Magyar nyelvű klinikai rekordok morfológiai egyértelműsítése

Orosz György, Novák Attila, Prószéky Gábor

MTA-PPKE Magyar Nyelvtechnológiai kutatócsoport
Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
1083, Budapest, Práter utca 50/a
e-mail:{oroszgy, novak.attila, proszeky}@itk.ppke.hu

Kivonat Cikkünkben azokat az eljárásokat mutatjuk be, amelyekkel a meglévő PurePos szóalaktani egyértelműsítő rendszert, valamint az abban alkalmazott HuMor morfológiai elemzőt egy klinikai dokumentumokból álló orvosi korpusz elemzésére adaptáltunk. Ismertetjük a rendszer fejlesztéséhez szükséges teszthalmaz létrehozásának lépéseit, a fejlesztés alatt álló egyértelműsítő építőelemeit, és az azokon végzett első doménadaptációs eljárásokat. Részletesen leírjuk a felhasznált morfológiai elemző tőtárának bővítési lépéseit, az egyértelműsítőben a morfológiafejlesztés egyes megoldásai mellékhatásaként fellépő hibákat és az azokra adott megoldásokat. Végezetül megmutatjuk, hogy az így kapott eszközzel relatív 41,86%-kal sikerült csökkenteni a címkéző által vétett hibák számát, megvizsgáljuk a fennmaradó hibákat, s javaslatokat teszünk azok javítására.

1. Bevezetés

A legtöbb kórházban az orvosi feljegyzések tárolása csupán archiválás és az egyes esetek dokumentálása céljából történik. Ezen adatok felhasználási lehetősége így csupán az egyes kórtörténetek manuális visszakeresésére korlátozódik. Korábban bemutattunk [1,2] egy olyan automatikus eljárást, amely az orvosi (azon belül is a szemészeti) rekordok helytelen szavait nagy százalékban javítani tudja. Ezen előfeldolgozási lépés után a mélyebb szemantikai összefüggések automatikus kinyeréséhez szükséges a dokumentumok mondatainak (morfo-)szintaktikai annotálása is.

A szófaji és ezzel együtt a morfológiai egyértelműsítés a nyelvtechnológia egyik alapfeladata, mely a hagyományos szövegfeldolgozási lánc elején áll. Eredményének használatához – az egészségügy esetén pedig még inkább – annak nagy fokú pontossága szükséges. Angol nyelvterületen számos alkalommal vizsgálták már statisztikai tanuló algoritmusok orvosi doménre való adaptálását, míg a magyar nyelvű klinikai dokumentumok ilyen típusú feldolgozására nem ismerünk hasonló eredményeket.

Kutatásunkhoz szükség volt egy manuálisan annotált kis méretű korpusz létrehozására – immár nem csak szemészeti típusú klinikai dokumentumokat feldolgozva – melyet a bemutatott egyértelműsítő módszerek finomhangolására, tesztelésére és mérésre használtunk. Az ellenőrzött és javított morfológiailag címkézett szöveg elkészítéséhez a rekordokat automatikusan főbb alkotórészekre bontottuk, melyekből a kinyert szöveges bekezdésekhez adaptáltuk a központozási hibákat javító és tokenizáló rendszert, a morfológiai elemzőt és az egyértelműsítő rendszert.

Írásunkban a fenti lépéseken túl ismertetjük a HuMor morfológiai elemző [3,4] adaptálása során alkalmazott eszközöket, eljárásokat. Bemutatjuk az egyértelműsítő rendszer orvosi doménre történő alkalmazása során felmerült tipikus hibaeseteket és az erre adott megoldásokat. Végezetül áttekintjük az így kapott rendszer és részeinek eredményességét.

2. A tesztkorpusz létrehozása

A [2] cikkben korábban ismertetett helyesírásilag korrigált tesztkorpusz morfológiai egyértelműsítő fejlesztéséhez nehezen alkalmazható. Ez az anyag elsősorban szószintű problémák vizsgálatára lett létrehozva, és csak egy nagyon szűk domén kis méretű korpuszából vett szókincsével rendelkezik. Jelen kutatásunk keretében a korábbinál szélesebb domént lefedő és nagyobb terjedelmű tesztkorpuszt hoztunk létre. Az újonnan előállított korpusznak az alábbi feldolgozási lépéseken kellett keresztülmennie: a dokumentumok önálló strukturális egységekre tagolása, a központozási hibák automatikus javítása, mondatokra bontás és tokenizálás, a helyesírási hibák javítása, a szöveg automatikus morfológiai annotálása és annak manuális ellenőrzése, javítása. A munkánk során célunk olyan algoritmusok, módszerek készítése volt, mely segíti, támogatja morfoszintaktikailag egyértelműsített korpusz előállítását.

A korábbi XML-struktúrát létrehozó szabályalapú rendszer nem volt alkalmazható a szemészeti doménen kívül, mivel a a dokumentumok struktúrája osztályonként és akár orvosonként más és más. Így a szemantikai egységek meghatározásakor úgy döntöttünk, hogy a bekezdéseket tekintjük önálló összetartozó egységeknek. A bekezdésekre bontást egy, a formai jellemzők alapján működő egyszerű szabályalapú rendszer végezte, mely már általánosan alkalmazható volt. A bekezdéseket a további feldolgozás érdekében két osztályba kellett sorolni: főként nyelvi szöveget tartalmazó és egyéb, nem szöveges adatot tartalmazó bekezdésekre. Az osztályozáshoz az alábbi jellemzőket nyertük ki az egyes szakaszokból: sorok hossza, átlagos sorhossz, a legrövidebb sor hossza, átlagos soronkénti szószám, átlagos szóhossz, szavak száma, leghosszabb szó hossza, (feltételezhető) orvosnevek száma, egy szóból áll-e a bekezdés, whitespace karakterek aránya, írásjelek aránya, nagybetűk aránya, számszerű tokenek aránya, alfanumerikus karakterek aránya. Bár végeredményként a dokumentumok két osztályát kívántuk látni, azt tapasztaltuk, hogy a rendelkezésre álló adatokon ez a legtöbb közkedvelt gépi tanulásos algoritmusnak csak alacsony

eredményességgel sikerül, így a szövegek struktúrájához jobban illeszkedő alábbi osztályozást választottuk:

1. szöveges bekezdések,
2. fejlécek, szakaszcímek,
3. numerikus, illetve táblázatos adatok.

Egy kézzel ellenőrzött 500 bekezdésből álló teszthalmazon a klasszifikációs feladatra a J48 [5,6] döntési fa algoritmus bizonyult a legeredményesebbnek 93,2%-os keresztvalidált pontossággal.

Mielőtt a szövegeket a Huntoken rendszerrel [7] tokenizáltuk volna, az alábbi központozási hibákat javítottuk:

– a mennyiség és a mértékegység egybeírása,
– dátumok tagolatlansága,
– számszerű kifejezések egybeírása,
– jobbról tapadó írásjel (pont, vessző stb.) következő tokenhez való tapadása,
– központozási jeleknél whitespace-ek hiánya.

A szövegeinkben gyakori jelenség volt még a mondatvégi írásjelek hiánya, így a mondatokra bontás hibáját minimalizálva az egyértelmű helyeken tovább daraboltuk a bekezdést, így elkerülve, hogy több mondatot összevonva hibás határok kerüljenek megállapításra. (Pl. olyan sorok, amelyek csak rövid szöveget tartalmaznak a sor elején, nem vonandóak össze a következővel.) A mondatokra bontó alrendszerhez szükség volt még egy rövidítéslistára is, mely olyan – formai jegyeknek megfelelő – gyakori szavakból áll, melyeket automatikus módszerekkel illetve manuálisan is ellenőriztünk. (Pl.: a pont nélkül a HuMor által helyes szónak talált szóalakokat külön ellenőriztük.)

A véletlenszerűen választott 600 mondatból álló tesztanyag helyesírását, a már ismertetett [2] rendszerrel automatikusan javíttattuk, majd kézzel ellenőriztük és tovább javítottuk, majd a bemutatott egyértelműsítő alaprendszer kimenetét használva manuálisan annotáltuk a korpuszt.

3. Az egyértelműsítő rendszer kialakítása

3.1. A PurePos rendszer

Korábban ismertettük a PurePos [8] morfológiai egyértelműsítő rendszert, mely hatékonyan képes szófaji egyértelműsítésre és lemmák automatikus meghatározására. Bemutattuk, hogy a készített rendszer mind sebességben, mind pedig teljesítményben felveszi a versenyt társaival. A Szeged Korpuszon [9] tanítva és mérve 98,35%-os teljes pontosságról számolhattunk be. Integrált módon képes morfológiai elemzőt használni, mely a címkézés pontosságát – kis méretű tanítóanyag esetén is – minden tekintetben jelentősen növeli. Az eszköz nyílt forráskódú, Javában íródott, így működése könnyen módosítható. A rendszer alapjait a Brants [10] és Halácsy et al. [11] által ismertetett algoritmus

képezi, melyet úgy alakítottunk át, hogy képes legyen a morfológiai elemző integrált és hatékony használatára. Nagy előnye még a taggernek, hogy tanuló algoritmusának tanítási ideje – más maximum entrópia vagy CRF-alapú eljárásokhoz képest – nagyon alacsony, másodpercekben mérhető.

1. táblázat. A egyes szófaji egyértelműsítő modulok pontossága.

	PP	PP+	ME	PE	HuLaPos
Pontosság	83,82%	86,88%	80,14%	79,34%	81,59%

Az alábbiakban (1. táblázat) összehasonlítjuk a PurePos integrált HuMor morfológiai elemzőt tartalmazó változata (PP+), az integrált elemzőt nem használó (PP) és három további szófaji címkéző, az OpenNLP maximum entrópia (ME) és perceptronalapú taggere [12] (PE) és [13]-ban leírt, Moses dekoderen alapuló, Laki László által fejlesztett eszköznek (HuLaPos) a fenti tesztkorpuszon mért címkepontosságát. Valamennyi eszközt a Szeged korpusz Humor tagekre konvertált változatán tanítottuk be. A eszközök közül csak a PurePos és a HuLaPos ad lemmát is tartalmazó teljes morfológiai elemzést.

Egybevetve a szabadon elérhető nyelvfüggetlen PoS taggerek eredményességét, a PurePos hatékony eszköznek tűnik doménadaptációs eljárások fejlesztésére, tesztelésére. A korábbi és jelen eredmények alapján is elmondható, hogy a magyarhoz hasonlóan komplex morfológiájú nyelvek esetében a morfológiai tudás kulcsszerepet játszik egy magas pontosságú egyértelműsítőben, így szükségesnek ítéltük a HuMor morfológiai elemző orvosi doménre való adaptációját.

3.2. A morfológia adaptálása

Az egyértelműsítő rendszer egyik alkotóeleme a HuMor morfológiai elemző. Hogy az elemző orvosi szövegek elemzésében nyújtott teljesítményét növeljük, úgy határoztunk, hogy első körben lehetőleg viszonylag megbízható minőségű forrásból származó anyaggal bővítjük az elemző tőtárát.

A tőtár bővítésének egyik fontos forrása az 1992-ben megjelent Orvosi helyesírási szótár [14] volt. A helyesírási szótár semmiféle információt nem tartalmaz sem a benne szereplő szavak szófajára, sem azok nyelvére, illetve kiejtésére vonatkozólag, ezen információkra azonban a morfológiai adatbázisba való felvételükhöz szükség volt (illetve az összetett szavak esetében az összetételi határ helyét kellett megállapítanunk). Mivel több tízezer szót kellett annotálnunk, úgy döntöttünk, hogy a szavak kategorizálását és a hozzáadandó információk előállítását megpróbáljuk automatikus módszerekkel segíteni.

A szófaji kategorizációban egyrészt egyszerű formai jegyekre támaszkodhattunk (pl. a szótárban szereplő neveket és rövidítéseket ilyen alapon könnyen meg lehetett különböztetni az egyéb szavaktól). Másrészt a

szavak egy részének kézzel való szófaji kategorizációja után ezen az anyagon a PurePos-ban is alkalmazott végződésguesser-algoritmust tanítottuk be és alkalmaztuk, majd a kapott címkéket átnéztük és javítottuk, illetve ezt az eljárást iteráltuk. A latin-görög szókincs elemeinél bizonyos végződéstípusok esetében különösen nehéz volt eldönteni, hogy egy-egy szó főnév vagy melléknév, esetleg mindkettőként használatos. A kérdéses esetekben egyenként kellett utánanéznünk a szó jelentésének, illetve használatának, ami nagyon időigényes volt.

Ezért az automatikus szófaji osztályozásnál még egy szempontot figyelembe vettünk: a szótárban szereplő több tagú latinos kifejezések esetében az utolsó elem gyakran melléknév (hacsak nem birtokos szerkezetről van szó), a első elem pedig leginkább főnév, a elemek sorrendje tehát szisztematikusan különbözik a magyar jelzős szerkezetekétől. A latin melléknevek elsősorban emiatt jelentenek külön problémát a magyar nyelvű orvosi szövegek címkézése szempontjából. A magyarul írt megfelelőjük (amely a latin szó hímnemű alanyesetű alakjával áll alaki kapcsolatban) egyértelműen melléknév, amely a magyarban szokásos módon melléknév–főnév sorrendben áll. A valódi többszavas latin kifejezésekben a sorrend főnév–melléknév, és a két elem egyeztetve van. A nem hímnemű vagy esetleg nem alanyesetű szerkezetben álló latin melléknévi alakok a magyar címkézés szempontjából gyakorlatilag főnévnek tekinthetők. Elvileg ugyanez lenne a helyzet a hímnemű alanyesetűek szempontjából is, ha nem lenne a korpusz tele olyan szerkezetekkel, amelyek sorrendjükben a magyar névszói szerkezet mintáját követik (mivel azok), helyesírásukban azonban latinos írásmódú elemekből vannak összeállítva.

Ezért úgy döntöttünk, hogy a latin helyesírású főneveket és mellékneveket megkülönböztető címkével látjuk el a morfológiában, és ezek közül a hímnemű alanyesetű mellékneveket alapvetően melléknévként, a többit pedig főnévként címkézzük, hogy ha lesz elegendő kézzel ellenőrzött annotációt tartalmazó orvosi szöveget tartalmazó tanító anyagunk, a tagger ebből megtanulhassa a hímnemű alanyesetű latin melléknevek jellegzetes eloszlását. Sajnos a rendelkezésünkre álló idő egyelőre csak a tesztkorpusz létrehozására volt elegendő, ezért ezt a lehetőséget munkánk jelen fázisában nem tudtuk kihasználni, a latin szavakat megkülönböztető címkék tanító anyag híján egyelőre inkább problémát okoztak a taggernek, semmint segítséget.

A szófaj eldöntésén kívül tehát meg kellett különböztetnünk az idegen és a magyar helyesírású elemeket. Erre azért is szükség volt, mert az előbbiekhez a kiejtést is meg kellett határoznunk, hogy a szavak helyesen toldalékolódjanak. Ebben részben segítséget nyújtott, hogy a szótár utalásként sok olyan szópárt tartalmaz, amelyek ugyanannak a szónak vagy kifejezésnek a helyesírási változatai. Ezek legnagyobb részénél az egyik változat a magyar helyesírású, a másik az idegen helyesírású változat. Az esetek nagy részében a magyar volt preferált változatként megjelölve. Volt azonban az anyagban rengeteg kivétel is. Részleges manuális kategorizáció után erre a feladatra a TextCat algoritmus [15] egy adaptált implementációját használtuk, amely rövid stringekre is képes elég jól használható választ adni a magyar vagy nem magyar kérdésre. Viszonylag

egyértelmű volt a helyzet, ha egy szópár egyik tagját a rendszer inkább idegennek, a másikat pedig inkább magyarnak minősítette. A párok nagy része a szótárban ugyanakkor olyan, hogy mindkét eleme idegen, amelyek ugyanannak a szónak különböző írásváltozatai. Ezek kiszűrésében ugyancsak jó szolgálatot tett a fenti algoritmus. A korábban említett iteratív szótárbővítő eljárásnak ezt a nyelvmegállapító eljárást is részévé tettük. A szótár rengeteg olyan idegen (főleg görög-latin, emellett angol és francia) szót is tartalmaz, amelynek a magyar ortográfiával írt megfelelője nem szerepel a szótárban. Ezeket is fel kellett ismernünk, és itt nem támaszkodhattunk olyan implicit extra információra, amit a szópárok esetében a másik elem adott.

Amellett, hogy el kellett döntenünk, hogy az elem idegen vagy magyar, a konkrét kiejtést is hozzá kellett rendelni. Ez a hivatkozási rendszer folytán párban álló elemek esetében részben adott volt, bár az elemek nagy részének a magyaros mellett a latinos kiejtésére is szükségünk volt (különös tekintettel az s betűre végződő szavakra), hiszen sokszor önállóan is, több szavas latin frázis elemeiként viszont elvileg mindig a latinos kiejtés a mérvadó a toldalékolás szempontjából. Mivel rengeteg szó kiejtését kellett megadnunk, ezt sem kézzel csináltuk, hanem algoritmikusan állítottuk elő őket (az s végűeknél mindkét változatot), és az így előállított kiejtést javítottuk kézzel, ha szükséges volt. Erre a feladatra nem valamilyen általános gépi tanulás alapú G2P (grapheme-to-phoneme) algoritmust használtunk, hanem egyszerűen írtunk egy reguláris kifejezéseken alapuló heurisztikus algoritmust, amelynek kimenetét némi csiszolgatás után viszonylag keveset kellett javítgatni. Ezt akár a lexikon szerkesztésére használt editorból közvetlenül is meg lehetett hívni akár egy egyszerre kijelölt több szóból álló blokkra is, ha olyan szót találtunk, amelyet a korábbi algoritmusaink esetleg tévesen nem ítéltek idegennek.

További feladat volt az összetételi határok megállapítása, és az összetételekben gyakran szereplő elemek kiemelt kezelése: ezeket előrevettük a szavak feldolgozása során, így az ezeket tartalmazó összetételek kezelését a morfológiára bízva hatékonyabban csökkenthettük a feldolgozásra váró szótári tételek számát, illetve minimalizálhattuk a esetleges inkonzisztens manuális adatbevitel esélyét. Ehhez egy olyan algoritmust implementáltunk, amely az általános helyesírási szótárban és az orvosi helyesírási szótárban szóként szereplő legalább két karakter hosszú és magánhangzót is tartalmazó elemeket szófában eltárolva és azokat utótagként keresve a szótár szavaiban statisztikát készített az így felbontott szavak elemeiből, és a megtalált prefixumokat több szempontból osztályozta: külön megjelölte egyrészt a 4 karakternél rövidebbeket, a szótárban szóként létezőket, a belül kötőjelet tartalmazókat és azokat az eseteket, ahol a felbontott szó maga is utótagja volt a szótár valamelyik másik szavának. Ennek az eredményét felhasználva és a gyanúsnak tűnő elemekkel alkotott összetételeket külön kézzel ellenőrizve a leggyakoribb valódi elő- és utótagokat felvettük a szótárba, majd második körben az ezekkel képzett valódi összetételeket is, így hozzájutottunk a szótárban szereplő összetételek összetételi tagokat is jelölő reprezentációjához, amelyeket a szótárba felvettünk.

A szótár meglepő módon sok olyan igéből képzett szót (leginkább melléknévi igenevet és nomen actionist) tartalmaz, amelyek (általában latin-görög tőből képzett) alapigéje ugyanakkor nem szerepel benne. Ezek helyett a szavak helyett az alapigét vettük fel, hiszen így kapunk a képzett elemekre normális elemzést. A munka egyik fázisa az volt, amikor ezekre vadásztunk. Emellett sok olyan s-képzős melléknév szerepel a szótárban, amelyeknek alapszava is benne van. Első körben az ilyennek látszó szavakat is kihagytuk a feldolgozásból, mert az alapszó felvétele automatikusan a képzett szó bekerülését is jelentette. Ami még különös körültekintést indokolt a szótár feldolgozásakor, az az volt, hogy meglepően sok nyilvánvaló nyomdahibával találkoztunk benne, ezért nem lehetett készpénznek venni a szótárban szereplő adatokat.

A helyesírási szótár mellett a másik fontos feldolgozott szóanyag az OGYI[1] honlapjáról letöltött gyógyszernév- és hatóanyag-adatbázis volt. Itt a szavak kategorizálása és a szófaj eldöntése kevésbé okozott problémát. A kiejtés viszont itt is fontos volt. Az ezt kiszámoló algoritmusunkat annyiban adaptálnunk kellett, hogy mivel a hatóanyagok elnevezésére az jellemző, hogy bár azok alapvetően latinos-görögös elemekből épülnek fel, de az írásmódjuk az angolban szokásos képet mutatja, így a latin/görög végződések helyett, szinte mind ki nem ejtett *-e*-re végződik.

A szótárbővítés harmadik forrása természetesen maga a korpusz volt. Már a szótár feldolgozásakor előnyben részesítettük azokat a szavakat, amelyek a korpuszban is szerepeltek. De emellett az előbbi forrásaink feldolgozása után továbbra is elemzetlenül maradt gyakori szavak feldolgozása is fontos volt. Ezek túlnyomó része rövidítés volt. A gyakori rövidítések feloldását, és ez alapján a rövidítés szófaji besorolását (ha az nem volt a szóalak alapján teljesen nyilvánvaló) korpuszkonkordanciák alapján végeztük. Amire nem ügyeltünk eléggé (és ez nagyon jelentős negatív hatással volt a tesztek során a rendszer címkepontosságára), az az volt, hogy a feldolgozás során figyelmen kívül hagytuk azokat a pontra végződő szavakat (potenciális rövidítéseket), amelyre az elemzőnek már volt valamilyen elemzése, és így a korpuszban gyakori címkéjükkel a morfológiába nem kerültek bele.

Az orvosi szótár (egyelőre korántsem teljes) feldolgozása és a korpuszban szereplő leggyakoribb rövidítések felvétele együttesen 36000 tétellel bővítette a morfológia tőtárát (még mintegy 25000 szót nem dolgoztunk fel). A gyógyszernév-adatbázisból 4860 tétel került bele.

Az így javított elemzővel ellátott rendszer szófaji egyértelműsítésre számolt pontossága 93,25%, mellyel mintegy 6,4%-kal sikerült redukálni a korábbi rendszer hibáinak számát.

Közelebbről szemügyre véve a hibákat, azt tapasztaltuk, hogy a rendszer gyakori hibáinak egy része olyan jelenség, melyek a további szintaktikai, szemantikai feldolgozás szempontjából érdektelen. Ezek azon esetek, amikor a morfológia különbséget tesz latin, illetve magyar eredetű főnevek és melléknevek között, továbbá az igenevek és az ezekből lexikalizálódott melléknevek között.

[1] http://www.ogyi.hu/listak/

Ezen hibákat a továbbiakban nem számolva, a fenti eredmények 90,55%-ra és 93,77%-ra módosulnak az eredeti és a bővített morfológiát tekintve.

3.3. Az egyértelműsítő adaptálása

Az egyértelműsítő rendszer adaptálása során megoldandó első probléma az új, eddig a tanító anyagban nem látott címkék elérhetővé tétele a tagger lexikális és kontextuális modellje számára. A PurePos és minden más szófaji egyértelműsítő rendszer a tanulási fázisában a tanító anyagból a szófaji címke és a szó kontextusa alapján modellezi az adott szófaji kategória eloszlását. Így természetes módon, a tanítás során nem látott tagről semmilyen előzetes információval nem fog rendelkezni a modell. Mint ahogy azt a morfológia építésénél láttuk, a főnevek és melléknevek egy új kategóriáját vezettük be azon szavakra, melyek a latin morfológia szabályai szerint ragozandók. A morfológiához hozzáadott szavak jelentős hányadának csupán egyetlen elemzése van, s ha ez a fenti osztályok egyikébe tartozik, akkor bár az adott szóhoz ezen kategória fog tartozni, de az utána következő szavak címkézése során a kontextuális modell nem képes eloszlást rendelni. Továbbá, amikor egy szóhoz a HuMor több elemzést is ad, s ezek egyike egy újonnan létrehozott címke, akkor ehhez sem tartozik a megtanult modellek egyikében sem valószínűségi információ. Úgy találtuk, hogy a legjobb becslés, amit – egy új tanító anyag létrehozása nélkül – tehetünk, hogy a latin főneveket és mellékneveket a magyar főnevek eloszlásával becsüljük. (Így pl.: a *diagnosis* szó `[FN|lat][NOM]` címkéjét és a *sin.* szó `[MN|lat][NOM]` elemzését is az `[FN][NOM]` eloszlásával becsüljük.)

Az orvosi nyelvezet egyik sajátossága a rövidített szavak nagy mennyisége és változatos használata, nem beszélve ezek a normától különböző használatáról, helyesírásáról. Összehasonlításképpen: míg a Szeged Korpuszban a rövidítések a tokenek 0,36%-át teszik ki, addig az általunk javított anyag 8,49%-a rövidítés. Fontos különbség még, hogy ebben a speciális nyelvezetben az orvosok – eltérve a helyesírási normáktól – sokszor a toldalékokat nem kötik kötőjellel a rövidített szótőhöz, hanem egyszerűen lehagyják azt. (A tanító anyagban szereplő rövidítésekből a kötőjellel írottak aránya 9,36%, míg az egyértelműsítendőben 3,87%.) Pl.: a *jo*, *jo.*, *j. o.* „rövidítések" mindegyike a különböző kategóriájú *jobb oldal*, *jobb oldali*, *jobb oldalon* kifejezések bármelyikét jelentheti, az adott szövegkörnyezetben persze általában egyértelműen azonosíthatóan az egyikre utal.

A PurePos eredetileg sem a tanítási, sem pedig a címkézési fázisban nem kezeli különlegesen a rövidítéseket, mert egyrészt a norma szerint írott köznyelvi szövegeken a toldalékos alakok elemzésében nagy mértékben tud támaszkodni a POS-tageket tippelő suffix guesserre, másrészt általában nem kell ilyen mennyiségben és ennyire ad hoc módon létrehozott rövidítéstömeggel megbirkóznia. Ezzel a megközelítéssel jelen anyag esetén sokszor hibás következtetésre jut a tagger, így az alábbiak szerint módosítottuk a működését. A rendszer képes bizonyos előre definiált formai jegyeknek megfelelő szóalakokhoz külön lexikális eloszlást megtanulni, amit az alaprendszer a számjegyeket tartalmazó tokenekre, HTML entitásokra és írásjelekre alkalmaz. A fenti felsoroláshoz hozzáadtuk még

a toldalékolatlan alakú rövidítéseket, továbbá ezeket a tanítási fázisban elhagytuk a standard tokenekhez megtanult lexikális modellből. Így sikerült azt elérni, hogy a megtanult lexikális eloszlás ne az egyes tokenek eredetijéből fakadjon, hanem egy általánosabb, rövidítésekhez tartozóból. Mivel az adaptált morfológia számos rövidített alakot már ismer, ezért ezt a tudást is kívánatos volt alkalmazni. Az eredeti PurePos-ban a tanítóanyagban már látott szavak esetén az egyértelműsítő nem egyezteti a tanult tudást az integrált morfológiával, a rövidítések ilyen típusú kezelése, viszont szükségessé tette ezt. Az egyeztetés úgy történik, hogy a morfológia által javasolt latin típusú címkék a magyar megfelelővel való becsült valószínűséggel kerülnek be az egyértelműsítési folyamatba.

A fent bemutatott – a szófaji osztályok és a bizonyos tokenek reprezentációjának módosításával járó – doménadaptációs eljárással további javulást értünk el a taggelés területén, így 94,49%-os tokenszintű pontosságról számolhatunk be.

3.4. Hibaanalízis

Az összehasonlítási alapnak tekintett alaprendszer hibáit megvizsgálva, a hibákat az alábbi csoportokba lehet sorolni:

1. Az egyik leggyakoribb hiba, hogy a rövidítések hibás osztályba kerülnek, azok különleges írásmódja és nagyon változatos használata miatt. Ezen belül is tipikusan a főnévi és melléknévi szerepek keverése jellemző.
2. A hibák egy másik osztálya a latin, illetve latin eredetű kifejezések szófajának fentihez hasonló rossz meghatározása. Mivel ezen szóalakokat a korábban használt morfológiai elemző nem tudta megelemezni, így a guesserre maradt a feladat. A guesser rossz működése – a benne implementált tanulási algoritmus jellemzői miatt – nagyobbrészt a más doménen történő tanításból fakadnak.
3. A korpuszt alkotó orvosi szövegekben jellemző a melléknévi igenevek állítmányként történő használata, amely a köznyelvben meglehetősen ritka. Többek között ehhez kapcsolódóan a rendszer egyik gyakori hibaosztályát azok az esetek alkotják, amikor melléknévi igeneveket múlt idejű igékként annotál a rendszer. Ilyen tipikus rosszul elemzett szavak a *javasolt*, *kifejezett*, *igazolt*. Rendszeresen hibás analízist adott a PurePos a melléknévi igenév–melléknév ambiguitási osztály esetén is (pl.: *ismert*, *jelzett*). Hozzá kell tennünk, hogy ezeknek az eseteknek a megítélése a humán annotátorok számára is gyakran kétséges.
4. A fentieken kívül nagy számban vannak még jelen olyan hibák, melyek egyszerűen az orvosi nyelvhasználat egyediségéből fakadnak. Ilyen hibáson osztályozott szavak pl.: a *jobb*, mely a tanítóanyagban alapfokú melléknévként gyakorlatilag nem szerepel, vagy a *beteg*, melyet a tanulás során a PurePos soha nem látott főnévként. Ezen hibaesetek közös vonása, hogy a két korpuszban a kapcsolódó ambiguitási osztályok elemeinek eloszlása teljesen más.

Míg a 3.2 és 3.3 részekben részletezett megoldásokkal elsősorban az 1. és 2. típusú hibák javítását céloztuk meg, addig a 3. és 4. típusúak javításához szükségesnek látjuk a megtanult lexikai valószínűségek változtathatóságának a lehetőségét. Ehhez a továbbiakban úgy módosítjuk a PurePos rendszert, hogy a bemeneti mondatok egyes tokenjeinek elemzéseihez a címkézési folyamat segítésére a felhasználó által előredefiniált eloszlást rendelhessünk. Így a rendszer képessé válhat arra, hogy néhány egyszerű szabályt használva, nagyon gyakori tévesztések célzott javításával kis erőfeszítéssel nagy mértékben javítsuk az annotálás pontosságát. További tervünk, hogy a korpusz mellett további egyéb orvosi adatbázisokat is felhasználva olyan rövidítésfeloldó rendszert hozzunk létre, amely különösen a több elemből álló rövidítések esetében a jelenleginél jóval nagyobb pontossággal képes a rövidített szavak címkézésére.

4. Összegzés

Írásunkban ismertettük egy folyamatban lévő kutatási projekt aktuális állását, melynek részeként bemutattuk a rendelkezésünkre álló orvosi rekordokon végzett azon előfeldolgozási lépéseket, amelyeket szükségesnek véltünk egy gold standard korpusz létrehozásához. Azt is láttuk, hogy az így létrehozott eszközök egy későbbi orvosi rekordokra épülő szövegbányászati rendszer fontos építőkövei lehetnek. Bemutattuk azon lépéseket, amelyekkel a HuMor morfológiai elemzőt az orvosi doménre adaptáltuk, továbbá megvizsgáltuk, hogy az így előállt megnövekedett morfológiai tudást mily módon lehetséges mélyebben integrálni a PurePos morfológiai egyértelműsítő rendszerbe. Részletes hibaanalízist végeztünk, s a felmerülő hibák egy részére teljes, illetve részleges megoldást mutattunk be.

A jövőben folytatjuk a rendszer doménadaptálását, s ennek keretében a rövidítések kezelésére bevezetünk egy olyan alrendszert, mely prefixegyezés alapján statisztikai módszerrel próbálkozik a rövidítések feloldásával, hogy az azokhoz tartozó lexikális eloszlást a rövidített szó eredetijéből nyerjük ki. Célunk még, hogy folytassuk a manuális annotálást, hogy a PurePos elemzővel végzendő további doménadaptációs kísérletekhez megfelelő tanítóanyag is rendelkezésünkre álljon, illetve hogy korábban semmilyen szempontból sem látott tesztanyagon is validálhassunk eredményeinket.

Hivatkozások

1. Siklósi, B., Orosz, Gy., Novák, A., Prószéky, G.: Automatic structuring and correction suggestion system for Hungarian clinical records. In De Pauw, G., de Schryver, G.M., Forcada, M.L., M. Tyers, F., Waiganjo Wagacha, P., eds.: 8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages, Istanbul (2012) 29–34
2. Siklósi, B., Orosz, Gy., Novák, A.: Magyar nyelvű klinikai dokumentumok előfeldolgozása. In Tanács, A., Vincze, V., eds.: Magyar Számítógépes Nyelvészeti Konferencia 2011, Szeged (2011) 143
3. Novák, A.: Milyen a jó humor? In: Magyar Számítógépes Nyelvészeti Konferencia 2003, Szeged (2003) 138–145

4. Prószéky, G., Novák, A.: Computational Morphologies for Small Uralic Languages. In: Inquiries into Words, Constraints and Contexts., Stanford, California (2005) 150–157
5. Quinlan, J.R.: C4.5: Programs for Machine Learning. Volume 1 of Morgan Kaufmann series in Machine Learning. Morgan Kaufmann (1993)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software. ACM SIGKDD Explorations Newsletter **11**(1) (2009) 10
7. Miháczi, A., Németh, L., Rácz, M.: Magyar szövegek természetes nyelvi előfeldolgozása. In: Magyar Számítógépes Nyelvészeti Konferencia 2003, Szeged (2003) 38–43
8. Orosz, Gy., Novák, A.: PurePos – an open source morphological disambiguator. In Sharp, B., Zock, M., eds.: Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science, Wroclaw (2012) 53–63
9. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In: Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora LINC 2004 at The 20th International Conference on Computational Linguistics COLING 2004. (2004) 19–23
10. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the sixth conference on Applied natural language processing. Number i, Universität des Saarlandes, Computational Linguistics, Association for Computational Linguistics (2000) 224–231
11. Halácsy, P., Kornai, A., Oravecz, C.: HunPos: an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic, Association for Computational Linguistics (2007) 209–212
12. Baldridge, J., Morton, T., Bierner, G.: The OpenNLP maximum entropy package (2002)
13. Laki, L.J.: Investigating the Possibilities of Using SMT for Text Annotation. In: SLATE 2012 - Symposium on Languages, Applications and Technologies, Braga, Portugal, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2012) 267–283.
14. Fábián, P., Magasi, P.: Orvosi helyesírási szótár. Akadémiai Kiadó, Budapest (1992)
15. Cavnar, W.B., Trenkle, J.M.: N-Gram-Based Text Categorization. Ann Arbor MI **48113**(2) (1994) 161–175