Ismeretlen kifejezések és a szófaji egyértelműsítés

Zsibrita János[1], Vincze Veronika[1], Farkas Richárd[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{zsibrita, vinczev}@inf.u-szeged.hu
[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: A jelenleg használt magyar morfológiai elemző és szófaji egyértelműsítő eszközök számos esetben nem működnek megfelelően, elsősorban az ismeretlen (szótárban nem szereplő) szavak és kifejezések kezelése miatt. Előadásunkban bemutatunk egy új (teljesen JAVA-ban implementált) szófaji egyértelműsítő rendszert („magyarlanc"), amely a morphdb.hu nyelvi erőforrásra épülő morfológiai elemzőn és számos, ismeretlen kifejezések kezelésére kidolgozott szabályon alapul.

1 Bevezetés

Ebben a munkában bemutatjuk a `magyarlanc`-nak keresztelt szegmentáló és szófaji egyértelműsítő rendszerünket. A rendszer a morphdb.hu nyelvi erőforrásra [8] épül, de számos ponton kiegészíti (alternatívája) a hunpos rendszernek [4]. A legfontosabb eltérések:

- a harmonizált KR-MSD kódrendszert használja [3], így a Szeged Korpuszon [1] közvetlenül tanítottuk,
- relatív szótöveket ad eredményül,
- teljesen JAVA nyelven implementált, így könnyen integrálható nagy (akár webszerver) alkalmazásokba,
- számos szabályt tartalmaz ismeretlen kifejezések kezelésére.

A következő fejezetekben röviden bemutatjuk az egész elemző láncot, majd az utolsó pontot tárgyaljuk részletesen.

2 Kapcsolódó munkák

Számos magyar nyelvre kidolgozott szófaji egyértelműsítő rendszer látott már napvilágot. A Szegedi Tudományegyetemen két szófaji egyértelműsítő is készült korábban: egy, a rejtett Markov-modellre épülő statisztikai módszer, illetve a szabályalapú RGLearn algoritmus [5]. A két módszert kombinálták a TnT taggerrel is: a hibrid algoritmus körülbelül 1%-os javulást eredményez a szófaji egyértelműsítésben a

Szeged Korpusz 2.0-n mérve. A BME MOKK fejlesztése a hunpos, egy ingyenes és nyílt forráskódú HMM-alapú szófaji egyértelműsítő [4], egy nyílt forráskódú implementációja a TnT-nek. Itt az elsődleges cél az ismeretlen szavak morfológiai kódjának minél pontosabb megállapítása volt. A hunpos OCaml nyelven készült, egy magasrendű nyelven, mely támogatja a tömör, könnyen érthető kódolási stílust[1]. A HuMOR morfológiai elemzőre[2] is épült egy ismeretlen szavakat elemző rendszer: a szimbolikus megszorításokon alapuló részleges elemző a Magyar Nemzeti Szövegtárból[3] nyert statisztikai információval egészül ki [7]. Ezen – kifejezetten szófaji egyértelműsítésre mint célfeladatra kidolgozott – rendszerek mellett a szófaji egyértelműsítőt mint köztes lépést használják a magasabb rendű magyar szintaktikai elemzők is, mint például az MTA Nyelvtudományi Intézetében magyarra átültetett NooJ[4] és a MorphoLogic kft. MetaMorpho MorphoParse-ja[5].

A bonyolultabb morfológiával rendelkező nyelvek esetében a HMM-alapú egyértelműsítés versenyképesnek bizonyul a többek között SVM vagy CRF módszereken alapuló tanuló algoritmusok jelenlegi generációjával szemben. A magyarban, mint más erősen ragozó nyelvekben igen fontos megőrizni a részletes morfológiai információkat a szófaji kódokban annak érdekében, hogy a magasabb rendű feldolgozási feladatokban is hasznosíthatóak legyenek. Ez az angolban használatosnál jóval nagyobb kódhalmazhoz vezet (kódrendszertől függően akár 1000 körüli is lehet a címkék száma az angol treebankekben rendszerint alkalmazott 36-hoz képest), azonban ez nem válik a tanítás és az egyértelműsítés hátrányára, noha a nem generatív modellek tanító folyamatát számítási szempontból megdrágítja.

3 magyarlanc

A `magyarlanc` programcsomag[6] magyar nyelvű szövegek alap nyelvi elemzésére szolgál. A csomag tisztán JAVA nyelvű modulokat tartalmaz, ami biztosítja a platformfüggetlenséget és a nagyobb rendszerekbe (például webszerverek) történő integrálhatóságot. A csomag magában foglal egy angol/magyar nyelvdetektort, magyar nyelvre adaptált mondat- és tokenszegmentálót[7], illetve egy szófaji elemzőt.

[1] http://mokk.bme.hu/resources/hunpos

[2] http://www.morphologic.hu/Morfologiai-elemzes.html

[3] http://corpus.nytud.hu/mnsz/

[4] http://corpus.nytud.hu/nooj/

[5] http://www.morphologic.hu/MetaMorpho-technologia

[6] A rendszer nyílt forráskódú, a Creative Commons licenc alatt szabadon hozzáférhető: http://www.inf.u-szeged.hu/rgai/magyarlanc

[7] Kiindulási alapként a morphadorner rendszer szegmentálóit használtuk: http://morphadorner.northwestern.edu/

3.1 Szófaji elemző

A szófaji elemző (lemmatizáló és POS-tagger) a Stanford POS-tagger[8] egy módosított változata, amely az ismeretlen szavakra a morfológiai elemző által adott lehetséges elemzéseket használja fel (az eredeti implementáció az ismeretlen szavakra az összes lehetséges morfológiai kódból választ). A POS-taggert a Szeged Treebanken [1] tanítottuk az automatikus morfológiai elemzéseket bemenetként felhasználva. A tanítás folyamán egy csökkentett MSD-kódhalmazt (42 kóddal) használtuk, hogy a lehetséges címkék számát kezelhető korlátok közé szorítsuk. A csökkentett kódhalmazban a szófaji alkategóriákat csak akkor vettük fel, ha a megkülönböztetés egyes szóalakok esetén szükségesnek látszott a Szeged Korpusz alapján (például megkülönböztetjük a főneveken belül a részes és birtokos esetben állókat). A kódhalmaz redukálásánál azt az irányelvet követtük, hogy a csökkentett kódkészletet használó szófaji egyértelműsítő modul kimenete egyértelműen megfeleltethető legyen az eredeti MSD-kódoknak. Tehát például az Nc-sd és Nc-sg kódok redukált alakja különbözik, míg a Nc-sd és Nc-sd---s3 ugyanarra a kódra redukálódik, mert soha nem fordulhat elő, hogy egy szóalaknak ez a két kód lehetséges elemzése (és a szófaji egyértelműsítőnek döntenie kell köztük).

3.2 Morfológiai elemző

Ahogyan az előző fejezetben bemutattuk, azon szóalakok esetén, amelyek nem szerepeltek a tanító adatbázisban, egy morfológiai elemző meghatározza a lehetséges elemzések halmazát, majd a szófaji egyértelműsítő modulnak ezen halmazból kell választania. Az alkalmazott morfológiai elemző a morphdb.hu nyelvi erőforrás [8] egy új változatára épül. Az új verzióban a KR és MSD kódrendszer harmonizált verziója található meg [3]. A nyelvi erőforrást mint bemenetet használva, Gyepesi György szoftvercsomagja egy véges állapotú (karakterátmeneteket használó) automatát állít elő. Az elemzés eredménye egy KR-kódhalmaz, mely visszaírási információkat is tartalmaz. A morfológiai kódharmonizációnak és a visszaírási információknak köszönhetően ezek a kódok egyértelműen megfeleltethetőek egy MSD-kódnak és a hozzá tartozó relatív szótőnek. A megfeleltetést végrehajtva már közvetlenül használhatjuk a morfológiai elemzőt a szófaji elemző tanítására és kiértékelésére a Szeged Korpuszon.

Természetesen egyetlen nyelvi erőforrás sem lehet tökéletes fedésű. A következő fejezetben bemutatunk néhány egyszerű megoldást azoknak az eseteknek a kezelésére, amelyekre a morphdb.hu erőforrásra épülő automata nem ad egyetlen morfológiai elemzést sem.

[8] http://nlp.stanford.edu/software/tagger.shtml

4 Ismeretlen szóalakok kezelése

Ismeretlen szóalakok kezelésére kidolgoztunk néhány egyszerű megoldást (amelyek a `magyarlanc`-ba beépítésre kerültek). A Szeged Korpusz 2.5-ben 143612 különböző szóalak fordul elő. A morphdb.hu jelen verzióira épült automata ezeknek nagyságrendileg (l. következő alfejezet) 75%-ára ad legalább egy elemzést. A fejezetben bemutatásra kerülő egyszerű módszerek segítségével az ismeretlen szavak (amelyekre az eredeti automata nem ad elemzést) háromnegyedére kapunk elemzést.

4.1 Tulajdonnév gazetteer a morfológiai elemzéshez

Első lépésben megvizsgáltuk azt is, hogy milyen hatásai vannak az alap nyelvi erőforrás (morphdb.hu) tulajdonnevekkel történő felbővítésének, ugyanis az ismeretlen szavak nagy része tulajdonnév. Az alábbi táblázatban láthatóak a Szeged Korpuszon tanított és kiértékelt POS-tagger eredményei, amelyek csak a morfológiai elemzőhöz felhasznált tulajdonnév gazetteerben térnek el egymástól (a kiértékelési módszertan pontos leírását l. az 5.3 fejezetben).

1. táblázat: Különböző méretű tulajdonnév gazetteerek eredményei.

#tulajdonnév	Ismeretlen szavak	főnév (P/R/F)	összes szófaj (P/R/F)
498	24,79%	70,50/85,53/77,29	77,04/79,11/78,06
339133	19,47%	72,89/88,87/80,09	79,43/79,65/79,54

A felbővített alapszótárral 111199 szóalakra kapunk legalább egy elemzést (80,53% az ismert szavak aránya) és a szófaji egyértelműsítő rendszereknek mind a pontosságát, mind fedését javította. Az alább bemutatásra kerülő kísérleteink során minden esetben ezt a felbővített alapszótárból kiinduló morfológiai elemzőt használtuk.

4.2 Arab és római számok

Az ismeretlen esetek egy jelentős részét az arab és római számok képezték. Ezek nyílt tokenosztályt alkotnak. A véges állapotú automata kiegészíthető lenne speciális állapotokkal és átmeneti szabályokkal ezeknek a felismerésére, ami tulajdonképpen egy független automatát jelentene. A `magyarlanc`-ban egyszerű reguláris kifejezésekkel ismerjük fel ezeket (megjegyezzük, hogy kifejezéseink nem kiterjesztettek, így reguláris nyelvet generálnak, azaz ekvivalensek egy determinisztikus véges állapotú automatával). A kidolgozott reguláris kifejezések megkülönböztetik a sorszámneveket, a tőszámneveket, a törtszámneveket és osztószámneveket, valamint ezek nyelvtani eseteit is, és összesen 5708 szóalakra adnak elemzést (az ismeretlen szóalakok 21,23%-a).

4.3 Összetett szavak

Az összetett szavak szótárban történő felsorolása soha nem lesz tökéletes fedésű, míg az összetétel tagjai általában ismertek (pl. *szárny+fesz+táv*). Elemzésükkor ki lehet ezt használni, oly módon, hogy ismert összetevőkre bontjuk azt és ellenőrizzük, hogy érvényes összetételről van-e szó (például a *vírusgazda* szó *vírus*ra és *gazdá*ra történő felbontása után mindkét összetevő értelmes, de a *futár fut+ár* felbontása nem értelmes). A balról jobbra haladó véges állapotú automatás morfológiai elemzők is alkalmassá tehetők az összetett szavak elemzésére, például ha megsokszorozzuk az állapotokat és megkülönböztetjük a *táv* elemzéseit aszerint, hogy a szó elején vagyunk vagy már egy elemzett főnév vagy ige megelőzi azt.

Az általunk javasolt eljárás ennél jóval egyszerűbb és hatékonyabb. Amennyiben egy szóalakra nincs elemzésünk, megvizsgáljuk, hogy az összetett szó-e. Ehhez megkeressük a szó minden lehetséges (legfeljebb háromtagú) felbontását. Azokat a felbontásokat tekintjük lehetségesnek, ahol minden egyes összetevőnek van legalább egy elemzése az eredeti automata szerint. Vannak azonban olyan pszeudoösszetételek, amelyek nem érvényesek. Ezek kiszűrésére szakértői szabályokat adtunk meg, mint például: ha az első összetevőnek csak igei elemzése van, és a másodiknak nincs igei elemzése, akkor nem érvényes az összetétel. Az eljárás végén minden érvényes összetételt lehetséges elemzésként ajánlunk fel az utolsó összetevő morfológiai kódjával, illetve az utolsó összetevőt lemmatizáljuk (például a *részrehajlónak* szóalak esetén a lemma *részrehajló*).

A Szeged Korpuszon ezzel a módszerrel 12012 olyan szó helyes elemzését kaptuk meg a lehetséges elemzések között, amelyet az eredeti automata nem elemzett (az ismeretlen szavak 44,67%-a), és mindössze 1654 szóra (ismeretlen szavak 6,15%-a) ad a módszer helytelenül összetett szavas elemzést.

4.4 Kötőjelet tartalmazó tulajdonnevek

Az összetételek egy speciális esete, amikor kötőjellel képzünk egy ismeretlen szóból (általában tulajdonnév) és ismert köznévből álló összetételt (például *Bush-kormány*), ahol tehát már nem is szükséges minden összetevő „ismerete”. Egy utófeldolgozó lépésben minden olyan szót megvizsgálunk, amely tartalmaz kötőjelet. Amennyiben a kötőjel utáni rész egy főnév, feltehetjük, hogy ez egy tulajdonnév-köznév összetétel, és főnévnek jelöljük a köznév morfológiai kódjaival és relatív lemmájával (a *Telenor-csoporttal*-nak *Telenor-csoport* lesz a lemmája).

Hasonló módon, ha a kötőjel után egy lehetséges főnévi toldalék áll, akkor feltesszük, hogy a kötőjel előtti rész egy tulajdonnév, és főnévnek jelöljük a toldalék által megadott esettel és a kötőjel előtti résszel mint szótő (például a *Vodafone-nak* szóalak lemmája *Vodafone*). Mivel az összes lehetséges főnévi toldalékot nem akartuk felsorolni, más módszerhez folyamodtunk: a különböző morfofonológiai osztályokra választottunk egy-egy főnevet (*lány*, *némber*, *sün*, *fal*, *holló*, *felhő*, *kalap*, *kert*, *köd, néni*) és ellenőrizzük, hogy a toldalékot a mintafőnév után írva főnévi elemzést kapunk-e. Előfordulhatnak azonban pszeudotoldalékok is a kötőjel után (például *Ray-Ban*). Ezek nagy része morfofonológiai és hangrendi összeférhetetlenségi szabályok alapján kiszűrhető.

A kötőjeles esetek vizsgálatával a Szeged Korpuszon 1085 esetben kapunk helyes elemzést (az ismeretlen szavak 3,17%-a).

5 Szófaji egyértelműsítés és többszavas kifejezések

A szófaji egyértelműsítés kapcsán egy érdekes kérdés, hogy mi az elvárt elemzése a többszavas kifejezéseknek. Szemléletes példa a *Magyar Nemzeti Bank* frázis, amely egy darab főnévként szerepel a korpuszban, Np-sn MSD-kóddal. Ha az ilyen és ehhez hasonló kifejezések szavait külön-külön vizsgálnánk, akkor a frázis minden egyes szavához tartozna egy-egy lemma és az ahhoz tartozó szófaji kód. E példát vizsgálva a *Magyar* és a *Nemzeti* szavakra egyaránt melléknévi elemzést kapnánk (Afp-sn), míg a *Bank* egy főnévi (köznév, Nc-sn) szófaji kóddal lenne ellátva. A jelenlegi nyelvi elemző megoldások azt a stratégiát követik, hogy első lépésben minden tokenre meghatározzák annak morfológiai elemzését (a POS-tagger kimenete), majd egy későbbi (általában független) lépés feladata a frázisok azonosítása.

Mivel korábbi névelem-felismerési kísérleteinkből [2] azt tapasztaltuk, hogy a szófaji kódok hozzáadott információtartalma a névelem-felismeréshez elhanyagolhatóan kicsi, ezért egy újszerű megközelítést javaslunk: első lépésben egy modul vonja össze a kifejezéseket, majd ezeken végezzük el a szófaji egyértelműsítést. Ily módon a *Magyar Nemzeti Bank* kifejezésről mint egyetlen egységről kell döntést hoznia egy szófaji egyértelműsítőnek, ami intuitíve kézenfekvőbbnek látszik (ez a szintaktikai egység ugyanúgy viselkedik, mint bármely más főnév).

5.1 Frázishatárok azonosítása

Megvizsgáltuk, hogy ha a nyelvi elemző első lépésben meghatározza a frázisokat, majd ezeken hajtja végre a szófaji egyértelműsítést, jobb eredményeket érhetünk-e el, mint a hagyományos megközelítéssel.

A frázisok azonosításához szekvenciális tanulást (CRF, Conditional Random Fields [6]) használtunk. A rendszer a Szeged Korpuszban jelölt frázisokon (olyan termek, amelyek tartalmaznak szóközt) tanult[9]. A frázisok esetünkben a több tokenből álló tulajdonnevek, de a módszer tetszőlegesen kiterjeszthető (a tanító adatbázis módosításával), bármely, egy logikai egységet alkotó tokensorozat összevonására, mint például mennyiségek (*3 millió Ft*) vagy dátumok (*2012. december 21.*).

A frázishatár-jelölő tanuló algoritmus egyszerű jellemzők halmazát (kb. 100 ezer dimenzió) használta fel. A felhasznált jellemzőcsoportok az alábbiak voltak (részletesen l. [2]):

- felszíni jellemzők (a szóalakra mint betűsorozatra vonatkozó információk)
- környezeti jellemzők
- gyakorisági adatok

[9] A Szeged Korpusz 2.0-ban a több tokenből álló tulajdonnevek egyetlen tokenként vannak jelölve, és a lehetséges morfológiai kódok és lemmák is frázisszinten lettek meghatározva.

- tulajdonnévszótárak
- egyértelmű tulajdonnevek listája

Ezen egyszerű jegyeknek felhasználásával már 90% körüli pontosságú eredmény érhető el. Az így kapott modell segítségével ismeretlen (korábban nem látott) szövegekből tudjuk detektálni az összevonandó frázisokat.

5.2 Szófaji egyértelműsítés a frázisokon

Ha már ismertük az összevonandó frázisokat, minden frázist lecseréltünk annak utolsó szavára, tehát a *Magyar Nemzeti Banknak*-ot egyszerűen *Banknak*-ra cseréltük. Ezt követte a szófaji egyértelműsítés és a lemmák meghatározása.

Vegyük az alábbi példamondatot: *Levélben fordult az Országos Magyar Méhészeti Egyesülethez*. Egy egyszerű elemzés során az eredmény: `[levél/N, fordul/V, az/Tf, országos/A, magyar/A, méhészeti/A, egyesület/N, ./.]` lenne, melyben ugyan ha külön-külön vesszük a szavakat, akkor valóban helyes az elemzés, de a valamely szervezetre utaló jelentéstartalom teljesen elvész.

A fent ismertetett módszer alapján, ha sikerült helyesen felismerni fráziskénti az *Országos Magyar Méhészeti Egyesülethez* tokensorozatot, akkor az elemzés eredménye: `[levél/N, fordul/V, az/Tf, Országos Magyar Méhészeti Egyesület/N, ./.]`, ahol a szervezetre való utalás nem vész el, illetve a szervezetet jelölő valamennyi token egy egységet alkot, és főnévi kóddal kerül az elemzés eredményébe.

5.3 Kiértékelési módszertan

Ahhoz, hogy a standard megközelítéssel összevethető legyen a módszer, először automatikusan lemmatizáltuk a Szeged Korpuszt (`magyarlanc` felhasználásával), és a szótöveken tanítottunk egy frázishatár-felismerő CRF rendszert, minden egyéb paraméterében a korábban bemutatott módszerrel megegyező módon. Az így – immár lemmákon – tanult modell lemmatizált szövegek frázishatárainak meghatározására lesz alkalmas.

Ebben a megközelítésben tehát először szófajilag egyértelműsítjük a mondatokat, majd ennek eredményét felhasználva célozzuk meg a frázisok azonosítását (intuitíve a szótárakon alapuló frázishatár-felismerőnek jobban kell teljesítenie a szótövek ismeretében). Az előző példa szerint, ha a rendszernek sikerül detektálnia az `országos/A, magyar/A, méhészeti/A, egyesület/N` lemmasorozatot mint egy négy szóból álló frázist, akkor a tokensorozat a második lépésben összevonódik, így az a későbbiekben egy frázist fog alkotni. A frázis utolsó szava lemmatizált formában fog szerepelni a frázisban, a többi token azonban az eredeti formában kerül be, szófaji kódként pedig a frázis utolsó tokenjének szófaji kódja kerül az elemzésbe: `Országos Magyar Méhészeti Egyesület/N`.

Tehát a nyelvi elemzés kimeneteként mindkét módszernél szófajilag elemzett és frázishatárokkal annotált mondatot várunk el. A szófaji egyértelműsítőt és a frázisha-

tár-felismerőt is a Szeged Korpusz egy véletlenül választott 80%-án tanítottuk, majd a kiértékelést a maradék 20%-on végeztük el. A kétfajta megközelítést két különböző módon értékeltük ki. Az egyik esetben a névelem-felismerésben használatos frázisszintű pontosság/fedés/F-mértéket számoltuk ki. Ebben az esetben ha egy frázishatár nem jól lett meghatározva vagy annak típusa nem egyezett, azt mind hibás illesztésnek tekintettük. A másik kiértékelés tokenalapon történt, itt az egy egységként azonosított (és szófajilag egyértelműsített) többszavas frázisokat tokenekre bontottuk, és minden token a frázis szófaji kódját kapta meg (ezt a szétbontást az etalon és a predikált halmazon is végrehajtottuk).

5.4 Eredmények

Az alábbi táblázat tartalmazza a kétfajta frázishatár- és szófaji egyértelműsítő módszer eredményeit, valamint a 4. fejezetben tárgyalt utófeldolgozási lépések hozzáadott értékét.

2. táblázat: Szófaji egyértelműsítő rendszerek eredményei.

		frázisalapú kiértékelés P/R/F	tokenalapú kiértékelés P/R/F
1. POS-tagger 2. frázishatár	N	83.50/92.72/87.87	90.41/95.45/92.87
	A	93.92/89.66/91.74	94.04/89.67/91.81
	összesen	88.40/89.61/89.01	90.93/90.64/90.79
1. frázishatár 2. POS-tagger	N	89.00/95.07/91.93	90.49/95.75/93.04
	A	95.07/89.58/92.24	95.11/89.58/92.26
	összesen	90.38/90.27/90.33	91.06/90.79/90.93
1. frázishatár 2. POS-tagger +utófeldolgozás	N	88.96/95.19/91.97	90.50/96.04/93.19
	A	95.05/89.61/92.25	95.10/89.62/92.28
	összesen	92.25/90.31/90.36	91.15/90.88/91.02

Az eredmények alapján mindkét kiértékelő módszer szerint a frázishatárok előzetes detektálása, majd a frázisok egy egységként történő kezelése szignifikánsan jobb eredményt ér el (McNemar-teszt alapján), mint a klasszikus megközelítés. Ez elsősorban a főnevek és melléknevek pontosságának javulásának köszönhető, ami arra enged következtetni, hogy a frázisösszevonásokkal sok tévesen főnévnek/melléknévnek jelölt tokent javítani tudunk (például a *Magyar Nemzeti Ban*k esetében a két melléknévi token helyett – ha a frázishatárokat sikerül azonosítani és a frázist főnévnek jelölni – két főnévi jelölésünk lesz).

Az ismeretlen szavak elemzésére adott utófeldolgozási megoldásaink hozzáadott értéke a végső rendszerhez a tokenalapú kiértékelés alapján szignifikáns. A főnevek és a melléknevek esetén ezek alkalmazásával a fedés nő, míg a pontosság tulajdonképpen nem változik. Előbbi természetesen annak a következménye, hogy több főnevet és melléknevet azonosítunk utófeldolgozással, mint a nélkül.

6 Konklúzió

Ebben a munkában bemutattuk a `magyarlanc` nyelvi elemző rendszert. Ennek jellegzetességei, hogy JAVA nyelven implementálódott, szabadon hozzáférhető, MSD-kód és relatív szótő alapú, számos utófeldolgozási lépést tartalmaz ismeretlen szavak kezelésére, a frázishatárok felismerését is elvégzi (méghozzá a szófaji egyértelműsítés előtt).

A végső rendszer a klasszikus szófaji egyértelműsítő modulnál 1,3%-kal jobb F-mértéket ér el a Szeged Korpuszon.

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND és a MASZEKER kódnevű projektek keretében az NKTH támogatta.

Bibliográfia

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD). LNAI series Vol. 3658 (2005) 123–-131
2. Farkas R., Szarvas Gy.: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 22–31
3. Farkas R., Szeredi D., Varga D., Vincze V.: MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 354–357
4. Halácsy P., Kornai A., Oravecz Cs.: HunPos — an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (2007) 209–-212
5. Kuba A., Bakota T., Hócza A., Oravecz Cs.: A magyar nyelv néhány szófaji elemzőjének összevetése. In: Alexin Z., Csendes D. (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 16–22
6. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML (2001)
7. Novák A., Nagy V., Oravecz Cs.: Magyar ismeretlenszó-elemző program fejlesztése. In: Alexin Z., Csendes D. (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 45–54
8. Trón V., Halácsy P., Rebrus P., Rung A., Vajda P., Simon E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of 5th International Conference on Language Resources and Evaluation (2006)