Morfológiai egyértelműsítés maximum entrópia módszerrel

Halácsy Péter[1], Kornai András[1], Varga Dániel[1]

[1] Budapesti Műszaki Egyetem -- Média Oktató és Kutató Központ,
1111, Budapest, Stoczek u. 2.
{hp, kornai, daniel}@mokk.bme.hu

Kivonat: Cikkünkben olyan magyar nyelvű statisztikai morfológiai egyértelműsítő modelleket hasonlítunk össze, amelyekbe a korpusztól független morfológiai elemzőt is beleépítettünk. Ismeretes, hogy magyar nyelvre a morfológiai elemző alkalmazása megnöveli a pontosságot a tisztán statisztikus módszerekhez képest. Modelljeink ugyanakkor a maximum entrópia módszer segítségével hatékony becslést adnak a morfológiai elemző által fel nem ismert szavakra is, tehát robusztusan viselkednek olyan tesztkorpuszokon is, amelyekhez a morfológiai elemző nem lett adaptálva.

1. Bevezetés

A morfológiai analízis (MA) a magyar, és általában az összetettebb morfológiájú nyelvek számítógépes kezelésének egyik központi feladata: a helyesírás-ellenőrzéstől a gépi fordításig szinte nincs is olyan gyakorlati alkalmazás, amelyhez valamilyen formában ne lenne szükséges MA. De még ha tökéletes (minden szót ismerő, és hibát soha nem vétő) MA algoritmus állna is rendelkezésünkre, akkor is szembe kell néznünk azzal a ténnyel, hogy a magyarban számos szóalak többértelmű, és hogy melyik elemzés a helyes, azt csak a szövegkörnyezet alapján lehet eldönteni.

Cikkünkben a morfológiai egyértelműsítés problémáját a statisztikai módszerek szemszögéből tárgyaljuk: ennek fő előnye, hogy a kontextus vizsgálatát egyértelműen korpusznyelvészeti alapokra helyezi. A címkézési feladatra a legjobb eredményt nyelvünkre tudomásunk szerint eddig Oravecz és Dienes [10] érte el 98.11% pontossággal. Ők a *TnT* rejtett Markov modell (HMM) alapú rendszert [2] módosították: a legnehezebb feladathoz, a tanítókorpuszban nem látott szavak helyes címkézéséhez a Humor morfológiai elemzőt hívták segítségül.

Cikkünk első részében bevezetjük a valószínűségi MA (WMA, weighted MA) fogalmát, és ennek segítségével a morfológiai egyértelműsítési probléma nehézségére adunk előzetes becslést. A második részben egy a magyar nyelvre eddig még nem alkalmazott, a maximum entrópia elvén alapuló szófaji címkéző módszert ismertetünk. Ehhez morfológiai elemző komponensként a hunmorph rendszert [12] alkal-

maztuk a morphdb.hu nyelvi erőforrással [14]. Az eredményeket a harmadik részben ismertetjük és értékeljük.

Magyar nyelvre a korábbi vizsgálatok elsősorban egy idealizált (a tesztanyag minden szavát garantáltan ismerő) morfológiai elemzőre támaszkodtak, ezért általános felhasználási értékük némileg megkérdőjelezhető, különösen akkor, amikor olyan kicsi és stilisztikailag homogén korpuszon alapulnak, mint a MULTEXT-East *1984* anyaga [3]. Munkacsoportunk az itt bemutatott algoritmus tanításához és teszteléséhez a Szeged Korpusz 2. változatát [4] használta, ennek az *1984* csupán 8%-a, és az Oravecz és Dienes [10] által használt korpusszal (280 ezer szövegszó) stílusában leginkább összemérhető `wholenews` szekció (ezt a sajtó és az üzleti rövidhír részkorpuszok összevonásával hoztuk létre) is némileg nagyobb a Szeged Korpuszban (350 ezer szövegszó).

Bár az `1984` anyagon elért 97.91%, a `wholenews` anyagon elért 98.38%, és Szeged Korpusz egészén elért 98.17% numerikusan nem jelentenek hatalmas javulást, úgy véljük, hogy rendszerünk a gyakorlatban jobban használható lesz. Nem csak azért, mert kritikus komponensei, beleértve a `WMA`-t, nyílt forráskódúak és szabadon módosíthatóak, hanem mert az általunk javasolt algoritmus robusztusan ellenáll a korpuszhoz nem igazított MA algoritmusok gyakorlatban nem ritka lefedettségi hiányosságainak, és mint ilyen, lehetővé teszi az eddiginél nagyobb változatosságú, pl. a dinamikusan növekvő magyar web kiaknázásával épült korpuszok [6] morfológiai elemzését is.

2. A címkézési feladat

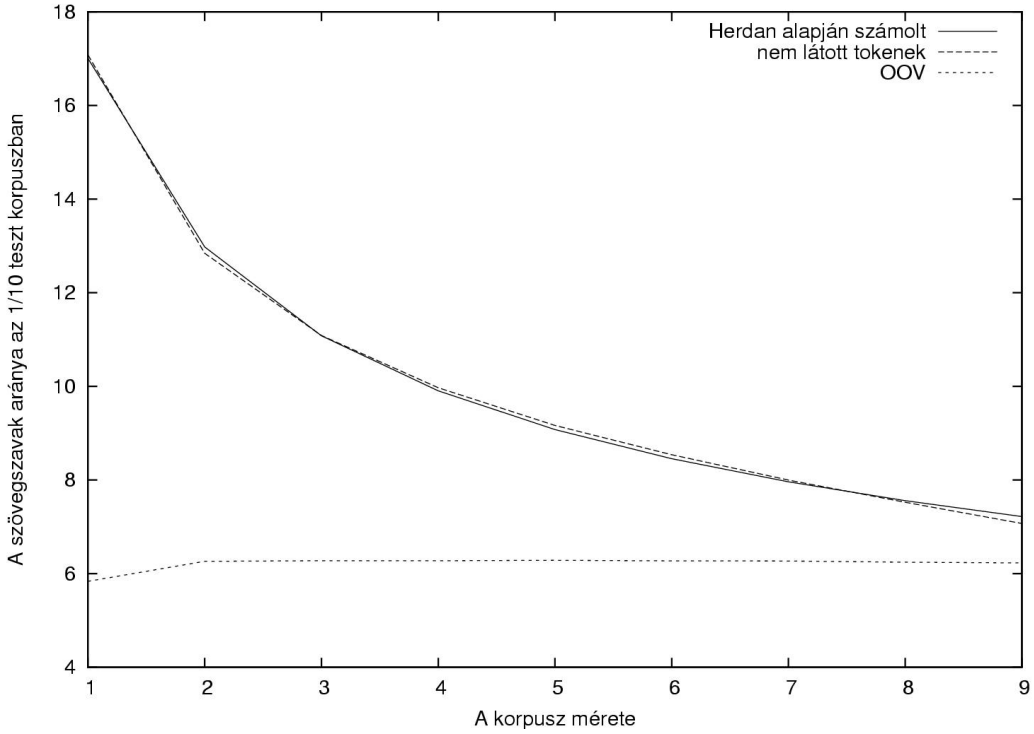
A morfológiai egyértelműsítés központi feladata a több elemzéssel rendelkező szavak esetében a helyes elemzés kiválasztása: ennek a feladatnak a nehézségét szokás a többelemzésű szövegszavak arányával [4], illetve az egy szövegszóra jutó elemzések átlagos számával [13] mérni. Ezeket a számokat azonban erősen torzítják a gyakori, de nem minden elemzést egyforma valószínűséggel nyerő szövegszavak (pl. *az* tipikusan névelő de lehet mutató névmás is, *én* tipikusan névmás, de pszichológiai szakszövegben gyakran főnév), hiszen a legegyszerűbb maximum likelihood címkézési stratégia számára ezek nem igazán problémásak.

A feladat nehézségének helyes mérőszáma tehát az egy szó egyértelműsítéséhez átlagban szükséges információmennyiség. Ha a *w* szó a T_i címkét $P(T_i \mid w)$ valószínűséggel kapja (címkézett korpuszból ezt a $C(T_i, w) / C(w)$ hányadossal becsülhetjük empirikusan, ahol *C* az előfordulások száma) akkor e szó címke-entrópiája $H(w) = -\sum_i P(T_i \mid w) \log P(T_i \mid w)$, és a címkézési feladat egészének nehézségét ezen entrópiáknak a *w* szavak gyakorisága szerint súlyozott átlaga adja, vagyis: $\sum_w P(w) H(w)$. Ez a Szeged Korpuszon durván 0.1 bit/szó (a pontos érték a választott címkerendszertől függ), tehát messze nem olyan nagy, mint azt a többelemzésű szavak aranyából gondolhatnánk: ha a lehetőségek mindig éppen egyformán valószínűek és a korpusz fele kétértelmű [4], akkor az entrópia akár 0.5 bit/szó.

A gyakorlatban természetesen a morfológiai elemző nem tökéletes, az egyes szavak gyakoriságát és címke-entrópiáját pedig csak becsülni tudjuk. Különösen érdeke-

sek számunkra azok a módszerek, amelyek e becsléseket a morfológiai elemző kiküszöbölésével, egyenesen a korpuszból végzik, hiszen ezek a morfológiai analízis (MA) nélkül működő, csak a korpuszból tanuló címkéző algoritmusoknak felelnek meg. A címkézési feladatot már ilyen algoritmusokkal is meglehetősen sikeresen meg lehet oldani: ha például minden adott szövegszóhoz a tanítókorpuszban látott szövegszavak esetén a típus leggyakrabban előforduló címkéjét, a nem látott típusok esetén pedig a nyílt kategóriák közül a leggyakoribb (egyes szám alanyesetű főnév) címkét rendeljük, akkor a Szeged Korpuszon (90% tanítás, 10% teszt, 10-szeres keresztvalidáció) 92% pontosságot érünk el. Ugyanezt az algoritmust tekinti alapszintnek (baseline) [10], de ott csak 81.2% pontosságot mérnek. A különbségnek az az oka, hogy a mi tanító- és tesztkorpuszaink egy nagyságrenddel nagyobbak, és így esetünkben csupán 10.7% a nem látott szövegszavak aránya, szemben az általuk tapasztalt 17.13%-kal.

Általában, ha a tanítókorpusz mérete *N*, a tesztkorpuszé ennek konstans hányada (pl. *N/10*), akkor Herdan törvénye szerint a tesztben az új szavak aránya cN^{q-1} ahol q a Zipf konstans reciproka. Az 1. ábrából látható, hogy a korpusz méretének növekedésével a fix arányú tanító- és tesztkorpusz esetén a nem látott szavak száma folyamatosan csökken: a mért és a Herdan-törvény segítségével számolt értékek megdöbbentően közel állnak egymáshoz (*q* és *c* paramétereket a korpusz alapján becsültük).



1. ábra. A tesztkorpuszban nem látott szavak arányának csökkenése eredeti korpuszon és a kevert változaton.

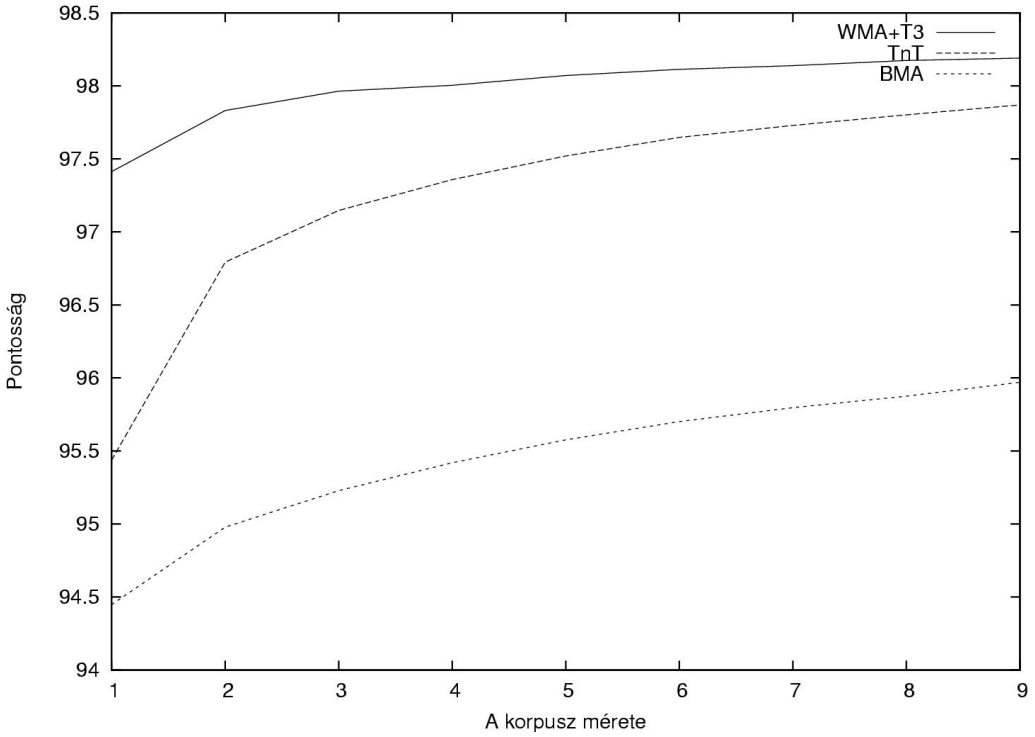
A Szeged Korpusz több, egymástól műfajában és nehézségben teljesen különböző szekcióból áll. Hogy az 1. ábra és 2. ábra görbéit elég nagy korpuszra is fel tudjuk rajzolni, a korpuszt még tanító- és tesztkorpuszra bontás előtt összekevertük. Az ezen

a korpuszon mért pontosság (2. ábrán) nem vethető össze a hagyományos 10-es keresztellenőrzéssel nyert eredményeinkkel, mert a keverés hatására a nem látott szavak aránya nagyon lecsökken a tesztkorpuszban (akár 30%-kal is).

Már [10] is kiemeli, hogy a produktív magyar morfológia miatt a magyar nyelvű korpuszokon nagyobb a nem látott szavak aránya, mint egy ugyanakkora méretű angol korpuszon. (270,830 szövegszó esetén mértek magyarra 17.13%, angolra 4.5%-ot.) Miután a nem látott szavak aránya igen erősen befolyásolja az alapszintűnél összetettebb módszerek hatékonyságát is, alapjában három utat követhetünk:

(A) növeljük a tanítókorpusz méretét, hogy az ilyen szavak arányát csökkentsük,
(B) a nem látott szavakat a már látott szavakkal rokonítjuk, vagy
(C) a nem látott szavakra vonatkozó heurisztikát javítjuk, pl. MA igénybevételével.

Közhelyszámba megy, hogy a gyakorlatban a leghatékonyabb az (A) módszer „there is no data like more data”, és ezt mutatják a mi vizsgálataink is.



2. ábra. Különböző algoritmusok tanulási görbéje kevert korpuszon.

Jó példa a (B) módszerre az alapszintű algoritmus alábbi módosítása (ehhez hasonlót javasol [7] is), amire a későbbiekben `BMA`-ként (baseline MA) hivatkozunk:

1. Ha *w* a tanítókorpuszban szerepel, akkor a $T = \arg\max(T_i \mid w)$ címkét kapja, egyébként
2. ha az MA ismeri és egy címkét rendel a szóhoz, akkor ezt kapja,
3. ha az MA ismeri, de nem egyértelmű a szó, akkor az MA által kiadott $T_{w,i}$ címkék közül a tanítókorpuszban leggyakoribb címkét adjuk, minden egyéb esetben
4. a címkét `NOUN`-nak vesszük.

Ez a módszer a Szeged Korpuszon 95.40%, az *1984*-en pedig 95.84% pontosságot ér el, ami összemérhető a transzformáció-alapú tanuló-rendszerek eredményeivel ([7], [1], [9]), de messze marad a Markov modellel elérhető 98.11%-tól [10]. Mivel a módszer a látott szavakra igen magas pontosságot ad, és a nem látott szavak aránya monoton csökken a korpusz méretének növelésével, a teljes pontosság monoton növelhető a korpusz méretével, ahogy a 2. ábra mutatja.

Ugyanezen az ábrán látható a morfológiai elemző hatása is. Az MA nélkül működő rejtett Markov modellen alapuló `TnT` [2] a `BMA` modell felett teljesít, mert figyelembe tudja venni a szó környezetét is. Ugyanakkor, ha a rejtett Markov modellezést kiegészítjük úgy, hogy a nem látott szavaknál az MA kimeneti címkéire támaszkodjon, hasonlóan [10]-hez, akkor jelentősen megnő a pontosság. Ezt a módszert mi `WMA+T3`-ként jelöltük, mert tekinthető egy súlyozott MA (weighted morphological analyzer) és a három szó méretű kontextust figyelembe vevő Markov-lánc együttesének. Ezt a modellt a következő fejezetben részletesebben mutatjuk be.

A 2. ábrából az is kiolvasható, hogy az MA jótékony hatása a korpusz növekedésével, és így a nem látott szavak arányának csökkenésével egyre kisebb lesz. Ahogy növeljük a korpusz méretét, a `TnT` és a `WMA+T3` hibaszázalékai közötti különbség egyre csökken. Közöttük a fő különbség csupán az, hogy a nem látott szavakra a `WMA+T3` az MA kimeneti címkéi közül tud választani.

A morfológiai egyértelműsítők hibája értelemszerűen a tesztkorpusz olyan szövegszavainál a legnagyobb, amelyek sem a tanítókorpuszban nem szerepeltek (mint láttuk ezek aránya a korpusz növekedésével csökken), sem az MA nem ismeri őket (out of vocabulary, OOV). Ezek aránya a korpusz méretétől független: az ilyenek teszik ki a tesztkorpusz 2%-át. Egy adott korpuszon az OOV tetszőlegesen csökkenthető, sőt akár ki is küszöbölhető az MA tőtárának növelésével (különösen hasznos lehet ez az eljárás az 1984 újbeszédének lefedéséhez). De hosszú távon, dinamikusan növő korpuszon (amilyen például a magyar web) 2% alatti OOV nemigen várható, hiszen a köznyelv állandóan bővül új szavakkal, különösen tulajdonnevekkel. A magyar szófaji címkéző szakirodalomban eddig egységesen követett eljárás, hogy az MA építést előre, a tanító- és a tesztkorpusz különválasztása előtt, a teljes korpusz alapján elvégzik. Ez azonban csupán az OOV problémát a mérésből kiküszöbölő egyszerűsítésnek tekinthető, és ezért az eddigi eredményeknek egy új korpuszon való reprodukálhatósága megkérdőjelezhető.

3. A maxent modell

A maximum entrópia (maxent) modellt szófaji címkézésre először Ratnaparkhi [11] javasolta. Ebben a keretben minden osztályozandó objektumhoz (esetünkben szövegszóhoz) úgynevezett jegyek (predikátumok, angolul features) halmazát rendeljük, és a rendszer ezek alapján tanulja meg a kimeneti címkéket (melyeket szintén jegyként kezel). A jegyek meghatározásakor nemcsak az éppen aktuális szót, hanem annak környezetét (rendszerünkben a közvetlen szomszédait) is figyelembe vehetjük. A maximum entrópia modellezéshez az OpenNLP maxent programkönyvtárat (`http://maxent.sourceforge.net/`) alkalmaztuk.

Míg az előző szakaszban tárgyalt (B) eljárás a morfológiai elemzést csak a tesztszót a már látott tanítószavakkal való rokonítására használja, az alábbiakban javasolt

architektúra inkább a (C) úthoz áll közelebb, amennyiben túllép az MA által adott ambiguitási osztályokon, és a címke-valószínűségekre explicit becslést tesz.

A következőkben a mondatokat szavak $w_1,\ldots w_n$ sorozatának tekintjük, amelyhez tanításkor ismert a $t_1,\ldots t_n$ címke-sorozat. A maximum entrópia modell egy együttes eloszlást határoz meg a lehetséges t_i címkék és az aktuális c_i kontextus között,

$$p(t_i, w_i) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(t_i, c_i)}$$

ahol π egy konstans normalizációs faktor, $\{\alpha_1,\ldots,\alpha_k\}$ a modell paraméterei és a $\{f_1,\ldots f_k\}$ a modellben használt bináris jegyek, amik minden címkére és kontextusra {0,1} értéket vehetnek fel (az 1 érték jelenti az adott predikátum teljesülését). Gyakorlatban a bináris jegyek helyett egyértékű predikátumokat is meg tudunk adni, amik bináris jegyekké alakíthatóak át. Jelenleg a következő jegyeket használjuk:

1. a szóalak kisbetűsítve[59]
2. nem mondatkezdő szó esetén a megelőző szó kisbetűs alakja
3. nem mondatzáró szó esetén a következő szó kisbetűs alakja
4. az MA elemzéseiből alkotott ambiguitási osztály
5. tartalmaz-e a szóalak számot, nemalfabetikus karaktert
6. csupa nagybetűs-e, nagy kezdőbetűs-e
7. ha 5 karakternél hosszabb a szó, akkor az utolsó 2, 3, és 4 karaktere külön-külön

Nem nyilvánvaló, hogy az MA elemzéseit hogyan kell jegyekké alakítani. A legjobb eredményt úgy értük el, ha az MA elemzéseinek halmazát (az ún. ambiguitási osztályt) egyetlen jegyként vettük fel. A szó utolsó néhány karakterére és a felszíni alakra vonatkozó jegyek alapjában az OOV probléma megoldását szolgálják: amikor a szót sem az MA nem ismeri sem a tanítókorpuszban nem szerepelt, akkor a modell csak a környező szavak és végződés adta jegyeket használja.

A tesztkorpusz címkézésénél a maxent modell által meghatározott együttes eloszlás alapján kiszámoljuk, hogy mi a kontextusra jellemző címke-eloszlás, azaz a mondat *i*. szavára, minden egyes lehetséges címkére kiszámoljuk a

$$P(t_i = T_k \mid c_i) = \frac{P(t_i = T_k \mid c_i)}{\sum_{t \in T} P(t_i = T_k, c_i)}$$

[59] A szó, előző szó, következő szó, a szuffixumok, az ambiguitási oszály, stb. mind predikátumok, amelyekből annyi különböző jegy lesz, ahány különböző szótípus, megelőző szótítpus, stb. található a korpuszban; a továbbiakban ezt a megkülönböztetést nem jelöljük.

valószínűséget. A maxent modell tehát nem hoz döntést, csupán minden egyes lehetséges címkére megadja annak valószínűségét. A maxent modell – bár jegyként megkapja az MA által adott címkéket – a tanítókorpuszban látott minden címke-típushoz pozitív valószínűséget rendel.
Első modellünk, a továbbiakban `MA+ME`, egyszerűen a fenti maxent modell alapján egy szóhoz a következő címkét rendeli:

1. Ha az MA ismeri a szót, akkor ezek közül választjuk a maxent modell által legvalószínűbbnek tartott címkét. (Speciálisan, ha az MA csak egyetlen elemzést ismer, akkor azt választjuk.)
2. OOV szóalak esetében a maxent modell választ.

Ez a modell csak lokális információkra hagyatkozik: egy adott szó címkézésénél nem veszi figyelembe a szó kontextusában lévő szavak címkéjét, ellentétben például a HMM alapú `TnT`-vel. Ezért két további modellt javasolunk.

A `WMA+T3`-nak nevezett modell a maxent modell és egy trigram-simítás kombinálása. A maxent modell és az MA kombinálásával súlyozott MA-t (Weighted Morphological Analyzer, WMA) építhetünk, amely a szóhoz hozzárendeli címkék egy valószínűségeloszlását, az alábbi módon:

1. Ha a szó szerepelt a tanítókorpuszban, akkor a szó címkéinek valószínűségét maximum likelihood módszerrel becsüljük, mint az alapszintű módszereknél.
2. Ha az MA ismeri a szót, akkor pontosan az általa kiadott címkéket engedjük meg, és a maxent által ezekre adott valószínűségeket egyre normalizáljuk. Speciálisan, ha az MA csak egyetlen elemzést ismer, akkor annak egy valószínűséget adunk.
   Előfordulhat, hogy az MA olyan címkét ad ki, amit a maxent modell a tanítókorpuszban nem látott. Ennek most mi egy konstans valószínűséget adunk normalizálás előtt.
3. OOV szóalak esetében a maxent modell által legvalószínűbbnek ítélt három elemzést engedjük meg, és ezeket normalizáljuk.

A WMA tehát minden egyes szóra megadja lehetséges címkéit súlyokkal. A címkék közül ki kell választani azokat, amik megadják a mondathoz rendelhető legvalószínűbb címke-szekvenciát. Formálisan:

$$\arg\max P(t_1,...,t_n \mid w_1,...,w_n) = \arg\max P(w_1,...,w_n \mid t_1,...,t_n)P(t_1,...,t_n),$$

ahol az első a szorzat első tagját a `WMA` kimenete, a másodikat a tanítókorpuszban látott címke-szekvenciák alapján épített másodrendű Markov modell szolgáltatja. A Markov modell építéséhez, és a legvalószínűbb szekvencia megkereséséhez (Viterbi algoritmussal), a `SRILM`[60] programcsomagot használtuk. Ennél a modellnél a maxent modellből ki kell hagyni a megelőző és következő szó jegyeit (tehát a `WMA`

[60] http://www.speech.sri.com/projects/srilm/

kontextusfüggetlen), hogy a kombinált modellben a két komponens független legyen. A `WMA+T3` modell gyakorlatilag analóg Oravecz és Dienes [10] modelljével.

Az utolsó modellünk, a `TNT+MA+ME`, szintén érzékeny a címke-szekvenciára. Az előbbiekben bemutatott `MA+ME` modell jegyei közé felvesszük még a szó, a megelőző, és a következő szó címkéit. Tanítási fázisban ezek adottak, címkézéskor pedig ezeket a jegyeket a tanítási korpuszon betanított `TnT` modell jósolja meg.

4. Értékelés

Ahhoz, hogy a Szeged Korpuszt, mint tanító- és tesztkorpuszt alkalmazni tudjuk, konverzióra volt szükség az MSD címkék és hunmorph által használt KR címkék [8] között. A konverzió nem teljesen triviális feladat, mert a két rendszer még az inflexiós kódok tekintetében sem vág teljesen egybe (pl. a marginális esetragok és a familiáris többes kezelésében).

A reziduális főkategóriájú (`X`, `Z`, `O`) MSD-címkéket tartalmazó mondatokat elhagytuk a korpuszból. A hunmorph ugyan számos `X` elemet (ismeretlen szó) felismer, és a vele közös tőtárú hunspell számos `Z` (sajtóhiba) elemet ki tud javítani, de célunk nem az előfeldolgozás, hanem a morfológiai egyértelműsítés vizsgálata, és ezekhez az elemekhez a Szeged Korpusz nem adja meg azt a javított kódot (ground truth), amivel rendszerünk eredményeit össze lehetne hasonlítani. Az `O` főkategóriájú nyílt címkeosztály esetében pedig úgy tapasztaltuk, hogy a Szeged Korpusz szerkesztési elvei még nem teljesen kiforrottak ezekre nézve, ezek az elemek még manuálisan sem különíthetőek el megfelelő pontossággal egymástól és más kategóriáktól.

Az eredeti Szeged Korpusz 82,098 mondatából így végül 70,084 mondatot tartottunk meg. A korpuszból elhagyott mondatokat későbbi robusztussági tesztjeinkhez alkalmaztuk, `hard` részkorpusz néven. Bár szemünkben a tulajdonnévi csoportok kijelölése (named entity recognition) is külön feladat lenne, megtartottuk a szóközt tartalmazó tokeneket, amelyek a korpusz 1.37%-át teszik ki. Mivel az általunk használt MA ezeket nem ismeri, ezek méréseinkben garantáltan az OOV szavak számát növelik.

Összességében 1001 MSD címkét 744 KR címkére konvertáltunk, ami látszólag egyszerűsíti a címkézési feladatot, valójában azonban nem, mert a KR címke és a tő ismeretében az MSD címke gyakorlatilag 100%-ban visszaállítható, azaz nincs két címke összevonásából adódó információvesztés. Másképpen fogalmazva: egy adott százalékban korrekt KR címkézés mechanikusan, egy statikus táblázat segítségével ugyanilyen, vagy még nagyobb százalékban korrekt MSD címkézéssé alakítható.

1. táblázat. A modellek pontossága a Szeged Korpusz szekcióin.

szekció	méret	oov	alapszint	BMA	TnT	MA+ME	WMA +T3	TNT+MA +ME
irodalom	209785	5.79	86.20	95.46	96.02	97.37	97.63	**97.83**
iskola	290167	1.62	90.17	96.34	96.97	97.73	97.80	**98.01**
Sajtó	355311	9.98	82.68	94.36	97.32	97.93	98.14	**98.38**
számtech	157969	8.43	86.06	94.44	97.02	97.53	97.91	**98.11**
Jog	147766	4.97	91.41	96.89	98.44	98.76	98.96	**99.04**
teljes	1161016	5.64	89.70	95.40	97.42	97.72	97.93	**98.17**

Az egyes részkorpuszokat jellemző méret és OOV adatok után a két alapszintű modell (MA nélküli és MA-val működő) és négy statisztikai modell eredményeit közöljük: `TnT` a Brants-féle trigram modell, `MA+ME` a tisztán maxenten alapuló, a `WMA+T3` egy MA-t használó saját trigram modell, `TNT+MA+ME` pedig a `MA+ME` modell, amely a `TnT` kimenetét is megkapja bemeneti jegyként. A rendszerek hatékonysági sorrendje a szekció kiválasztásától teljesen függetlennek bizonyult.

A táblázatban látható, hogy a morfológiai egyértelműsítésnél fontos a címkeszekvencia mint információforrás. A `MA+ME` modell csak lokális információk alapján dönt, a környező szavak címkéjét nem veszi figyelembe. Ezzel szemben a `WMA+T3` és a `TNT+MA+ME` modellek nem szavanként hoznak egymástól független döntéseket, hanem az egész mondatra határozzák meg a legjobb címke-szekvenciát.

A tisztán statisztikai `TNT+MA+ME` pontossága felülmúlja az összes általunk ismert szabálytanuló rendszerét: [9] 96.52% pontosságot ér el a teljes Szeged Korpuszra és 98.26%-t a hírekre. [7] 98.03%-os pontosságot ér el az *1984* feladaton, ahol mi jelenlegi módszertanunk mellett csupán 97.91%-ot mérünk. Ehhez a korpuszból idealizált (azaz a tesztanyag minden szavát garantáltan ismerő) MA-t épít az egyértelműsítés fázisa előtt. Ha a rendszerünkben használt független MA-t kicseréljük egy korpuszból épített morfológiai szótárra, akkor [7]-tel immáron azonos feltételek mellett 98.50%-os pontosságot érünk el.

A robusztusságukat ellenőrizendő a rendszereink pontosságát megmértük a teljes `hard` részkorpuszon tesztelve, a standard korpusz megfelelő méretű véletlenszerűen választott részén tanítva, a pontosságba nem mérve bele a kezelhetetlen címkéket. Azt tapasztaltuk, hogy a `TNT+MA+ME` pontossága ebben a felállásban 97.80%, ami csupán fél százalékpontnyi csökkenés az ugyanekkora, véletlenszerűen választott tanító- és tesztkorpusszal mért 98.31%-os eredményhez képest. A kontextust kevésbé figyelembe vevő `MA+ME` esetében a csökkenés nagyobb, itt 97.87%-ról 96.93%.-ra változik a pontosság.

Az eredményekből látható, hogy a tisztán statisztikai elven működő modellek eredményesen kombinálhatóak az erőforrás alapú morfológiai elemzővel. Magyar nyelvre ezt először [10] demonstrálta. Modelljeink előnye az általunk alkalmazotthoz képest abban áll, hogy az OOV szavakat is képesek robusztusan kezelni. Eredményeink nem teljesen hasonlíthatóak össze, mert méréseinket más (bár hasonló méretű és jellegű) korpuszokon végeztük. A legjobb rendszerünk teljes Szeged Korpuszon mért 98.17% pontossága az OOV szavak kezelésén túl azért is kiemelkedő, mert műfajában nagyon különböző összetevőkből álló heterogén korpuszon keresztértékeléssel értük el ezt az eredményt. Így módszerünk remélhetőleg lehetővé teszi az eddiginél nagyobb változatosságú, például a dinamikusan növekvő magyar web kiaknázásával épült korpuszok morfológiai elemzését is.

5 Köszönet

Szarvas Györgynek és Vajda Péternek a korpuszért és annak átalakításáért, Trón Viktornak a morfológiai elemző beépítésében nyújtott segítségéért és Oravecz Csabának értékes tanácsaiért.

Irodalomjegyzék

[1] Kuba András, Bakota Tibor, Hócza András, and Csaba Oravecz. A magyar nyelv néhány szófaji elemzőjének összevetése. I. Magyar Számítógépes Nyelvészeti Konferencia, pages 16–22. 2003.

[2] T. Brants. TnT – a statistical part-of-speech tagger, 2000.

[3] Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufis. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics, pages 315–319, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[4] Csendes Dóra, Hatvani Csaba, Alexin Zoltán, Csirik János, Tibor Gyimóthy, Prószéky Gábor, and Tamás Váradi. Kézzel annotált magyar nyelvi korpusz: a szeged korpusz. In II. Magyar Számítógépes Nyelvészeti Konferencia, pages 238–245. Szegedi Tudományegyetem, 2003.

[5] Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Szógyakoriság és helyesírás-ellenőrzés. In Proceedings of the 1st Hungarian Computational Linguistics Conference, pages 211–217. Szegedi Tudományegyetem, 2003.

[6] Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In Proceedings of Language Resources and Evaluation Conference (LREC04). European Language Resources Association, 2004.

[7] Tamás Horváth, Zoltán Alexin, Tibor Gyimóthy, and Stefan Wrobel. Application of different learning methods to Hungarian part-of-speech tagging. In ILP, pages 128–139, 1999.

[8] András Kornai, Péter Rebrus, Péter Vajda, Péter Halácsy, András Rung, and Viktor Trón. Általános célú morfológiai elemző kimeneti formalizmusa. II. Magyar Számítógépes Nyelvészeti Konferencia, pages 172–176. Szegedi Tudományegyetem, 2004.

[9] András Kuba, László Felföldi, and András Kocsor. Pos tagger combinations on hungarian text. In 2nd International Joint Conference on Natural Language Processing, IJCNLP, 2005.

[10] Csaba Oravecz and Péter Dienes. Efficient stochastic part-of-speech tagging for Hungarian. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002), pages 710–717, 2002.

[11] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey, 1996.

[12] Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: open source word analysis. In *Proceeding of the ACL 2005 Workshop on Software*, 2005.

[13] D. Tufis, P. Dienes, C. Oravecz, and T. Váradi. Principled hidden tagset design for tiered tagging of Hungarian. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, 2000.

[14] Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Simon Eszter, and Vajda Péter. morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In *III. Magyar Számítógépes Nyelvészeti Konferencia*, 2005. megjelenés alatt.