NYELVTECHNOLÓGIAI ALGORITMUSOK KORPUSZOK AUTOMATIKUS ÉPÍTÉSÉHEZ ÉS PONTOSABB FELDOLGOZÁSUKHOZ

DOKTORI (PH.D.) DISSZERTÁCIÓ

Endrédy István

Témavezető:
Dr. Prószéky Gábor,
az MTA doktora

PÁZMÁNY PÉTER KATOLIKUS EGYETEM

INFORMÁCIÓS TECHNOLÓGIAI ÉS BIONIKAI KAR

ROSKA TAMÁS MŰSZAKI ÉS TERMÉSZETTUDOMÁNYI DOKTORI ISKOLA

Budapest, 2016

1 Bevezetés

Jelen értekezés témája interdiszciplináris terület: a nyelvészet és az informatika közös pontjaira épül. Egyrészről szükség van az informatika módszereire és eszközeire ahhoz, hogy nagy mennyiségű szöveget gépileg feldolgozzunk, másrészt mindezt nyelvészeti céllal szeretnénk tenni. Az emberi nyelvtechnológia számos kutatásához szükséges, hogy nagy méretű, valós, ember által írt szöveg (korpusz) álljon rendelkezésre. A korpusz egyik fontos tulajdonsága a mérete, azonban a használhatóságát nagyban befolyásolja a kiegyensúlyozottsága: a szövegei mennyire reprezentatívak. A többféle forrásból és változatos műfajból származó szövegekből felépülő korpusz jobban reprezentálja a nyelvet, általános nyelvészeti kutatásokat jobban ki tud szolgálni. Arra vállalkozom, hogy bemutassam, hogyan lehet automatizáltan, webes forrásból nagy korpuszokat építeni, majd pedig az ezekben végzett kutatásokat mutatom be.

Az elérhető magyar korpuszok száma örvendetes módon egyre növekszik. Az egyik legnagyobb a 2004-ben készült a BME MOKK webkorpusza (Halácsy és mtsai. 2004), melynek mérete 600 millió szó. A Magyar Nemzeti Szövegtár a maga 187,5 millió szavas méretével ugyan ennél kisebb, de a tudatosan válogatott tartalom teljes mértékben szófaj-egyértelműsített (Váradi 2002). Ennek a korpusznak a felújított változata, az MNSZ2 folyamatosan növekszik (Oravecz, Váradi, és Sass 2014), jelenleg 784 millió tokenes.

Azonban azzal kellett szembesülnünk, hogy a magyar korpuszok világában eddig nem volt elérhető olyan igazán nagy méretű (milliárd tokenes) korpusz, amelyik mind méretével, mind többféle annotációjával performancia alapú elemzést megcélzó kutatási projektünk igényeit (Prószéky és Indig 2015) megfelelően ki tudja szolgálni. Az internet alapján épített, jó minőségű korpusz lehetővé teszi a korpusz méretének állandó növekedését, és lehetőséget ad arra is, hogy megvizsgáljuk az adott nyelv gyakori struktúráit és ezek időbeli változásait. Úgy éreztük, hogy szükség van egy nagy, átfogó, annotált és folyamatosan frissített adatbázisra.

A dolgozat érinti a korpuszépítés, a szótövesítés, az ékezetesítés, a főnévicsoport-felismerés és a mondatvázak témáját is. A munkafolyamat korpuszvezérelt volt: az ötleteket a korpuszból merítettük, és vele is ellenőriztük.

A dolgozat első részében a **korpuszépítéssel foglalkozom**. Egy letöltő robotot (crawler) készítettem egy olyan **szövegkinyerő algoritmus**sal, amely képes az egyes weboldalak egyedi sajátosságait automatikusan megtanulni. Az elérhető megoldások (Pomikálek 2011; Kohlschütter, Fankhauser, és Nejdl 2010) a HTML tartalom egyes jegyeiből (linkek, stopwordök, szavak és címkék száma alapján)

határozzák meg értékes részt. Ez jó minőséget biztosít, azonban az oldalakon számos olyan szöveg is található, amelyet ezen algoritmusok - tévesen - a fő tartalomhoz vesznek (kapcsolódó cikkek, kommentek, egyéb vezető hírek stb.). Ezek nem csak duplikátumok lesznek az épített korpuszban, hanem a szöveg kohézióját is rontják. Az erre a célra kifejlesztett Aranyásó algoritmus a szöveg jegyeinél magasabb szinten dolgozik: egy weboldal azon tipikus pontjait keresi és tanulja meg, ahol a fő tartalom szokott lenni. Ha ezt sikerült meghatározni, akkor a szövegkinyerő algoritmusoknak már csak ezt a lényeges részt adja át. Ezzel a módszerrel sikerült jobb eredményt elérni.

Az épülő korpusz pontosabb feldolgozásához szükség volt **lemmatizáló**ra. Az általam kifejlesztett **lemmatizáló minőségének méréséhez egy kiértékelési módszerre és alkalmazásra** volt szükség.

A dolgozat második felében **a főnévi csoport felismerését kutattam.** A főnévicsoport-felismerés fontos lépés a mondatelemzés során. A legjobb magyar eredményt produkáló HunTag eszköz továbbfejlesztett, moduláris változatával (HunTag3) lehetőségem nyílt több irányt is kipróbálni. Viterbi, bigram- és trigramátmenet, illetve CRFsuite vizsgálata, háromféle kódkészlet (MSD, KR, Humor) kipróbálása során magyar nyelvre a legjobb modellnek a trigramátmenet bizonyult. Majd számos futtatás során kikísérletezett új jegyek hozzáadásával sikerült több, mint 3% javulást elérni. A tesztadaton nem történt hibaelemzés, az eredmények szignifikanciáját pedig crossvalidációval ellenőriztem. A hibaelemzés és a mérések azt mutatták, hogy a legtöbb hiba a szomszédos NP-k felismerésében (hol tévesen egybe, hol tévesen külön), a birtok és a birtokos téves szétválasztásában vagy másik NP-vel való téves összevonásában volt. Ezeken történő javítások eredményezték a legjelentősebb javulást.

Angol nyelv esetén a CRFsuite hozta a legjobb eredményt, de ez nem lépte át az eddigi legjobb angol publikált eredményeket.

2 Új tudományos eredmények összefoglalása

A **fő eredményem**nek tartom az **Aranyásó algoritmus**t, amely weblapokról öntanuló módon nyeri ki a értékes tartalmat; a **lemmatizáló**t, amely számos termékben alkalmaznak és amelyet a **tövesítő kiértékelés**ek legjobbnak mértek magyar nyelvre; a **magyar főnévi csoport felismerés**ben elért eredményeket, és az elkészült **Pázmány Korpusz**t.

A dolgozatban bemutatott eredményeket a három téziscsoportra lehet osztani. Az első téziscsoport a korpuszépítésre, illetve az ennek során használt, weboldalak értékes szövegének kinyerését végző algoritmusra vonatkozik. A második téziscsoportban létrehoztam a - Humor elemzőre épülő - tövesítőt,

ékezetesítőt és ezek kiértékelését végző módszert. Végül a harmadik téziscsoport a főnévicsoport-felismerés terén végzett újításokat foglalja magában.

I. TÉZISCSOPORT

Az első téziscsoportban azzal a problémával foglalkoztam, miként lehet algoritmikusan kinyerni egy weblapról az értékes szöveges tartalmat. Egy weboldal számos ismétlődő és irreleváns sablonos tartalmai megnehezítik a fő szöveges tartalom kiszűrését. A menük, fej- és láblécek, reklámok, a minden oldalon ismétlődő struktúra nem csak doménenként, olykor aldoménenként, hanem időben is változhat. Így olyan algoritmust kellett létrehozni, ami ezt öntanuló módon követni tudja, emberi beavatkozás nélkül. A jelenlegi sabloneltávolító algoritmusoknál sikerült jobb eredményt elérni az Aranyásó (GoldMiner) algoritmussal (1. táblázat).

Az Aranyásó alapján működő szövegkinyerő az egyes domének egyedi sajátosságait egy mintavétel alapján tanulja meg: megjegyzi azt a leggyakoribb HTML címkesorozatot, amely közrefogja a legtöbb értékes tartalmat.

Algoritmus		**Kinyert mondatok száma**	**Egyedi mondatok száma**	**Összes karakterszám**	**Egyedi mondatok karakterszáma**	**Egyedi mondatok aránya**	**Egyedi mondatok aránya karakterszámban**
origo.hu	összes szöveg	264 423	63 594	16 218 753	7 048 011	24%	43%
	BTE	60 682	33 269	12 016 560	7 499 307	54%	62%
	JusText	58 670	30 168	8 425 059	4 901 528	51%	58%
	Aranyásó	22 475	21 242	3 076 288	3 051 376	**94%**	**99%**
nol.hu	összes szöveg	509 408	144 003	25 358 477	12 570 527	28%	49%
	BTE	154 547	107 573	24 292 755	13 544 130	69%	55%
	JusText	186 727	128 782	14 167 718	11 665 284	68%	82%
	Aranyásó	162 674	123 716	12 326 113	11 078 914	**76%**	**89%**
index.hu	összes szöveg	232 132	55 466	9 115 415	4 542 925	23%	49%
	BTE	51 713	26 176	5 756 176	4 061 697	50%	70%
	JusText	40 970	29 223	4 371 693	3 441 337	71%	78%
	Aranyásó	13 062	11 887	1 533 957	1 489 131	**91%**	**97%**

1. táblázat. Sablonszűrő algoritmusok összehasonlítása az egyes doméneken

Egy crawler is készült, ami az Aranyásó segítségével gyűjtötte a magyar nyelvű szövegeket a webről.

A letöltött szövegekből épült a 1,2 milliárd szavas Pázmány korpusz, amely a több mint 30 000 domén szövegeit tartalmazza (2. táblázat). A további felhasználás érdekében a korpuszból különválogattam a kommenteket, mert ezek eltérő jellegűek és szerkezetűek.

alkorpusz	tokenszám	mondatszám	NP-szám
fő korpusz	954 298 454	48 536 849	223 347 534
egyéb tartalmak	228 806 919	15 802 499	52 865 889
kommentek	58 985 126	3 505 818	13 867 066
összesen	**1 242 090 499**	67 845 166	290 080 489

2. táblázat. A Pázmány korpusz összetétele

1. tézis: Megalkottam a GoldMiner algoritmust, ami az eddigieknél hatékonyabban nyeri ki a weblapokról a cikkeket.

1.a tézis: Létrehoztam egy új crawlert, amely korpuszt tud építeni.

Kapcsolódó publikációk: [1], [9]

7. tézis: A crawler segítségével létrehoztam az 1,2 milliárd tokenes Pázmány Korpuszt.

II. TÉZISCSOPORT

A második téziscsoportban egy lemmatizálót illetve egy ékezetesítőt készítettem, mely előbbi kiértékeléséhez egy módszert és alkalmazást is megalkottam. A Humor elemzőre (Prószéky és Kis 1999) épülő ékezetesítő és lemmatizáló az elemzés morfémáiból számolja ki az eredményt.

A lemmatizáló algoritmus lényege az, hogy a morfcímkék által meghatározott szerep figyelembevételével számolja ki a szó tövét. Balról jobbra haladva az elemzésben szereplő morfémák felszíni alakját használjuk fel a tő építéséhez, kivéve az utolsó tőalkotó morfémát: ennek a szótári alakja szerepel a tőben. Természetesen kérdés, hogy mely morfémák számítanak tőalkotónak, és melyek nem. A képzők esetén ez különösen fontos kérdés. Például az *adósság* szó esetén a *-ság* képzőt levágva *adós* lesz a tő. De ha az *-s* illetve az *-ó* képzőt is levágjuk, akkor már *adó* illetve *ad* lesz a szó töve. Látható, hogy meghatározó a végeredmény tő szempontjából, hogy egy adott képzőt tőalkotónak tekintünk-e vagy sem. Általában érdemes figyelembe venni, hogy az adott alkalmazás szempontjából mi lehet a legkedvezőbb megoldás. Minél több képző levágása növelheti a tövesítés fedését (kevesebb potenciálisan releváns

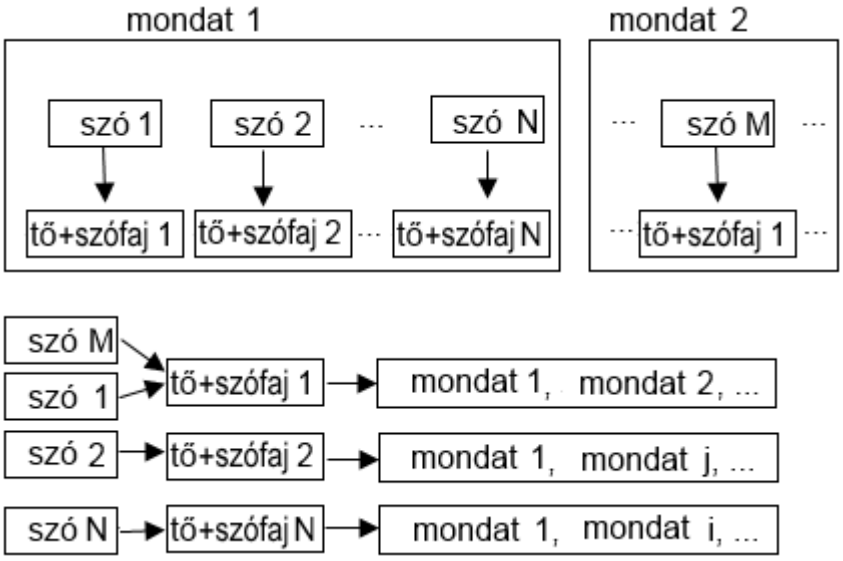
találatot veszítünk el), de ronthatja a pontosságát (több nem releváns találat áll elő).

A lemmatizáló több tulajdonsága konfigurációs fájlból hangolható: a tőalkotó morfémák, a képzők eredő szófajai, címkekonverziók, vagy akár a tő meghatározásának az algoritmusa is paraméterezhető, testre szabható.

A lemmatizáló kiértékelését két módon is elvégeztem: a közvetlen kimenetét illetve egy keresőrendszerbeli teljesítményét is megmértem. A közvetlen kiértékelés lehetővé tette, hogy gold standard lemmákhoz képest mérjem az egyes megoldások minőségét. Az IR kiértékelés ezzel szemben nem várta el a tövesítőktől, hogy a szótári tövet meghatározzák (stemmerektől ez nem is várható el), hanem itt elegendő, ha - a lemmapontosságtól eltekintve - jól vezeti vissza a ragozott alakokat egy közös alakra. Ez egyrészt képes megmutatni, ha egy adott nyelvre elegendő csak stemmert készíteni, lemmatizálóra nincs szükség, másrészt ez egy valós feladat közben méri a tövesítőket, nem laboratóriumi körülmények között. Mindkét kiértékelés gold standardjét lemmával annotált korpuszokból automatizáltan állítottam elő. Az eredmények összehasonlíthatóságához tíz tövesítőn (lemmatizálók és stemmerek vegyesen) végeztem el a kiértékelést, három nyelven (angol, lengyel, magyar).

A tövesítő közvetlen kimenetén végzett kiértékelésben többféle metrikát alkalmaztam. Megmértem a tőalternatívák első elemének a lemmapontosságát, továbbá több szempontból azt, hogy az egyes tőalternatívák önmagukban mennyire jók, illetve ezek sorrendezése mennyire helyes. Pontosabb képet adhatunk egy tövesítőről, ha tudjuk, hány százalékban helyes az első töve, mennyi hibás tőalternatívát képes előállítani vagy a töveit mennyire jó heurisztika alapján sorrendezi.

A tövesítők IR alapú kiértékelése tipikusan kézzel összeállított dokumentum-lekérdezés halmazok alapján szokott történni(Hull 1996; Tordai és De Rijke 2006; Halácsy és Trón 2007). Kiértékelésem arra az ötletre épül, hogy minden, tövekkel annotált korpusz gépi úton átalakítható IR kiértékelő gold standarddé. A módszer alapja az, hogy a korpusz mondatai (egy vagy több mondata) lesznek a IR találati egységek (dokumentumok), az egyes mondatok szavai pedig az ezekhez a dokumentumokhoz (mondatokhoz) tartozó lekérdezések. Az IR gold standard úgy generálható ki egy tövekkel annotált korpuszból, hogy a benne szereplő összes szóhoz tartozó (*mondatazonosító, eredeti szóalak, szótő, szófaj*) sorozat alapján elkészítjük azt a listát, ami tartalmazza az egyes szóalakok tő+szófaj értékeit, és azt, hogy ezek milyen mondatokban fordulnak elő. Más szavakkal, kilistázzuk a korpuszból a literális szóalakokat és a hozzájuk (tő és szófaj információn keresztül) kapcsolódó mondathalmazokat (lásd 1. ábra). Így bármelyik szóalakhoz meg tudjuk mondani, hogy mely mondatokban fordul elő. Ezek a szó–mondat összerendelések jelentik a gold standardet, amihez képest a tövesítők kiértékelhetők. A pontosság és fedés az alapján számítható, ha összevetjük egy adott tövesítő egyes szavakra adott találatait és a gold standardet.



1. ábra. Tövesítők IR-alapú kiértékelése tővel annotált korpusszal: mondatok=dokumentumok és szavak=lekérdezések, az eredmény mondathalmazok a gold standarddal kiértékelhető

Ez a kiértékelés egyrészt automatikusan állítja elő a gold standardet korpusz alapján, ezért a tesztesetek száma jóval magasabb (British National Corpus esetén 2 millió).

Másrészt két kiértékelést is elvégez: egyrészt kiértékeli az összes találatot, másrészt külön kiértékeli az első *n* találatot is. Ez utóbbi lehetőség azt a célt szolgálja, hogy szimulálja az emberi kiértékelést: kereséskor az első találatoknak nagyobb a súlya, ritkán szoktuk a sokadik találati oldalt is végignézni. Így az ott lévő esetleges hibás találatok nem annyira zavaróak. A tesztek során derült ki, hogy – bár nem ezt a célt szolgálta – az első *n* találat kiértékelése arra is alkalmas, hogy magának az IR-rendszernek a sorrendező algoritmusát (*ranking*) kiértékelje. Ha nem ebben a találati tartományban vannak a TP találatok, akkor nem jó a *ranking*.

alkorpusz	***stemmer nélkül***	***Hunspell*** *(első tő)*	***Hunmorph-foma*** *(leghosszabb tő)*	***Hunmorph-compound*** *(első tő)*	***Hunmorph*** *(első tő)*	***Ocastem***	***Snowball***	***Humor*** *(leghosszabb tő)*
szépirodalom	52,4	86,6	76,2	86,6	86,4	**88,7**	58,3	88,4
tanulók	52,9	**88,6**	78,1	88,2	88,1	88,0	57,0	88,3
újságcikk	57,3	84,5	75,5	83,1	81,8	88,6	64,7	**92,8**
IT	57,9	81,9	75,8	81,7	79,3	87,9	68,6	**92,5**
jogi	62,0	81,8	77,4	82,4	80,8	86,7	72,4	**93,8**
üzleti	55,5	78,1	68,9	80,2	78,9	87,6	65,2	**91,4**
összesen	56,2	83,9	75,6	84,0	83,0	87,9	64,0	**91,0**

3. táblázat. Tövesítő modulok első/leghosszabb javaslatának pontossága a Szeged Korpuszon

domén	tövesítő nélkül	Hu-light	Snowball	Hunspell	Humor
szépirodalom	25,2	61,9	66,7	67,8	**78,3**
tanuló	14,0	55,5	56,3	69,0	**75,4**
újság	16,5	79,6	81,1	77,1	**85,9**
IT	19,7	71,7	73,8	73,4	**81,8**
jogi	18,3	52,6	53,4	70,1	**75,3**
üzleti	23,2	73,1	73,5	44,4	**87,9**
összesen	18,4	65,6	67,3	66,3	**80,1**

4. táblázat. Tövesítők IR-kiértékelése magyar nyelvre a Szeged TreeBank mondatai alapján (Lucene-motorral)

Magyar nyelv esetén az elkészült lemmatizáló bizonyult a legjobbnak mindkét kiértékelésben (3. táblázat, 4. táblázat). IR kiértékelésnél 80,1%, lemmapontosságnál 91,0%, tőalternatívák kiértékelésénél 91-94% F értékekkel. Az erősen ragozó magyar nyelv tulajdonsága, hogy tövesítő nélkül 18% eredményt lehet elérni.

domén	token	tövesítő nélkül		Stempfel		Morfologik		Hunspell		Humor	
		F	oov	F	oov	F	oov	F	oov	F	oov
szépirodalom	54.205	53,80	0	78,36	2,7	**89,34**	2,7	84,45	5,6	88,71	7,8
tájékoztató	56.779	54,08	0	77,31	3,9	**89,09**	3,9	84,38	7,7	87,87	5,8
párbeszédek	59.024	68,26	0	72,60	9,5	**83,62**	9,5	78,58	11,3	79,63	12,8
regény	169.270	55,81	0	77,48	4,1	**88,98**	4,1	84,38	5,9	87,84	4,2
beszélt rádió	23.303	64,52	0	73,94	6,4	**86,11**	6,4	82,12	8,1	83,43	10,2
kutatás oktatás	20.229	50,23	0	79,56	3,6	**89,64**	3,6	85,90	7,1	89,38	5,8
internet	72.273	55,27	0	77,15	3,2	**88,78**	3,2	83,48	8,0	86,02	7,8
újság	506.214	51,78	0	79,23	2,8	**90,03**	2,8	86,10	6,1	89,16	4,6
írott parlamenti	66.315	51,24	0	78,77	4,3	89,85	4,3	85,48	7,4	**89,86**	2,7
eszközök	30.998	52,22	0	80,82	1,2	90,65	1,2	87,48	7,5	**91,38**	2,3
nem osztályozott	10.140	54,38	0	78,72	2,0	**90,11**	2,0	85,55	3,9	88,52	3,9
összesen	1.028.671	54,21	0	78,2	3,6	**89,25**	3,6	85,02	6,7	88,08	5,4

5. táblázat. A lemmapontosság kiértékelése lengyel nyelvre a PNC alapján

domén	tövesítő nélkül	Stempel	Morfologik	Hunspell	Humor
szépirodalom	47,5	64,3	67,3	**71,6**	70,2
tájékoztató	46,3	67,1	70,6	**73,9**	73,8
párbeszédek	56,8	60,3	59,8	**64,4**	59,4
regény	41,0	57,5	61,5	**65,4**	64,1
beszélt rádió	60,2	68,0	65,3	**71,6**	67,1
kutatás oktatás	55,0	76,9	80,1	**82,5**	81,5
internet	50,0	63,4	65,3	**70,3**	67,3
újság	30,7	58,7	64,6	**67,5**	67,4
írott parlamenti	42,9	66,7	79,9	82,6	**83,1**
eszközök	47,8	77,2	81,2	83,3	**85,0**
nem osztályozott	64,7	73,3	72,3	**77,9**	76,5
összesen	36,9	60,4	65,3	**68,7**	68,0

6. táblázat. Tövesítők IR-kiértékelése lengyel nyelvre a PNC mondatai alapján (Lucene-motorral)

A lengyel eredmények azt mutatták, hogy a tövesítés minősége nagy hatással van az IR kiértékelés F-mértékére: az IR minőségét kétszeresen is javíthatja egy tövesítő (6. táblázat). A lengyel nyelv esetén a Morfologic stemmer volt a legjobb a lemmapontosságban (5. táblázat), az IR kiértékelésben pedig a Hunspell érte el a legjobb eredményt.

2. tézis: Létrehoztam egy ékezetesítő modult, tervezésében egy szerzőtárssal együttműködve, ami egy módosított Humor-lexikonnal 94.3% pontossággal képes ékezet nélküli szövegek ékezetesítésére.

Kapcsolódó publikációk: [2], [11]

3. tézis: Létrehoztam egy lemmatizáló motort, tervezésében egy szerzőtárssal együttműködve, ami a Humor elemzéseiből kiszámolja a szó tövét, és ez a kiértékelések alapján a legjobb eredményeket adja a magyar nyelvre.

Kapcsolódó publikációk: [2], [4]

4. tézis: Létrehoztam egy kiértékelési módszert, amellyel tetszőleges, lemmával annotált korpusz alapján lemérhető egy tövesítő (i) pontossága, (ii) IR minősége, (iii) UI, OI, ERRT értéke, és (iv) egyéb metrikák szerinti kiértékelése. Mindezt angol, lengyel és magyar nyelvre 10 tövesítőre kimértem.

4.a tézis: Létrehoztam egy módszert, amivel korpuszból automatikusan létrehozható olyan gold standard, amely tövesítők közvetlen illetve IR-rendszerbeli kiértékeléséhez használható.

4.b tézis: A kiértékelést angol, lengyel és magyar nyelvekre elkészítettem

4.c tézis: Kimutattam, hogy az erősen ragozó nyelveknél (lengyel, magyar) a lemmapontosság és az IR-minőség korrelál.

4.d tézis: Kidolgoztam egy IR-tövesítőkiértékelést, ami natív IR rendszer nélkül képes IR-rel korreláló eredményt mérni.

4.e tézis: Definiáltam a tövesítő közvetlen kimenetére olyan kiértékeléseket, amelyek alkalmasak a tövesítők összehasonlítására, és visszajelzést is adhatnak a tövesítő hibáinak feltárására és azok kijavítására

4.f tézis: Kimutattam és megmértem, hogy az IR-tövesítőkiértékelés első *n* találatának kiértékelése alkalmas az IR-ranking algoritmus kiértékelésére is.

Kapcsolódó publikációk: [2], [4]

III. TÉZISCSOPORT

A dolgozat harmadik téziscsoportjában a főnévicsoport-felismeréssel foglalkozom. Egyrészről a felismerés pontosságát javító jegyek definiálását és finomítását kutattam. Létrehoztam egy eszközt, ami egy tanítóminta jegyeinek hasznosságát képes mérni és segíti a hangolásukat azzal, hogy áttekintést ad a jegyek és a kimeneti (IOB) címkék közötti kapcsolatra. Nem képes önállóan javítani a jegyeket, csupán egy segédeszköz szeretne lenni a nyelvész számára. A statisztika erejét és a nyelvész intuícióját kapcsolja össze.

Másrészről a HunTag hibaelemzésével, új jegyek definiálásával és a HunTag3 alkalmazásával sikerült az eddigi legjobb magyar NP felismerési eredményt elérni. A hibaelemzés és a mérések azt mutatták, hogy a legtöbb hiba a szomszédos NP-k felismerésében (hol tévesen egybe, hol tévesen külön), a birtok és a birtokos téves szétválasztásában vagy másik NP-vel való téves összevonásában volt. Ezeken történő javítások eredményezték a legjelentősebb javulást. Magyar esetén a trigram átmeneti modell bizonyult a legjobbnak, mindezek együttesen 93,59%-os eredményt hoztak, amint a 7. táblázat mutatja. Az eddig publikált legjobb eredmény 90,28% volt (Recski 2014). Angol esetén a CRFsuite megoldással mértem a legjobb eredményeket, azonban ez nem lépte át az eddigi legjobb eredményt (SS05: 95,23%).

	MSD	KR	KR + jobb jegyek
T'n'T	68,52	70,95	-
baseline	81,71	88,72	-
HunTag	93,20	88,96	90,78
HunTag3 – bigram	93,43	89.10	90,72
HunTag3 – trigram	**93,59**	**89,83**	**91,50**
CRF	92,27	89,12	89,77

7. táblázat. NP-felismerési eredmények, magyar nyelvre, F-érték különféle kódkészletekkel (Szeged Treebank)

5. tézis: Kimutattam, hogy az NP-felismeréshez használt jegykészlet tovább javítható, ha (i) a jegyek kevés értéket vehetnek fel, és (ii) ha létezik olyan finomabb osztályozás, amely mellett a kimeneti címkékkel jobban korreláló jegyeket kapunk.

5.a tézis: Az angol nyelv esetében a szófaj kategóriát olyan alkategóriákra bontottam, amelyek jobban korrelálnak az IOB-címkékkel (+2% javulás).

5.b tézis: Létrehoztam egy módszert, amivel az 5.a. tézisben leírt művelet a WordNet synsetjeinek segítségével gépi támogatással elvégezhető.

5.c tézis: Becslést adtam egy jegyhalmaz hasznosságára a címke- és a jegykorreláció kiszámolásával, méghozzá egy tanulóalgoritmus (NLTK unigram, bigram chunker; HunTag; SS05) futási idejénél gyorsabban.

5.d tézis: Kimutattam, hogy nagyon fontos az input annotáció szerepe, ami akár fontosabb lehet, mint maga a gépi tanuló algoritmus. Bármilyen külső információ segít, ami „korrelál” a leendő kimeneti annotációval.

5.e tézis: Kimutattam, hogy kevés olyan szószintű jegy van, amely a kimeneti címkével 100%-osan korrelál, viszont ezek mindegyike fontos.

Kapcsolódó publikációk: [5], [6]

6. tézis: Sikerült javítanom a főnévi csoport felismerésének minőségét új jegyek definiálásával, egy társszerzővel létrehozott trigramátmeneti modellel és a HunTag3 alkalmazásával, amelyek együttesen az eddigi legjobb magyar NP-felismerést eredményezték. (93,59%)

6.a tézis: Kimutattam a főnévicsoport-felismerésnél, hogy a trigram-átmenetvalószínűségi modell csak részletesebb, háromnál több elemű (típusos, vagy finomabb felosztású) IOB címkék esetén tud többet adni, mint a bigram-modell.

Kapcsolódó publikációk: [6]

2.1 Az eredmények alkalmazási területei

Az elkészített tövesítő modul számos cég alkalmazásaiba beépült: erre épül a Microsoft Indexing Service, az Országos Atomenergetikai Hivatalban tárolt dokumentumok tárolására és keresésére szolgáló rendszer, az MTI szerkesztőségi rendszere és a PolyMeta kereső. Az MTA Nyelvtudományi Intézete által készített Magyar Nemzeti Szövegtár második bővített kiadásának (MNSZ2) morfológiai annotálása ugyancsak ezzel az eszközzel készült.

A második téziscsoportban bemutatott tövesítő kiértékelési módszer angol, lengyel és magyar nyelveken túl más nyelvekre és további tövesítő modulok alapos vizsgálatára is használható. A kiértékeléshez készített alkalmazás felhasználásával más nyelvekre is automatikusan elkészíthető lemmapontosság- és IR minőség mérésére alkalmas gold standard. A kiértékelő script pedig további tövesítőkre is lefuttatható.

A dolgozat során készült 1,2 milliárd tokenes, lemmával, szófajilag és főnévi csoportokkal annotált Pázmány korpusz hasznos alapanyaga lehet további nyelvtechnológiai kutatásoknak.

Köszönetnyilvánítás

Szeretnék köszönetet mondani azoknak, akik nélkül ez nem sikerült volna.

Köszönöm felsős magyartanáromnak, Bukovits Mártának, aki a nyelvtant világosan tudta tanítani, és azt a benyomást keltette, hogy a nyelvtanban matematikai pontosság létezhet. Köszönöm Naszódi Mátyás egyetemi tanáromnak, hogy bevezetett a számítógépes nyelvészetbe és a MorphoLogicba. Az itt leírt munkáim és eredményeim szorosan kapcsolódnak a MorphoLogicban eltöltött 13 évemhez. Ezért a disszertációm a MorphoLogic előtti tisztelgésnek is tekinthető. Hálás vagyok a cégnél töltött évekért, sokat tanultam a munkatársaktól: Kis Balázs, a hatékony; Tihanyi Laci, a gyors; Pál Miki, az alapos; Kundráth Peti, a „kóder”; Novák Attila, a tudós; Sebestyén Zsolt, akire mindig számíthattam; Hubay Kati, a problémamegoldó; Aggod Andi, a kitartó tesztelő; Kincse Szabi, a kommunikátor, és mindenki más, akikkel jó volt együtt dolgozni.

Köszönöm témavezetőmnek, Dr. Prószéky Gábornak az atyai bátorítást és a jó légkörű beszélgetéseket. Bármikor örömmel fogadott, még ha nagyon elfoglalt volt is. Pótolhatatlan volt a szerepe ebben a dolgozatban.

Köszönöm a Bírálóknak az értékes észrevételeiket, javaslataikat.

Köszönöm családomnak, feleségemnek, Orsinak, gyermekeimnek, Balázsnak, Katának, Dorkának és Bencének, hogy támogattak ebben a munkában, és türelmesen elfogadták, hogy emiatt nem tudok velük annyi időt tölteni. (9 éves Dorka lányomnak ígértem, ha elkészülök ezzel a munkával, építünk együtt egy műhelyt a pincében. Kétnaponta megkérdezte, hogy hogyan állok, mikor kezdhetjük.)

Nem tudom hogyan megköszönni szüleimnek az életemet, és ahogy figyelemmel kísérték ezt az ügyemet is. Testvéreimnek ahogy szeretnek.

Köszönöm a doktoranduszt társaknak is a közös projekteket, az utazásokat. Elévülhetetlen érdeme van Indig Balázsnak, aki – a tudta nélkül – átsegített a holtponton, ahol fel akartam adni, és Dr. Wenszky Nórinak, hogy mindig készségesen lektorálta az írásaimat.

Köszönöm a Doktori Iskola korábbi és jelenlegi vezetőinek, Dr. Roska Tamás, Nyékyné Dr. Gaizler Judit és Dr. Szolgay Péter dékánoknak, hogy lehetővé tették és több módon is támogatták kutatói munkámat. Dr. Vida Katinkának a kiemelt figyelmet, az egyetem többi munkatársának is a háttérmunkát.

Köszönöm a szerda déli miséket az egyetem kápolnájában.

3 A szerző publikációi

Folyóiratcikk

[1] **Endrédy, István**, Attila Novák. 2013. “More Effective Boilerplate Removal—The GoldMiner Algorithm.” *Polibits Journal* 48: pp. 79–83.

[2] **Endrédy István**, Novák Attila. 2015. “Szótövesítők összehasonlítása és alkalmazásaik. ” In: Navracsics Judit (szerk.) *Alkalmazott Nyelvtudomány,* XV. évfolyam, 1-2. szám, pp. 7-27, Veszprém

Könyvfejezet

[3] Indig Balázs, **Endrédy István**. 2016. “Gut, Besser, Chunker - Selecting the best models for text chunking with voting” In: A. Gelbukh (Ed.) *Lecture Notes in Computer Science*: Computational Linguistics and Intelligent Text Processing Springer International Publishing, Berlin Heidelberg (*megjelenés folyamatban*)

Külföldi konferenciakötet

[4] **Endrédy István**. 2015. “Corpus based evaluation of stemmers”, *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 234-239, Poznań

[5] **Endrédy István**. 2015. “Improving chunker performance using a web-based semi-automatic training data analysis tool”, *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 80-84, Poznań

[6] **Endrédy István**, Indig Balázs. 2015. “HunTag3, a general-purpose, modular sequential tagger – chunking phrases in English and maximal NPs and NER for Hungarian”, *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 213-218, Poznań

[7] **Endrédy, István**. 2014. “Hungarian-Somali-English Online Dictionary and Taxonomy.” In *Proceedings on “Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era”* 38–43. Reykjavik, Iceland

[8] **Endrédy, István**, László Fejes, Attila Novák, Beatrix Oszkó, Gábor Prószéky, Sándor Szeverényi, Zsuzsa Várnai, and Beáta Wagner-Nagy. 2010. “Nganasan–Computational Resources of a Language on the Verge of Extinction.” In *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (LREC 2010),* pp. 41-44 Valetta, Malta

Hazai konferenciakötet

[9] **István Endrédy,** Novák Attila. 2012. “Egy hatékonyabb webes sablonszűrő algoritmus – avagy miként lehet a cumisüveg potenciális veszélyforrás Obamára nézve.” In: *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pp 297–301. SZTE, Szeged

[10] Bakró-Nagy Marianne, **Endrédy István**, Fejes László, Novák Attila, Oszkó Beatrix, Prószéky Gábor, Szeverényi Sándor, Várnai Zsuzsa, Wagner-Nagy Beáta. 2010. “Online morfológiai elemzők és szóalakgenerátorok kisebb uráli nyelvekhez”. In: *VII. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 345–348, SZTE, Szeged

[11] Novák, Attila, **István Endrédy**. 2005. “Automatikus Ë-jelölő program.” In: *III. Magyar Számítógépes Nyelvészeti Konferencia*, pp 453–54. SZTE, Szeged

4 Irodalomjegyzék

Halácsy, Péter, András Kornai, László Németh, András Rung, István Szakadát, és Viktor Trón. 2004. „Creating open language resources for Hungarian.” *Proceedings of 4th Conference on Language Resources and Evaluation (LREC)*, 203–10.

Halácsy, Péter, és Viktor Trón. 2007. „Benefits of resource-based stemming in Hungarian information retrieval.” In *Evaluation of Multilingual and Multi-modal Information Retrieval*, 99–106. Springer.

Hull, David A. 1996. „Stemming algorithms: A case study for detailed evaluation.” *JASIS* 47 (1): 70–84.

Kohlschütter, Christian, Peter Fankhauser, és Wolfgang Nejdl. 2010. „Boilerplate detection using shallow text features.” In *Proceedings of the third ACM international conference on Web search and data mining*, 441–50. WSDM ’10. New York, NY, USA: ACM. doi:10.1145/1718487.1718542.

Oravecz, Csaba, Tamás Váradi, és Bálint Sass. 2014. „The Hungarian Gigaword Corpus.” In *Proceedings of LREC*. Reykjavik.

Pomikálek, Jan. 2011. „Removing Boilerplate and Duplicate Content from Web Corpora.” PhD dissertation, Masaryk University, Faculty of Informatics.

Prószéky, Gábor, és Balázs Kis. 1999. „A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages.” In *ACL*, szerkesztette Robert Dale és Kenneth Ward Church. ACL. http://dblp.uni-trier.de/db/conf/acl/acl1999.html#ProszekyK99.

Recski, Gábor. 2014. „Hungarian Noun Phrase Extraction Using Rule-based and Hybrid Methods.” *Acta Cybernetica* 21 (3): 461–79.

Tordai, Anna, és Maarten De Rijke. 2006. *Four stemmers and a funeral: Stemming in hungarian at clef 2005*. Springer.

Váradi, Tamás. 2002. „The Hungarian National Corpus.” In *LREC*.