldolgozhatjuk), de a találatok becsült száma és a keresőszavak alapján relevánsnak vélt szövegkörnyezetek (snippet) is hordoznak igen hasznos információkat. Létezik már néhány megoldás egyszerű számítógépes nyelvészeti problémára melyek hasznosítják a WWW-et [15][18]. A tulajdonnév-felismerés problémájához legközelebb álló ilyen megoldások egy bizonyos név-osztályba tartozó listákat próbálnak az Internetről automatikusan összegyűjteni (ilyen például a Google Sets szolgáltatás vagy [4][17]). Úgy gondoljuk , hogy a jövőben ez a terület jóval nagyobb figyelmet fog kapni és egyre mélyebb elemzést végző alkalmazások is inputként fogják kihasználni az WWW-et.

3 Empirikus eredmények

Az előző részben bemutatott technikák közül az ön-tanulás, az együtt-tanulás és néhány Web alapú módszert magyar és angol tulajdonnév-felismerési adatbázisokon teszteltük. A kísérletek paramétereit, valamint az elért eredményeket mutatjuk be ebben a fejezetben.

3.1 Magyar és angol tulajdonnév-felismerési adatbázisok

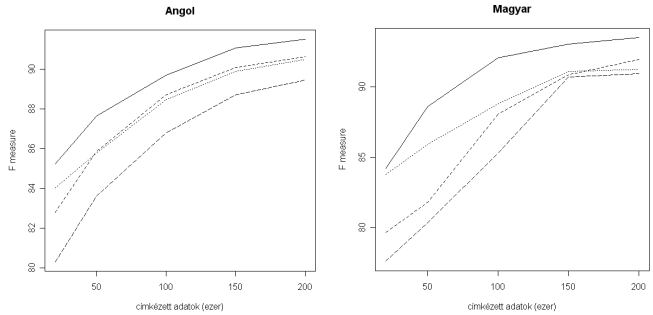
A tulajdonnevek azonosítása (és kategorizálása) folyó szövegben meghatározó fontosságú számos számítógépes nyelvfeldolgozó alkalmazás során. Példaként tekinthetjük a különböző információkinyerő rendszereket, ahol a tulajdonnevek általában jelentős, információt hordozó szerepet töltenek be a szövegben, vagy a gépi fordítási alkalmazásokat, ahol értelemszerűen más módon kell kezelni emberek, szervezetek neveit, mint a szöveg többi részét. Magyar és angol nyelvre a gazdasági témájú tulajdonnév-felismerési feladaton teszteltük a részben felügyelt tanulási modelleket. Itt a feladat egy szöveg minden egyes szavához a *szervezet/helység/személy/egyéb/nem-tulajdonnév* címkék valamelyikének hozzárendelése.

A CoNLL által kiírt nyílt versenynek 2003-ban [22] volt feladata ez a típusú klasszifikáció. Az adatbázis Reuters híreket[2] tartalmazott 1996-ból, amelyek felöleltek sport, politikai és gazdasági témákat egyaránt. Az akkori verseny szervezői azt szerették volna elérni, hogy a rendszerek hasznosítsák a jelöletlen adatból nyerhető információt is, ezért a címkézett adatbázisok mellé és egy közel 18 millió szavas jelöletlen korpuszt (szintén Reuters hírek) is elérhetővé tettek. Ezt a jelöletlen adathalmazt használtuk fel mi is kísérleteink folyamán. A 2003-as versenyre nem küldtek be olyan rendszert ami a címkézetlen adatokat hasznosította volna. A CoNLL testB adatbázisa hordoz néhány olyan speciális tulajdonságot ami a tanító adatbázistól megkülönbözteti, ezért ebben a munkánkban a testA halmazra közlünk eredményeket.

Magyar nyelvre a Szeged Korpusz 200 ezer szóból álló, gazdasági rövidhíreket tartalmazó szegmensét (SzegedNE korpusz) [21] használtuk, mint címkézett és kiértékelési adatbázis. Jelöletlen adatbázisnak gazdasági témájú újsághíreket (nem rövid-

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

híreket!) próbáltunk meg alkalmazni, azonban ez nem vezetett eredményre. A magyar adatbázisra közölt eredményeink a transzduktív megközelítést követték, azaz a kiértékelési adatbázist használtuk fel, mint címkézetlen korpusz.



1. ábra: Felügyelt tanulási modell által elért eredmények különböző jellemzőterek használata mellett a címkézett korpusz méretének függvényében

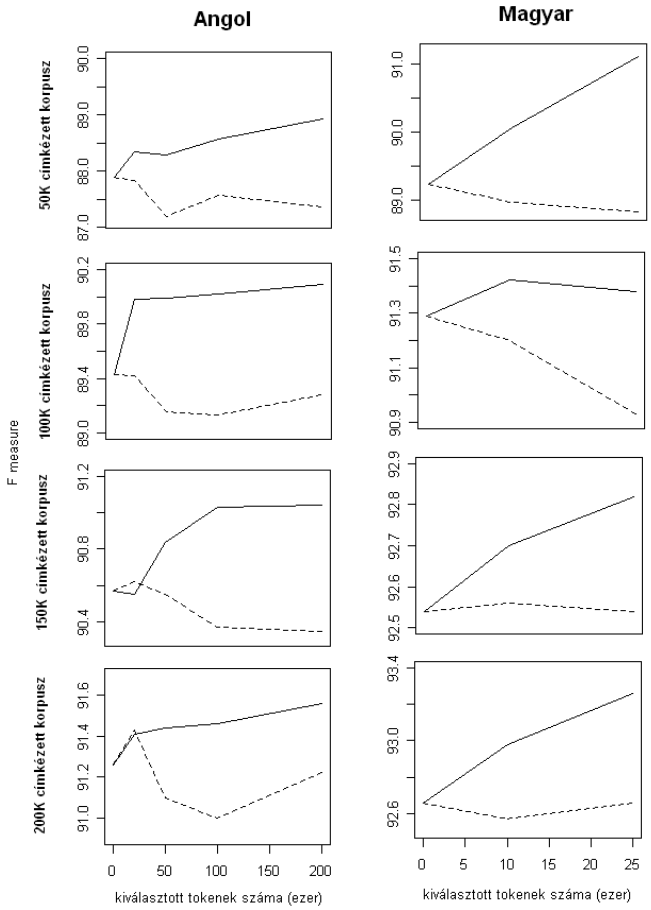
A modellépítés folyamán felhasznált jellemzőkészlet igen változatos volt [20]. A következő kategóriákba sorolhatjuk a jellemzőket (a magyar és angol adatbázison ugyanazokat a jellemzőket használtuk fel):

- 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, 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.
- 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.

3.2 A jelöletlen korpuszból származtatott jellemzők hozzáadott értéke

A kísérletekben elsősorban a Conditional Random Fields (CRF) [10] nevű osztályozó algoritmusra (MALLET implementáció [12]) támaszkodtunk, ami számos szekvenciális jelölési probléma megoldásában bizonyított és az utóbbi néhány évben a state-of-the-art-nak számít. Első kísérleteinkben a címkézett adatbázis méretének hatásait és a jelöletlen korpuszokból származtatott jellemzők hozzáadott értékét vizsgáltuk meg.

Ilyen címkézetlen korpuszból származó jellemzőcsoport a frekvenciainformációk és a szótárak. Előbbit több milliárd szavas Webkorpuszokból számítják ki. Angolra a Gigaword korpuszt, míg magyarra a Szószablya Gyakorisági Szótárat [8] használtuk. A tulajdonnév szótárak szintén az Internetről összegyűjthető listák. Egyes kategóriákhoz (tulajdonnév osztályokhoz) tartozó listákat gyűjthetünk automatikusan, keresőmo-



2. ábra: Együtt-tanulás (folytonos) és ön-tanulás (szaggatott) a jelölt és jelöletlen adatbázis méretének függvényében

torok és egyszerű szintaktikai keretek illesztésével [4], de a legalapvetőbb listák összeállítva elérhetőek, azokat legfeljebb csak szűrni és normalizálni kell. Az itt használt listákat az utóbbi módon gyűjtöttük, körülbelül egy embernapnyi ráfordítással.

Az 1. ábrán látható 4-4 görbe a teljes jellemzőtér (folytonos vonal), a frekvencia típusú jellemzők (pontozott vonal), a szótárak (szaggatott vonal) illetve mindkét jellemzőcsoport mellőzésével („teli-üres” vonal) nyert eredményeket mutatják a címké-

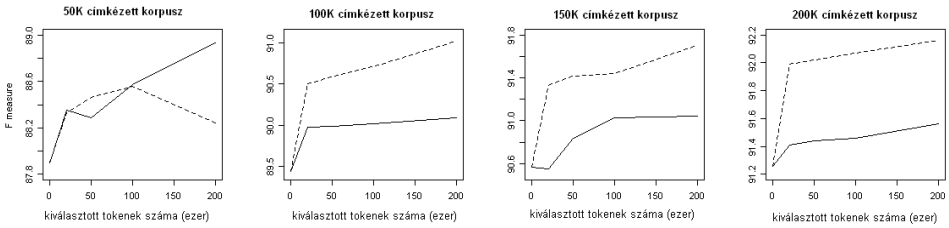
zett tanító adatbázis méretének függvényében. Az a tendencia egyértelműen megfigyelhető, hogy a szótárak megvonását egyre kevésbé érzi meg a modell ha nő a tanító adatbázis. Tehát a szótárak kis méretű adatbázisnál rendelkeznek komoly hozzáadott értékkel, nagyobb halmazokon ugyanezen információt meg tudják szerezni a statisztikai modellek a címkézet adatból is. A frekvencia jellemzők alkalmazásával átlagosan a hibák 19% eliminálhatóak, míg a szótárakkal 15%, együttes alkalmazásukkal 28%.

3.3 Bootstrapping módszerek

Az ön-tanulási és az együtt-tanulási algoritmusokat szimultán vizsgáltuk. Az öntanulásánál a CRF önmagát tanította a jelöletlen adatokkal, míg az együtt-tanulásnál egyetlen iterációban a korábbi munkáinkban alkalmazott és jó eredményeket elérő AdaBoostM1+C4.5 modellünk [20] által jelölt szövegekkel bővítettük a CRF tanító adatbázisát. Az alkalmazott két osztályozási modell teljesen másképp közelíti meg a címkézési problémát. Egyrészt a döntési fa alapú módszer az egyes szavakat egymástól függetlennek tekinti (a környezetre vonatkozó információk a jellemzőtérbe vannak beépítve), míg a CRF az egész mondatot (szekvenciát) egyben jelöli be. Másrészt a CRF az egyes jellemzők együttes eloszlása felett épít exponenciális modellt (logisztikus regresszió), míg a C4.5 algoritmus mohó módon választ minden lépésben egy jellemzőt információ elméleti metrikák alapján. Ezt a diverzitást tudja együtttanulás kihasználni.

A 2. ábrán szaggatott vonallal jelöltük az ön-tanulással, míg folyamatos vonallal az együtt-tanulás által elért eredményeket különböző méretű címkézett adatbázis mellett a jelöletlen adatbázis méretének függvényében (az x tengely 0 pontjánál lévő függvényértékek megegyeznek az 1. ábra értékeivel). Láthatjuk, hogy az együtt-tanulással minden esetben tudtunk javítani a felügyelt modellhez képest, míg az ön-tanulással nyert automatikusan címkézett példák csak összezavarták a modell-építést.

Az eredményeket javíthatjuk, ha nem minden automatikusan címkézett mondatot adunk hozzá a tanító adatbázishoz, hanem azok közül csak a legmegbízhatóbbakat. A 3. ábrán láthatjuk az együtt-tanulással elért eredményeket miután a döntési fa által jelölt mondatok közül csak azokat használjuk fel amelyek bizonytalansága kisebb, mint 10^{-3} illetve 10^{-10}. Természetesen minél alacsonyabb ez a küszöbérték annál több jelöletlen adat felhasználására van szükség, hogy szignifikáns módon bővíthessük a címkézett adatbázist. A 10^{-10} küszöb mellett 3 millió szövegszónyi nyers szöveget használtunk fel, 23407 „megbízható" mondatot kiválasztva. Együtt-tanulással sikerült hajszállal jobb eredményeket elérnünk 100 ezer jelölt adat felhasználásával (91,28% F érték), mint 200 ezer jelölt adattal jelöletlen adatok nélkül (91,26% F érték).



3. ábra: Performancia a jelölt és jelöletlen adatbázisok méretének függvényében különböző kiválasztási küszöbértékek mellett

A szép eredmények ellenére el kell mondanunk, hogy magyarra is hajtottunk végre jelöletlen gazdasági szövegek felhasználásával (a transzduktív megközelítés helyett) ön- illetve együtt-tanulást. Azonban ezzel – a CoNLL B adatbázisához hasonlóan – nem sikerült szignifikáns javítást elérnünk a felügyelt modellhez viszonyítva. Ez minden bizonnyal a jelöletlen adathalmaz és a kiértékelési adatbázis eltérő jellegéből fakad (hasonló következtetést vont le [9] is).

3.4 A WWW hasznosítása mint külső szakértői tudásbázis

Számos lehetőség kínálkozik arra, hogy az Internetről információt gyűjtsünk számítógépes nyelvészeti problémák megoldásához. [6] munkánkban három heurisztikát mutattunk be melyekkel az angol tulajdonnév-felismerő rendszerünk hibáit próbáltuk meg eliminálni a GoogleAPI és a Wikipedia felhasználásával.

A tulajdonnév-felismerő rendszerek hibáinak egy szignifikáns része abból fakad, hogy a rendszer nem jól találja meg a hosszabb frázisok határát (elejét vagy végét). Ezért megvizsgáltunk minden olyan egyedet ahol a címkézett frázis előtt (vagy után) közvetlenül nagybetűs szó állt vagy legfeljebb két stopword ékelődött nagy kezdőbetű szó és a jelölt egyed közé. A hipotézisünk az volt, hogy ha az ilyen módon kiterjesztett tulajdonnév előfordulási gyakorisága összemérhető még az eredeti jelöltével akkor a kiterjesztés végrehajtandó. Ennek eldöntésére Google keresést hajtottunk végre a címkézett tulajdonnévre és a kiterjesztett frázisra és ha a találatok számának aránya 0,1%[3] felett volt elfogadtuk a kiterjesztést.

A második heurisztikát arra a hipotézisre építettük, hogy a tulajdonnevek leggyakoribb jelentése (osztálya) statisztikailag hasznos információ. Ezért ha a rendszer nem tudott megbízható döntést hozni egy felismert tulajdonnév osztályáról akkor az Interneten megkerestük annak leggyakoribb szerepét és azt adtuk a frázis címkéjének. Módszerünk néhány kérdést küldött minden tulajdonnévhez (1. táblázat), hogy annak kategóriáját megtudjuk. Ezeket a Google snippetjeinek elemzéséből (főnévi csoportok azonosítása) nyertük ki.

1. táblázat: Felhasznált kereső-kifejezések

Angol	Magyar
NP such as NE	NE egyike NP
NP including NE	NE és más NP
NP especially NE	NE és egyéb NP
NE is a NP	NE vagy más NP
NE is the NP	NE vagy egyéb NP
NE and other NP	
NE or other NP	

[3] Ezt az értéket nem a kiértékelési adatbázis, hanem a fejlesztési adatbázis alapján határoztuk meg, az előbbi így ismeretlen maradt.

Azt hogy melyik kategória melyik tulajdonnév osztályba tartozik a tanító adatbázison egyértelmű egyedekre futtatott ugyanezen Google keresések eredményéből nyertük ki. Azt, hogy a címkézés megbízható-e különböző algoritmusok egyetértési rátájával mértük (comitee based learning).

A harmadik heurisztikánál az egymást követő, azonos típusú tulajdonnevek (pl. „Taleban | Míg-19") problémáját igyekeztünk orvosolni (ilyen esetben a tulajdonnevek határát B- kezdetű címke jelöli). A legtöbb ilyen esetben valamilyen írásjel választja el az egyedeket azonban például beszédfelismerés eredményeként előálló szövegben ezek nincsenek jelen. Az ilyen esetek kiszűrésére a Wikipediát alkalmaztuk: minden legalább kettő hosszúságú frázisra megnéztük, hogy létezik-e a nevet egy az egyben lefedő Wikipedia oldal, és ha létezett írásjelek nélkül elfogadtuk azt egy tulajdonnévnek. Ellenkező esetben a jelölt frázis darabjaira kerestünk, illetve elvégeztük az írásjelek menti vágást. Ily módon sikerült például elválasztanunk a „Golan Heights | Isreal"-t.

A [6] publikáció empirikus eredményei bizonyítják a WWW-ről, mint külső tudásbázisból, inputként gyűjtött információ felhasználásának hasznát. Azonban ezek magyarra történő adaptálásánál problémákba ütköztünk. Először is nem tudtunk minden kérdést lefordítani (a létigét a magyarban nem tesszük egyes szám harmadik személyben), újakat kellett kigondolnunk. De az igazán komoly gondot az okozta, hogy a kérdések mintegy 70%-ára nem érkezett Google találat vagy nem létezett Wikipedia oldal. A magyar Web (a *site:.hu* kifejezést használtuk) és a magyar Wikipedia (aminek mérete az angolhoz viszonyítva 3,5%) nem elég nagy ilyen jellegű feldolgozáshoz.

4. Összegzés

Ennek a cikknek az elsődleges küldetése az volt, hogy rávilágítson a jelöletlen korpuszok felhasználásában (részben felügyelt tanulási modellek) rejlő potenciálra. Eredményeket magyar és angol tulajdonnév-felismerési problémákra közöltünk. A különböző elméleti bázissal rendelkező felügyelt tanulók együttes-tanulásával sikerült 100 ezer szónyi címkézet szöveg és 3 millió szónyi jelöletlen korpusz felhasználásával ugyanolyan eredményeket elérnünk, mint 200 ezer szövegszónyi címkézett adatbázissal. De azt is bemutattuk, hogy a sztenderd részben felügyelt tanulási technikák vagy nem alkalmazhatóak (alacsony sűrűségnél vágás) a nagy méretű problémákra (ami általában a helyzet a számítógépes nyelvészetben) vagy nagyon körültekintő szövegválasztást igényelnek (jelöletlen adat és a kiértékelő adatbázis jellegében meg kell, hogy egyezzen). Ezért javasoljuk speciálisan nyelvtechnológiai problémákban alkalmazható módszerek alkalmazását.

Külön megvizsgáltuk azoknak a jellemzőknek a hozzáadott értékét melyeket jelöletlen korpuszokból származtattunk. Ezek alkalmazása átlagosan mintegy 28% relatív hibacsökkenést vontak maguk után. Végül három, WWW-en alapuló heurisztikát ismertettünk, melyekkel bizonyítottuk, hogy – annak ellenére, hogy a Web-en sokszor találkozhatunk elírással, valótlan információval, azaz zajjal – számítógépes nyelvészeti problémák megoldása során igen hasznos segítség lehet, a világ legnagyobb jelöletlen korpusza a WWW.

A jövőben szeretnénk a specifikusan számítógépes nyelvészeti problémák megoldására testreszabott részben felügyelt tanulási technikákat tovább vizsgálni, elsősorban olyan megoldásokat megcélozva, amelyek a WWW-et, mint külső szakértői tudást effektívebben és sokkal általánosabban tudják felhasználni nyelvtechnológiai problémák megoldása közben.

Bibliográfia

1. Baluja S.:Probabilistic modeling for face orientation discrimination. Learning from labeled and unlabeled data. In Neural Information Processing Systems (1998)
2. Chapelle, Olivier; Alexander Zien: Semi-Supervised Classification by Low Density Separation. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, 57-64 (2005)
3. Cozman, F.; I. Cohen; M. Cirelo:Semi-supervised learning of mixture models. ICML-03, 20th International Conference on Machine Learning. (2003)
4. Etzioni, Oren; Michael Cafarella; Doug Downey; Ana-Maria Popescu; Tal Shaked; Stephen Soderland; Daniel S.Weld; Alexander Yates.Unsupervised named-entity extraction from the web: an experimental study. In Artificial Intelligence Volume 165, Issue 1, 91-134 (2005)
5. Farkas Richárd, Szarvas György: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. Magyar Számítógépes Nyelvészeti Konferencia (2006)
6. Richárd Farkas, György Szarvas and Róbert Ormándi: Improving a State-of-the-Art Named Entity Recognition System Using the World Wide Web. Lecture Notes on Computer Sciences Vol. 4597. pp 163-172 (2007)
7. Goldman, S.; Y. Zhou: Enhancing supervised learning with unlabeled data. In Proceedings 17th International Confrence on Machine Learning. San Francisco, CA: Morgan Kaufmann (2000)
8. Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón: A Szószablya project. I. Magyar Számítógépes Nyelvészeti Konferencia (2003)
9. Ji, Heng; Ralph Grishman: Data Selection in Semi-supervised Learning for Name Tagging In Proceedings of ACL'06 Workshop on Information Extraction Beyond Document, Sydney (2006)
10. Lafferty, John; Andrew McCallum; Fernando Pereira: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of 18th International Conference on Machine Learning (2001)
11. Manning, Chris; Hinrich Schütze: Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. (1999)
12. McCallum, Andrew K: MALLET: A Machine Learning for Language Toolkit url: http://mallet.cs.umass.edu (2002)
13. Mitchell, T: The role of unlabeled data in supervised learning. In Proceedings of the Sixth International Colloquium on Cognitive Science. San Sebastian, Spain (1999)
14. Ke Lu, Jidong Zhao, Deng Cai: An algorithm for semi-supervised learning in image retrieval. Pattern Recognition, Vol. 39, No. 4. pp. 717-720 (2006)
15. Pasca, Marius; Dekang Lin; Jeffrey Bigham; Andrei Lifchits; Alpa Jain: Organizing and Searching the World Wide Web of Facts - Step One: the One-Million Fact Extraction Challenge. In Proceedings of American Association for Artificial Intelligence (2006)
16. L Rigutini, M Maggini: A semi-supervised document clustering algorithm based on EM. Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference pp. 200-206 (2005)

17. Shinzato, Keiji; Satoshi Sekine; Naoki Yoshinaga; Kentaro Torisawa: Constructing Dictionaries for Named Entity Recognition on Specific Domains from the Web. In Proceedings of ISWC'06 Workshop on Web Content Mining with Human Language Technologies (2006)
18. Sumita, Eiichiro; Fumiaki Sugaya: Using the Web to Disambiguate Acronyms. In Proceedings of NAACL '06 (2006)
19. Vapnik, V: Statistical learning theory. Springer (1998)
20. Szarvas György; Richárd, Farkas; András, Kocsor:. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In Lecture Notes on Artificial Intelligence Vol. 4265, 267-278 (2006)
21. Szarvas György; Richárd, Farkas; László, Felföldi; András, Kocsor; János, Csirik:. A highly accurate Named Entity corpus for Hungarian. In Proceedings of International Conference on Language Resources and Evaluation (2006)
22. Tjong, Erik F.; Kim Sang; Fien De Meulder: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of CoNLL-2003 (2006)
23. Yarowsky, D: Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 189–196 (1995)
24. Zhu, Xiaojin:Semi-Supervised Learning Literature Survey. Computer Sciences, University of Wisconsin-Madison, #1530 (2005)