Tisztán statisztikai alapú szófaji címkéző használata a Szeged Korpuszon

Kiss Géza, Németh Géza

Budapesti Műszaki Egyetem, Távközlési és Médiainformatikai Tanszék
{kgeza, nemeth}tmit.bme.hu

Kivonat: A cikkben bemutatott munka egy, eredetileg nyelvazonosításra kidolgozott, tisztán statisztikai alapú (morfológiai analízist igénybe nem vevő), automatikus gépi tanuláson alapuló címkéző eljárás alkalmazása szófaji címkézésre a Szeged Korpusz 2.0-ra. A megközelítés magyar nyelvre önmagában nem ad olyan pontos eredményt, mintha morfológiai elemzőt is használnánk, épp a tisztán statisztikai megközelítés miatt. Előnye, hogy nem látott szavak címkéjére is használhatóan jó becslést ad, viszonylag kis számításigényű, valamint tisztán statisztikai jellege miatt megfelelő tanítóhalmaz birtokában egy ismeretlen nyelvre is egyes gyakorlati alkalmazásokhoz elegendő pontossággal képes a szófaji címkézés feladatát ellátni. Ezek miatt más módszerek kiegészítőjeként is alkalmazható, pl. az ismeretlen szavak szófajának becslése céljából.

1 Bevezetés

A cikkben bemutatott munka egy tisztán statisztikai alapú, morfológiai analízist igénybe nem vevő, automatikus gépi tanuláson alapuló címkéző eljárás, amely egy korábban nyelvazonosítás céljára alkalmazott módszerünk [2] alkalmazása szófaji címkézésre a Szeged Korpusz 2.0-ra [1].

Mint látni fogjuk, ez a megközelítés magyar nyelvre önmagában nem ad olyan pontos eredményt, mintha morfológiai elemzőt is használnánk, épp a tisztán statisztikai megközelítés miatt. Viszont vannak olyan előnyei, amelyek akár más módszerek kiegészítőjeként való alkalmazásra is érdemessé teszik.

Egyrészt a tisztán statisztikai megközelítés miatt megfelelő tanítóhalmaz birtokában egy ismeretlen nyelvre is elegendő pontossággal képes a szófaji címkézés feladatát ellátni. Másrészt ezt viszonylag kis memória- és számítási igénnyel teszi. Példának okáért szöveg-beszéd átalakítás során is fontos információ a szintetizálandó szöveg szavainak szófaja, viszont valós idejű alkalmazásában ezt az információt a lehető legkisebb erőforrás igényű eszköz használatával szeretnénk megkapni.

Morfológiai elemzők alkalmazásakor is szükségszerűen találkozunk ismeretlen (out of vocabulary, OOV) szavakkal, amelyek szófajának megállapításához szükséges kiegészítő megoldáshoz folyamodni. Az itt bemutatott módszer ezt a problémát is kezeli. A módszer másik összetevője a szó környezetének figyelembevételével a legvalószínűbb nyelvi címkesor közelítése.

A munka során a korpuszt XML formából egy egyszerűsített, csak a megoldás szempontjából releváns információkat tartalmazó formára konvertáltuk, közben konzisztencia ellenőrzést is végezve. Megvizsgáltuk a korpuszban egy egységnek címkézett kifejezéseket is, és úgy módosítottuk a saját címkéző algoritmusunkat, hogy a detektált kifejezések minél inkább egyezzenek a korpuszban található felbontással, bár a szóközt is tartalmazó egységek miatt ezt nem tudtuk teljesen megvalósítani. Utóbbi probléma kezelésére készítettünk egy olyan szövegváltozatot, amely minden szót (szóközökkel szeparált részt) külön címkével címkéz; ehhez a több szóból álló kifejezések minden szava külön-külön a teljes kifejezés címkéjét kapta. Ezt használtuk a betanítás során, az eredmények ellenőrzéséhez ezt és az eredeti formát is használtuk; az utóbbin való kiértékelés rosszabb eredményt ad.

A korpusznak az X (ismeretlen szavak), Z (korpuszhibák), valamint O (nyílt tokenosztály) kategóriájú elemeket is tartalmazó mondatait különválasztottuk. A mondatokat összekevertük, és különválasztottunk ennek tizedét teszthalmaznak, a fennmaradó részt pedig tanítóhalmazként használtuk. Az ellenőrzéseket a teszthalmazon és a teljes halmazon is elvégeztük.

A következőkben leírjuk a módszer elméleti hátterét, részletesebben bemutatjuk a Szeged Korpuszon való munkánkat, majd számszerűsített eredményeket adunk a módszer hatékonyságáról, végül néhány következtetést is megfogalmazunk.

2 A módszer elméleti háttere

A szófaji címkézés problémáját az irodalomban több különböző eszközkészlettel veszik munkába, amelyek alkalmazását magyar nyelvre már többen számba vették [4][5]. A nagyobb pontosságra törekvő módszerekben két összetevő található meg a kívánt végeredmény eléréséhez: az elsőben megállapítjuk, hogy adott szó melyik kategóriákba eshet, vagy legalább egy kiindulási címkét kap (pl. TBL esetén); a másodikban a szó környezetét figyelembe véve döntünk a szóbajövő alternatívák között.

A szó lehetséges szófajainak megállapítása történhet egyszerűen szólista alapján: a tanítóhalmazban előforduló szavakhoz tároljuk a lehetséges kategóriákat. Ennél lényegesen kifinomultabb módszer, amikor morfológiai elemzőt veszünk igénybe a szavak elemzéséhez. Mindkét esetben gondot okoznak azok a szavak, amelyek nem szerepeltek a listában (nem látott szavak), illetve az elemző szótárában (OOV szavak). Ennek a nehézségnek az orvoslására is több módszer létezik, melyek a szó egyes jellemzői (jellemzően az utolsó karakterei, stb.) alapján próbálnak egy/több jó becslést adni a szó szófajára, pl. a maximum entrópia módszer [3].

A [2]-ben bemutatott módszerünkben egy szó különböző osztályokhoz való tartozására a P(szó | osztály) valószínűségét döntési fával becsüljük. A becslés pontosságára adott küszöb az, hogy a tanítóhalmaz adott részét helyesen osztályozza, a szó leggyakoribb kategóriájának adva a legnagyobb valószínűséget.

Így nem szükséges előzetes feltételezés alapján kiválasztani a szó azon részét (pl. utolsó *n* karakter), amelynek várhatóan hatása lesz az osztályozás szempontjából, hanem az rátanul a probléma megoldására a rendelkezésre álló minták alapján: olyan irányban bővíti a döntési fát, hogy az a lehető legkevesebb feltétel vizsgálatával adott pontosságú helyes szétválasztást érjen el.

A második lépés elvégzéséhez is számos módszert használnak, pl. Markov-láncot [6], vagy TBL-t (Transformation Based Learning) [7]. A mi módszerünk leginkább a

Markov-lánccal való megközelítése hasonlít, mivel a szófajnak az adott helyen való előfordulása valószínűségét becsli a környezetében lévő szavak címkéje alapján. Viszont nem csupán a szót megelőző, hanem az azt követő szavak osztályát is figyelembe veheti. Ehhez szabály-sablonokat definiálunk, amelyekből a tanítóhalmaz alapján szabályokat hozunk létre: ezek azt adják meg, hogy adott környezetben milyen valószínűséggel fordulnak elő a különböző címkék. A szóra kapott címkevalószínűséget ezzel az értékkel módosítva egy pontosabb becslést kapunk a címkék valószínűségére, ami segít megtalálni a legvalószínűbb címkesort.

A módszer részletesebb leírását az érdeklődő olvasó megtalálhatja pl. a [2]-ben. A következőkben ennek a szófaji címkézésre való alkalmazását mutatjuk be.

3 Munka a Szeged Korpusz 2.0-val

A módszert a Szeged Korpusz 2.0-n [1] próbáltuk ki (2005. december 5-i állapot). Ez tartalmazza a szöveghatárok jelölését, ezen belül a mondatokat, a mondaton belül pedig annak kifejezéseit MSD morfo-szintaktikai címkékkel ellátva, a szintaktikai egységek jelölésével együtt. Ezt a korpuszt használtuk a betanításhoz és a teszteléshez.

3.1 A korpusz előzetes vizsgálata

A korpuszban az egyes szófaji kategóriákba eső szavakat megvizsgálva azt láttuk, hogy a már korábban felsorolt három kategóriába (X, Z, O) tartozó szavak automatikus címkézésével nem tudunk foglalkozni, mivel nincs megadva javított kód a korpuszhibákhoz, valamint az O kategória elemei időnként más osztályoktól nehezen megkülönböztethetők (pl. létezik számnév kategória, de ebben is van szám típus). A Type-Token Ratio (a kategóriába tartozó szavak és ezek különböző fajtáinak aránya) is jelzi a feladat bonyolultságát: erre a három osztályra a legnagyobb, mivel nagyon kevés közöttük a többször ismétlődő.

A gépi címkézés úgy működik, hogy szeparátor karakterek mentén egységekre tagolja, majd címkézni a szöveget. A szóköznek mindenképpen szeparátor karakternek kell lennie, a csak alfanumerikus karakterekből álló szövegrészeknek pedig mindenképpen egy egységet kell alkotnia. Ezen túl viszont további munkát igényelt a a szeparáló karakterek körének meghatározása úgy, hogy minél inkább a korpuszban előforduló egységekkel dolgozzon az algoritmusunk. Ehhez kigyűjtöttük a nem csak alfanumerikus karaktereket tartalmazó kifejezéseket, és azoknak a központozásoknak az előfordulását megengedtük az egységben, amelyek ezeket a kifejezéseket nem bontották meg. Így a ’.’, ’-’, ’—’, ’,’, ’/’, ’"’, ’&’, ’+’, ’:’, ’\’, ’@’ karakterek egy előfordulását alfanumerikus karakterek között nem vesszük szeparátornak.

A korpusz egy jellegzetessége, hogy szóközt is tartalmazó egységeket is tartalmaz (pl. „OTP Bank Rt.” mint főnév), ami a címkézés eredményének kiértékelését nehezíti, hiszen ezeket a címkéző nem tudja egy egységként értelmezni (bár egy utólagos feldolgozási lépés során ezekből előállíthatók a szintaktikai egységek, de ez számottevően megnehezíti a feladatot). Hogy lehetőleg mégis jól összehasonlítható eredményt kapjunk, használható megközelítés egy olyan szövegváltozat létrehozása, amelyben a kifejezések szófaji címkéjét az azt alkotó, szóközzel vagy más szeparátor

karakterrel határolt egységek mindegyikének elejére elhelyezzük (a fenti példával: „{N}OTP {N}Bank {N}Rt.”). Jóllehet ez nem ad minden esetben helyes címkét a különálló szavakra, mégis esélyt teremt arra, hogy több adattal tanítsuk az algoritmusunkat, valamint hogy a címkézés eredményét értékeljük. Az értékelést ezzel a szöveggel, és az eredeti címke-elhelyezést megtartó szöveggel is elvégeztük.

3.2 A tanító és teszt halmaz előállítása

A kiértékeléshez a korpuszt szétválasztottuk az X, Z, O kategóriákat nem tartalmazó és tartalmazó részekre, ezeknek mondatait összekevertük, és 9 részt tanításra, 1 részt tesztelésre használunk. Ennek megvalósítása úgy történt, hogy a korpusz XML formátumú tartalmából létrehoztunk egy címkék nélküli egyszerű szöveget, amely az automatikus címkéző bemenete lesz, valamint egy egyszerűsített, csak a szófaji címkékkel címkézett kifejezéseket tartalmazó szöveget. Kiértékeléskor a kimenetet ezzel hasonlítjuk össze, annak érdekében, hogy csökkentsük a munkához használt anyag méretét, és ne legyen további szükség XML feldolgozásra.

A tanító és teszt szövegek létrehozásakor ellenőriztük, hogy az XML fájlokban megadott mondatokat kiadja-e a mondathoz tartozó szavak listája. A korpusz használt változatában csak néhány apróbb eltérést találtunk, általában abból fakadóan, hogy a szerkesztők a mondatot eredeti állapotában, érintetlenül hagyták, míg a mondatot alkotó kifejezések felsorolásában helyenként javították a szöveg szóköz és központozás hibáit. Az eltérések tehát a szóközök más elhelyezésében (pl. „D. J.” helyett „D.J.”), illetve három egymást követő pontnak a „…” speciális karakterre való cseréjében jelentek meg; egy esetben jelent meg egy plusz ’.’ központozás a mondatvégi „stb.” kifejezés után. A címkéket tartalmazó szöveget az XML fájlok címkézett kifejezéseiből állítottuk elő. Itt is külön figyelmet igényeltek a szóköz-szeparátorok, mivel ezek az XML címkék között nem jelennek meg, ezért csak az eredeti mondatnak és a kifejezések sorozatának az egybevetéséből tudtuk őket kinyerni, hogy a tanító és a teszt-halmaz csak a címkékben térjen el egymástól. A feldolgozás során végzett konzisztencia-ellenőrzésekkel sikerült néhány hibás HTML címkét, felesleges XML címkét, hibás ill. rosszul zárójelezett MSD kódot, valamint véletlenül beszúrt karaktert is megtalálni, ami remélhetőleg segített még tovább javítani az addig is nagyon jó minőségű korpuszt.

3.3 Kiértékelés módja

A kiértékelés során megvizsgáljuk, hogy az algoritmus milyen arányú egyezést ad a tanítóhalmazban látott szavakra, a tanítóhalmazban nem látott szavakra, valamint ezek együttesére (a korpusz egészére). A teljes MSD címkékre való tanítás mellett fontosnak találtuk megnézni, hogy az MSD fő (szófaji) kategóriáira mennyire hatékonyan működik, hiszen egyes alkalmazásokban ez a kevésbé részletes jellemzés is elégséges lehet (pl. szöveg-beszéd átalakítás hangsúlyozásának javítására).

Ha egy szóra megállapított címke hibás, akkor is érdekes lehet az algoritmusunk jóságának értékelése szempontjából, hogy más környezetben előfordulhat-e a szó ezzel a címkével. Más szóval az is fontos, hogy olyan szófajúnak címkéztük-e, amilyen szerepben előfordulhat, csak a szövegkörnyezet másfajta egyértelműsítést tett volna szükségessé, vagy egyáltalán nem fordulhat elő a szó a megállapított szerep-

ben. Ennek megállapításához azt is ellenőrizzük, hogy a megadott alternatív MSD címkék között előfordul-e az általunk adott besorolás. Ha szólistákkal illetve morfológiai elemzővel dolgozunk, akkor ez az eset nem fordulhat elő, viszont ezek önmagukban nem is képesek ismeretlen szóhoz szófajt javasolni.

Az ellenőrzést elvégezzük a feldolgozási egységekre címkézett és az eredeti címke-pozíciókat tartalmazó szövegekre is.

4 Számszerű eredmények

Az eredményeket a következő pontokban adjuk meg, először a szavak önmagukban való címkézésének feladatára, azután a szavak környezetét is figyelembe vevő megoldásra. Mindenütt megadjuk annak az eredményét, ha csak az MSD főkategóriákkal (a szófajokkal) dolgozunk, ill. ha a teljes morfo-szintaktikai címkékkel, valamint hogy milyen arányban találtuk el a helyes címkét, és milyen arányban csak a lehetséges címkék egyikére döntöttünk.

4.1 A módszer erőforrásigénye

A szükséges erőforrásigény mind memóriakapacitásban, mint számításigényben csekély. A főkategóriák becslésére használt adatbázisunk nem egészen 1 megabájt méretű, és átlagosan 4,1 mélységű döntési fa bejárását igényli karakterenként, valamint a címke-kategóriák számának (11) megfelelő számú összeadást. A szófajra való döntés szavanként a legnagyobb szám megtalálásával történik (10 összehasonlítás). Ezért az algoritmus az itt megadott eredményt kis futási idővel éri el. A részletes címkékhez 37,2 megabájtos, átlagosan 2,9 mélységű döntési fával végeztük a tanító halmazban előfordult 980 kategóriára.

A környezet hatásának figyelembevétele megnöveli a futási időt, de ez sem számottevően. Itt is egy sekély mélységű döntési fa bejárására van szükség a címke valószínűségek kiszámításához használandó szabály megtalálásához, ez esetben lépésenként egyszer minden szóra; a végleges címkesorhoz a szavak számával arányos számú lépésben eljuthatunk.

4.2 Gyakorlati és elvi korlátok

Az, hogy legfeljebb milyen mélységűre nőhet az alkalmazott döntési fa, a tanítás elvégzése előtt eldöntendő. Mi az MSD főkategóriára való tanítás esetén maximálisan 5 mélységet, a teljes MSD címkére való tanításkor 4 mélységet engedtünk meg. Ilyen szintű korlátozásra jelen esetben elsősorban a tanítás memóriában való limitálásához volt szükség, bár természetesen a túltanítás (az általánosító képesség elvesztése) elkerülése végett is érdemes küszöböt megszabni erre a mélyégre. Ha a teljes MSD címkék betanítása esetén nagyobb küszöböt engedtünk volna, számottevően megnőtt volna a betanítás erőforrásigénye, viszont az elkészült adatbázisé nem, vagy nem számottevően. Ekkor a helyes azonosításának aránya megnőtt volna, legkevesebb, hogy a látott szavakon való javulás miatt. Az itt látható eredmények tehát nem a módszer képességeinek elvi korlátját jelentik.

A látott szavak helyes azonosításának arányát a tanítóhalmazban előforduló leggyakoribb címke használatával magasabbra lehetne emelni. Ezt a lehetőség valóban szükséges lehet kihasználni a pontosabb eredmény eléréséhez, bár ez számottevő adatbázisméret növekedést jelent, ha explicit címke valószínűségeket is tárolunk a környezethez igazodó címkék kiszámításához. Emellett a szólistában elő nem forduló szavakra két feldolgozási lépést: egy meghiúsult keresést egy szófában, majd egy új számítási lépést a nem látott szó szófajának becslésére. Azonban látni fogjuk, hogy pl. csak a főkategóriákra való döntés esetén a szólista kiküszöbölésével is egy azzal közel egyenértékű megoldásra jutunk.

4.3 Szavak szófajának környezettől független becslése

Az 1. táblázatban látjuk a MSD főkategóriák becslésének eredményét. A táblázat első két sorában azt az esetet látjuk, amikor több szóból álló kifejezéseknél minden szóra a teljes kifejezés címkéjét helyezzük, erre végezzük a tanítást és az ellenőrzést. A harmadik és negyedik sorban az ellenőrzést a korpusz eredeti címkéire végezzük, ami számottevően rosszabb eredményt ad, hiszen az több szóból álló kifejezések címkéje ilyenkor nem lehet helyes. Láthatjuk, hogy a tanítóhalmazban nem látott szavak több, mint 91%-ára helyes címkét helyezünk azoknak a látott szavakhoz való hasonlósága alapján, és csak kb. 7% kapott olyan címkét, amivel nem fordulhat elő.

A 2. táblázatban a teljes MSD címkék becsléséről láthatunk adatokat. Nem látott szavak több, mint 54%-ához a helyes címkét sikerült rendelni, a szó környezetének figyelembe vétele nélkül. Ahogy fentebb írtuk, a látott szavak helyes azonosításának arányát nagyobb mélységű döntési fa készítésével, vagy egyszerűen szólista használatával is megoldhatjuk, ami az utolsó előtti oszlopban 100%-os eredményt adna, a többi oszlop adatai pedig ennek az aránynak megfelelően változnának.

1. Táblázat: azonosítási arányok szavanként az MSD főkategóriákra

	ellenőrizve valódi címkékre			ellenőrizve alternatív címkékre		
	nem látottra	látottra	összesre	nem látottra	látottra	összesre
szavakra címkézett, csak teszt	91.73	95.70	95.40	94.07	99.10	98.72
szavakra címkézett, egész	91.39	95.94	95.90	93.81	99.06	99.02
eredetire címkézett, csak teszt	88.61	93.97	93.57	90.90	97.38	96.88
eredetire címkézett, egész	82.65	94.03	93.93	84.78	97.15	97.05

2. Táblázat: azonosítási arányok szavanként a teljes MSD kategóriákra

	ellenőrizve valódi címkékre			ellenőrizve alternatív címkékre		
	nem látottra	látottra	összesre	nem látottra	látottra	összesre
szavakra címkézett, csak teszt	54.27	81.82	79.67	58.09	91.15	88.58
szavakra címkézett, egész	54.02	82.66	82.43	57.86	91.29	91.02
eredetire címkézett, csak teszt	52.92	80.80	78.62	56.63	90.14	87.51
eredetire címkézett, egész	49.48	81.54	81.27	52.96	90.16	89.85

3. Táblázat: azonosítási arányok környezettel az MSD főkategóriákra

	ellenőrizve valódi címkékre			ellenőrizve alternatív címkékre		
	nem látottra	látottra	összesre	nem látottra	látottra	összesre
szavakra címkézett, csak teszt	92.64	96.22	95.95	94.64	98.79	98.48
szavakra címkézett, egész	92.34	96.45	96.42	94.39	98.79	98.76
eredetire címkézett, csak teszt	89.50	94.48	94.10	91.44	97.05	96.62
eredetire címkézett, egész	83.46	94.52	94.43	85.27	96.86	96.77

4. Táblázat: azonosítási arányok környezettel a teljes MSD kategóriákra

	ellenőrizve valódi címkékre			ellenőrizve alternatív címkékre		
	nem látottra	látottra	összesre	nem látottra	látottra	összesre
szavakra címkézett, csak teszt	57.37	90.45	87.88	60.74	95.18	92.49
szavakra címkézett, egész	57.11	91.16	90.89	60.50	95.32	95.05
eredetire címkézett, csak teszt	55.60	88.97	86.35	58.87	93.69	90.96
eredetire címkézett, egész	51.99	89.63	89.31	55.06	93.79	93.47

4.4 Szavak szófajának környezetet figyelembe vevő becslése

Ebben a részben a szavak környezetét is figyelembe vevő módszer kiértékelését adjuk meg. Láthatjuk a 3. és 4. táblázatban, hogy az előző pontban megadottakhoz képest számottevő javulást kapunk, különösen a teljes MSD címkékkel való munka esetén. A nem látott szavak helyes azonosításának eredménye kb. 3%-kal megnőtt, a látott szavaké 9%-kal. Ennek a különbségnek az oka, hogy a látott szavak valószínűségét sokkal pontosabban tudjuk becsülni, így a címke valószínűség szabályok használata is jobb eredményeket ad.

Az eredmények nem a módszerrel elérhető legjobb eredményt tükrözik (hiszen szólista használatával az összesített eredmények számottevően jobbak lennének), hanem hogy használható becslést kapunk vele a nem látott szavakra, ill. a látott szavakra is magas recall mértéket kapunk azok külön számontartása nélkül.

4 Konklúziók

A cikkben bemutattuk egy, eredetileg nyelvazonosításra kidolgozott automatikus címkéző módszerünk szófaji címkézésre való alkalmazását a Szeged Korpusz 2.0-n.

A kapott eredmények elmaradnak az egyéb, magyar nyelvre publikált azonosítási rátáktól. Ennek oka egyrészt, hogy nem használtunk morfológiai elemzőt, valamint a teszt során nem használtuk a tanítóhalmazban látott szavak listáját sem, hanem a látott és nem látott szavak címkéjére is közelítő címkét alkalmaztunk. Emellett a nyelvi címkézés nyelvi kategóriái sokkal jobban illeszkednek az alkalmazott valószínűségszámítási modellhez, míg a morfológiai címkézésre a morfológiai elem-

ző és a szólisták használata elvileg jobban illeszkedik a probléma pontos megoldásához.

Az eljárás azonban több előnyös tulajdonsággal rendelkezik: az MSD főkategóriák (szófaji címkék) előállítását csekély erőforrás igénnyel is egyes alkalmazások számára elégséges pontossággal végzi, valamint a nem látott szavak címkéire jó becslést ad. Ezek miatt alkalmas lehet valósidejű alkalmazásokban való használatra, pl. a szövegbeszéd átalakítás problémakörében, illetve más módszerek kiegészítésére.

A kutatásokat részben a Promóció (GVOP – 3.1.1. – 2004 – 05-0245/3.0) projekt támogatásával végeztük.

Bibliográfia

1. Csendes, D., Hatvani, Cs., Alexin, Z., Csirik, J., Gyimóthy, T., Prószéky, G., Váradi, T.: Kézzel annotált magyar nyelvi korpusz: a szeged korpusz. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged. (2003) 238-245
2. Kiss, G., Németh, G.: Machine learning algorithm for automatic labeling and its application in text-to-speech conversion. Híradástechinka, Vol. LXI, Scientific Association for Infocommunications. (2006) 28–35
3. Halácsy, P., Kornai, A., Varga, D.: Morfológiai egyértelműsítés maximum entrópia módszerrel. Magyar Számítógépes Nyelvészeti Konferencia 2005 (2005) 180–189
4. Kuba, A., Felföldi, L., Kocsor, A.: Pos tagger combinations on Hungarian text. 2nd International Joint Conference on Natural Language Processing, IJCNLP (2005)
5. Horváth, T., Alexin, Z., Gyimóthy, T., Wrobel, S.: Application of Different Learning Methods to Hungarian Part-of-speech Tagging. Proceedings of Ninth Workshop on Inductive Logic Programming (ILP99) Bled, Slovenia (1999)
6. Tsujii, S., J., Rim, H.: Part-of-Speech Tagging Based on Hidden Markov Model Assuming Joint Independence. Proceedings of the 38th Annual Meeting of the ACL (2000)
7. Ramshaw, L. A., Marcus, M. P.: Text chunking using transformation-based learning. Proceedings of the Third Annual Workshop on Very Large Corpora (1995)