Az elérhetőségi elmélet névmási anaforafeloldásra gyakorolt hatása

Kovács Viktória

SZTE BTK Nyelvtudományi Doktori Iskola
viktoria.kovacs12@gmail.com

Kivonat: A névmási anaforafeloldás egy szövegben megtalálható összes visszautaló névmás és a hozzá tartozó legközelebbi antecedens azonosítását jelenti. A cikkben bemutatott gépi tanítási kísérletek az elérhetőségi elmélet (Ariel 2014) névmási anaforafeloldásra gyakorolt hatását mutatják meg két osztályozó kiértékelésén keresztül. Ehhez egy blogbejegyzésekből álló saját korpuszt használok fel, amelyben kizárólag a névmási visszautalások vannak kézzel annotálva. Az első kísérlethez a lehető legkevesebb előelemzési lépéssel, morfológiai és szintaktikai információkat felhasználva építettem a tanítófájlt, míg a második kísérlethez ezekhez az információkhoz hozzáadtam az elérhetőségi elméletben megfogalmazott, automatikusan is kinyerhető elveket.

1 Bevezetés

A névmási anaforafeloldás az anaforafeloldás és a koreferenciafeloldás részfeladata. Ezeknek a viszonyoknak a pontos felismeréséhez hozzájárulnak a morfológiai, a szintaktikai, a szemantikai és a pragmatikai információk is. Az anaforafeloldással kapcsolatos leggyakoribb probléma a referenciális többértelműségen alapul, azaz a kifejezés több antecedensre is visszautalhat a szövegben. Ennek ellenére a kommunikáció során megértik egymást a kommunikációs partnerek, ugyanis az antecedens felismeréséhez segítségül hívhatják a megnyilatkozás kontextusát, azaz a fizikai teret és időt, a diskurzusban korábban elhangzott információkat és a mentális enciklopédiájukat. További problémát jelentenek a zéró névmások, amelyek a szövegben ugyan nem jelennek meg, de visszautalhatnak egy korábbi objektumra. Ez utóbbi probléma megoldásához olyan, a felszíni szerkezetből is kinyerhető információkat használhatunk fel, amelyeknek a segítségével következtethetünk a zéró névmás jelenlétére. A szöveg kontextusára, illetve a szövegalkotó és a címzett mentális állapotára is következtethetünk bizonyos felszíni szerkezeti jegyek alapján, azonban ezeknek a jegyeknek a száma csekély, ezért nagy kihívást jelent az anaforafeloldás mind az elméleti, mind a számítógépes nyelvészet területén. A magyar nyelvvel kapcsolatban számos kezdeményezésről olvashatunk, amelyek egyre jobb eredményeket produkálnak a koreferencia és anaforafeloldás tekintetében (Lejtovicz–Kardkovács 2006; Miháltz 2012; Varasdi et al. 2007), azonban egyik kutatásnak sem kifejezetten a névmási anaforafeloldás áll a fókuszában.

Kovács Viktória: Az elérhetőségi elmélet névmási anaforafeloldásra gyakorolt hatása. In Váradi Tamás (sorozatszerkesztő), Ludányi Zsófia, Gráczi Tekla Etelka (szerkesztő): *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019. XIII. Alkalmazott Nyelvészeti Doktoranduszkonferencia.* Budapest: MTA Nyelvtudományi Intézet. 2019. 117–125. DOI: 10.18135/Alknyelvdok.2019.13.9

Az anaforafeloldás történhet szabály alapú rendszerek segítségével, de emellett lehetőségünk van gépi tanítási kísérleteket is végezni pozitív és negatív példák alapján, ez pedig egyszerűbbé teheti a feladat megoldását. A cikkben is egy ilyen gépi tanítási módszert mutatok be, amely Soon, Ng és Lim 2001-es munkáját veszi alapul (Soon et al. 2001), ennek a rendszernek egy, a feladathoz átalakított változata. A gépi tanítási kísérlet célja egyrészt az, hogy megvizsgáljam mennyire eredményes egy olyan osztályozó, amely a lehető legkevesebb előelemző lépéssel és manuális annotációból származó információk nélkül működik, másrészt pedig az, hogy az elérhetőségi elmélet (Ariel 2014) automatikusan is kezelhető elveinek tanításra gyakorolt hatását megvizsgáljam.

1.1 Elérhetőségi elmélet

Az elérhetőségi elmélet (Ariel 2014) a kifejezések formáját hozza összefüggésbe a referensük mentális elérhetőségével. Az elérhetőségi elmélet szerint a beszélő olyan kifejezést választ, amiről feltételezi, hogy a hallgató értelmezni tudja, tehát a kiválasztás során figyelembe veszi, hogy a hallgató mentális állapotában éppen mennyire van központi pozícióban az adott objektum, amire utalni akar. Az elméletben a különböző fokú mentális elérhetőséget mutató kifejezéseket Ariel egy skálán helyezi el, így a kifejezések a formájuk és nyelvtani tulajdonságaik alapján összehasonlíthatók az elérhetőség szempontjából. Azok a kifejezések, amelyek kevés nyelvi információt tartalmaznak, ilyen a névmás is, magas elérhetőséget mutatnak, azaz a referensük a hallgató mentális állapotának középpontjában helyezkedik el.

Az Ariel (2014) által legfontosabbnak vélt jellemzők az informativitás mértéke, a kifejezés rigidsége és a hossza. Az informativitás mértéke arra vonatkozik, hogy az adott kifejezés mennyire jellemzi hiánytalanul az adott dolgot, amire utal. A rigidség azt mutatja, hogy az adott kifejezés mennyire merev jelölő, a hossz pedig a nyelvi forma hossza, amely a kifejezés írott méretére és fonológiai méretére is vonatkozik. Minél informatívabb, rigidebb és hosszabb egy nyelvi kifejezés, annál alacsonyabb az értéke az elérhetőségi skálán. Ez azt jelenti, hogy ha a beszélő azt feltételezi, hogy a hallgató mentális állapotában nehezen elérhető a referens, akkor nagyon specifikusnak kell lennie a kifejezésnek ahhoz, hogy a hallgató értelmezni tudja az információt. Ezzel szemben, ha könnyen elérhető a címzett számára a referens, elegendő kevésbé rigid és rövidebb kifejezéssel utalni rá, mivel kevesebb erőfeszítésre van szüksége a hallgatónak ahhoz, hogy azonosítsa a kifejezés referensét. Az, hogy egy objektum a beszélgetésben a figyelmi állapot középpontjába kerüljön, és elegendő legyen névmással utalni rá, számos módon kiváltható. Elérhetőbbé teszi az objektumot az, ha jelen van a fizikai kontextusban, vagy ha korábban már szó volt róla. Szintén magas az elérhetősége azoknak az objektumoknak, amelyek szorosan illeszkednek a diskurzusuniverzumba, vagy épp feltűnően nem illeszkednek oda. További tényező még, hogy az adott objektum élőlény vagy tárgy, az elmélet szerint ugyanis az élőlények egyszerűbben elérhetőek.

Tehát amikor a visszautaló névmáshoz tartozó antecedenst keressük, olyan kifejezést keresünk, amely már önmagán hordozza az elérhetőség jeleit. A két kifejezés közötti elérhetőséget külön, relációként jellemzi Ariel (2014), melyben szerepet játszik a távolság is. Azt a megállapítást tette, hogy abban az esetben, ha az antecedens is magas elérhetőségű, messzebbről is vissza lehet rá utalni. Az elméleti keret szerint minél közelebb van az antecedens, annál elérhetőbb, annál egyszerűbb azonosítani.

2 Korpusz

A gépi tanítás egy interneten található blogbejegyzésekből, rövid cikkekből álló korpusz segítségével történt. A korpusz szövegei a következő blogokról származnak:

A Prohardver Fujitsu blogjáról (W1), a Webisztán blogról (W2), a Könyvkritikák blogról (W3), a Filmvilág blogról (W4), a Városikonyha blogról (W5), az Egyedikutya blogról (W6), a Játéknapló blogról (W7), a Nesze!szer blogról (W8), az Otthonédes blogról (W9). A korpuszban található szövegek választását az indokolta, hogy ezek a szövegek egyszerűen hozzáférhetőek és kevéssé bonyolult a szerkezetük. A választás másik oka az volt, hogy egyre több az online tartalom, ezért egy ilyen rendszer felhasználása is valószínűleg hasonló szövegen történne. A szövegek tartalmi szempontból változatos képet mutatnak. A korpuszban minden egyes névmási visszautalás teljesen manuálisan lett annotálva az MMAX2 annotációs szoftver (Müller–Strube 2006) segítségével. A visszautalásokon kívül azonban semmilyen információ nem lett kézzel a szövegekhez hozzáadva.

A korpuszban található szövegek a magyarlánc (Zsibrita et al. 2013) parse moduljával lettek előelemezve, ezeknek az információknak a segítségével generálja le egy, az adatokat megfelelő formátumra alakító algoritmus a tanítófájlt. A tanítófájlban minden visszautaló névmás és hozzá tartozó antecedensjelölt pár megtalálható, a párhoz hozzárendelt információkkal együtt.

A korpusz 60 db szöveget tartalmazott, összesen 430 névmási visszautalást, ebből 216 vonatkozó névmási, 126 személyes névmási, 88 mutató névmási visszautalás volt.

2.1 A tanításhoz használt párok

A gépi tanításhoz lehetséges visszautaló névmásokból és a hozzájuk tartozó lehetséges antecedensjelöltekből álló párokra van szükség. Ezeket a párokat a magyarlánc parse moduljának segítségével határoztam meg. A magyarlánc parse moduljának kimenete tartalmazza a szóalakot, a lemmát, az MSD-kódot, a POS taget, a morfológiai információkat, valamint a dependencia és konstituens elemzést. A konstituens elemzés segítségével a főnévi csoportok és mondatok, valamint a hozzájuk tartozó elemzések kinyerhetők az elemzett fájlból. A párok első eleme olyan főnévi csoport, amelyhez az elemző a PRON címkét rendeli. Kivételt képeznek ez alól azok a névmások, amelyeknek PronType címkéje *Art* (névelők), *Ind* (határozatlan), *Int* (kérdő), *Neg* (negatív), *Tot* (általános), mivel ezek a névmások nem referálnak egy konkrét objektumra a kontextusban. Kivételt képeztek továbbá azok a személyes névmások, amelyeknek a Person címkéje 1 vagy 2 a morfológiai elemzés során, mivel ezek a névmások a mindenkori beszélőre illetve címzettre referálnak. Tehát az ilyen típusú névmásokhoz nem tudunk antecedenst rendelni a szövegben. Az antecedensjelöltek a névmásokat a szövegben megelőző főnévi csoportok (NP) és teljes propozíciók (CP). Minden egyes lehetséges visszautaló névmás párt alkot az őt megelőző NP-kel és CP-kel a kézzel is annotált valódi antecedensével bezárólag. Tehát minden esetben annyi pár jön létre, ahány NP és CP található a névmás és az antecedense között, plusz egy, maga az antecedense.

2.2 A tanításhoz felhasznált tulajdonságok

A névmásokból és antecedensjelöltekből álló párokhoz hozzárendeltem még a két kifejezésre vonatkozó morfológiai, szintaktikai, illetve szemantikai és pragmatikai tudásból adódó információkat. A kísérlet célja az volt, hogy viszonylag egyszerűen, manuális annotáció nélkül is kinyerhető információkat használjon fel az osztályozó. Jelenleg a tanítófájlban a névmási anafora és antecedens párokat 21 tényező jellemzi, ezek mind a magyarlánc előelemzésének kimenetéből, illetve a nyers szövegből automatikusan kinyert információk. A tanítófájlba maguk a kifejezések nem kerültek bele, kizárólag az őket jellemző tulajdonságok.

A tanítófájlban a párok a következő módon vannak jellemezve:

```
[antLen, CPdist, NPdist, wDist, antCat, antPOS, anaTyp, anaCas,
antCas, casAgr, anaNum, antNum, numAgr, anaPer, antPer, perAgr,
antNom, antPrp, antPron, antDef, ant3, anaphoric]
```

A jellemzők két fő csoportra oszthatók. Az első csoportba morfológiai és szintaktikai jellemzők kerültek, amelyeket közvetlenül a magyarlánc által a szavakhoz rendelt címkékből nyertem ki. A másik nagyobb csoportba azok a jellemzők tartoznak, amelyeket szintén a magyarlánc kimenetéből, de nem az előelemzés segítségével nyertem ki. Az utolsó jellemző pedig a manuális annotációból származó adat, amely azt mutatja, hogy a pár anaforikus-e vagy sem. A következő két fejezetben ezen jellemzők részletes ismertetése található.

2.2.1 A tanításhoz felhasznált morfológiai és szintaktikai tulajdonságok

Az `antCat` jellemző a konstituens elemzésből származó címke, azt mutatja meg, hogy az antecedensjelölt CP vagy NP címkét kapott-e a konstituens elemzés során, tehát az antecedensjelölt teljes propozíció vagy főnévi csoport. Ennek megfelelően a *CP* vagy az *NP* értéket veheti fel.

Az `anaTyp` jellemző a névmás típusát jelöli. A magyarlánc morfológiai elemzőjének PronType típusú címkéit veheti fel értékként. A párokba rendezés során kizárt címkéken kívül ezért a *Dem* (mutató) a *Prs* (személyes) a *Rel* (vonatkozó) és *Rcp* (visszaható) értékeket veheti fel.

Az antecedensjelölt esete `antCas`, száma `antNum`, és személye `antPer` három különböző jellemzőként jelenik meg a tanítófájlban, és szintén a morfológiai elemzésből származik. A főnévi csoportnál a csoport fejéhez rendelt Case típusú morfológiai címkével egyezik meg az érték, a mondatoknál pedig hiányzó értéket jelez.

Az anaforikus névmás esete `anaCas`, száma `anaNum`, személye `anaPer` szintén három különböző jellemző, amelyek a morfológiai elemzésből származnak.

Az egyeztetés eset szerint `casAgr`, szám szerint `numAgr`, személy szerint `perAgr` jellemzők azt vizsgálják meg, hogy a névmáshoz rendelt eset, szám és személy címke és az antecedenshez rendelt eset, szám és személy címke megegyezik-e. Abban az esetben, ha megegyezik a jellemző, az 1-es értéket veszi fel, ha pedig nem, a 0-t.

Az `antPrp` jellemző azt jelöli, hogy az antecedensjelölt tulajdonnév-e. Abban az esetben, ha az antecedens a PROPN POS taggel rendelkezik, az érték 1, ha nem, az érték 0.

Az `antPron` jellemző azt jelöli, hogy az antecedensjelölt névmás-e. Abban az esetben, ha az antecedens a PRON címkével rendelkezik, az érték 1, ha nem, az érték 0.

Az `antPOS` jellemző a magyarlánc elemzésében az antecedensjelölt fejéhez rendelt POS taget írja ki értékként.

2.2.2 A tanításhoz felhasznált további tulajdonságok

A szintaktikai és morfológiai jellemzőkön kívül további tulajdonságokat is figyelembe vett az osztályozó, amelyeket szintén az előelemzés kimenetéből nyertem ki. Az elérhetőségi elméletben (Ariel 2014) találhatóak olyan alapelvek, amelyek a névmáshoz tartozó antecedens azonosítására vonatkoznak, és automatikusan is meghatározhatók.

Az `antNom` jellemző két értéket vehet fel, az értéke 1, ha alany esetű az antecedens, és 0, ha egyéb. Ennek a jellemzőnek az alapja az elérhetőségi elméletben megfogalmazott egyik alapelv: az elérhetőségi elv alapján az alany esetű főnévi csoportok könnyebb elérhetőséget mutatnak a mentális állapotban, ezért valószínűbb, hogy a későbbiekben is visszautalunk rájuk.

Az `antDef` jellemző szintén két értéket vehet fel attól függően, hogy az antecedensjelölt főnévi csoport a morfológiai elemzés során a Definite tulajdonság szerint milyen értéket kapott. *Def* értéket vesz fel, ha határozott, és *Ind*, ha határozatlan. Abban az esetben, ha a morfológiai elemzésben nincs a határozottságra utaló címke, hiányzó értéket jelöl a rendszer. A határozatlan névelős kifejezések gyakran diskurzusreferenst szoktak bevezetni, a frissen bevezetett diskurzusreferens által jelölt dolog pedig elérhetőbb, mint a már bevezetettek, ezért ez az információ is befolyásolhatja az osztályozó eredményességét.

Az `antLen` jellemző azt mutatja, hogy az antecedensjelölt hány szóból áll, tehát ez egy skaláris érték. Az elérhetőségi elmélet szerint a kifejezések hossza is mutatja az általuk jelölt dolog mentális elérhetőségét, a hosszabb kifejezések az általuk jelölt dolog nehezebb elérhetőségét mutatják. Ariel (2001) szerint éppen ezért az antecedens hossza és a visszautaló névmás és az antecedense közötti távolság nagysága között összefüggés van.

Az `ant3` jellemző két értéket vehet fel: ha több mint három szóból áll az antecedensjelölt, akkor 1, ha nem, akkor 0. Ez a jellemző kapcsolódik az `antLen` jellemzőhöz, Ariel (20014) elméleti keretében a három vagy annál több tartalmas szóból álló főnévi csoportok alacsony mentális elérhetőséget mutatnak.

Az antecedensjelölt és a névmás közötti távolságot három érték alapján is figyelembe veszi a rendszer.

A `CPdist` jellemző azt mutatja, hogy a névmástól számított hányadik tagmondatban helyezkedik el az antecedensjelölt, tehát ha azonos tagmondatban vannak, 0 az értéke, ha szomszédos tagmondatban akkor 1. Minden tagmondati határátlépés eggyel növeli az értéket.

Az `NPdist` jellemző azt mutatja meg, hogy a névmás és az antecedensjelölt között hány darab főnévi csoport helyezkedik el.

A `wDist` azt mutatja meg, hogy a névmás és az antecedensjelölt között hány szó található. Ebből az értékből következtethetünk az antecedens és a névmás között elhelyezkedő főnévi csoportok hosszára, tehát arra, hogy milyen elérhetőségű főneveket ítélhet nem anaforikusnak az osztályozó.

2.3 A gépi tanítási kísérletek

A meghatározott jellemzők segítségével a tanulófájlon több osztályozót is építettem a Weka szoftver (Eibe et al. 2016) segítségével, a legeredményesebb ezek közül a Random Forest (Breiman 2001) volt. Ez az algoritmus nem tipikusan ehhez a feladathoz készült, azonban a baszk nyelvvel kapcsolatban is hasonlóan ez az osztályozó ért el jobb eredményeket (Arregi et al. 2010).

2.3.1 Az osztályozó tesztelése

A két osztályozó teszteléséhez az alacsony számú visszautalás miatt a keresztvalidálás módszerét alkalmaztam. A korpuszt a szövegek alapján tíz részre osztottam, kilenc részből készült el a tanítófájl, és egy részből a tesztfájl, ezt a módszert pedig tízszer megismételtem, a végleges kiértékeléshez pedig az egyes tesztek átlagát használtam fel.

A tesztfájlokban a párok úgy jöttek létre, hogy a névmást közvetlen megelőző főnévi csoport volt az első lehetséges antecedensjelölt, amelyet a névmáshoz rendeltem. Egészen a névmást tartalmazó tagmondatot megelőző tagmondat határáig a főnévi csoportokat rendeltem a névmáshoz, ezután a következő antecedensjelölt a névmást tartalmazó tagmondatot megelőző tagmondat volt. Ez alól kivételt képeztek azok az esetek, amikor a mondat első szava névmás volt, ebben az esetben a megelőző tagmondat volt az első antecedensjelölt. Tehát a főnévi csoportok és a tagmondatok felváltva a névmástól számított távolságuk alapján lettek a névmáshoz rendelve egészen a szöveg elejéig. A párok sorbarendezésére azért volt szükség, mert amikor az osztályozó egy névmáshoz és a hozzá rendelt lehetséges antecedensjelölt párhoz az anaforikus viszonyt ítélte, továbblépett a következő névmásra, így egy névmáshoz csak egy jelöltet talált.

3 Eredmények

Ahhoz, hogy megtudjam, hogy az általam épített osztályozó mennyire eredményes, több tanítási és tesztelési kísérletet is elvégeztem. A Base Line-ként használt legegyszerűbb antecedensmeghatározáshoz egyszerűen figyelmen kívül hagytam a fent említett jellemzőket, és minden névmáshoz az első lehetséges, őt megelőző antecedensjelöltet rendeltem valódi antecedensként, ennek az esetnek az eredményei a következő táblázatban a *firstmatch* oszlopban találhatóak. Az összes többi tanítási kísérletet ehhez hasonlítottam, így mutatva meg, hogy a teljes tanítófájl és jellemzőkészlet mennyire hatékony.

A pusztán morfológiai és szintaktikai jellemzőket felhasználó osztályozó eredményeit mutatja a *withoutAccesibility(w/oA)* oszlop. Ehhez építettem a tanítófájlok segítségével olyan osztályozókat, amelyek nem vették figyelembe az `antLen, antDef, ant3, antNom antPron` *és* `wDist` jellemzőket.

Az elérhetőségi elmélet automatikusan is kezelhető elveinek a tanításra gyakorolt hatását mutatják a teljes jellemzőkészletet és a hozzájuk rendelhető összes értéket felhasználó osztályozók, ezek eredményei a *trainFull* oszlopban találhatók.

A tíz teszt átlagát tekintve tehát az elérhetőségi elmélet számítógéppel automatikusan is kezelhető elvei a pontosságon (precision) 8,09%-ot, a fedésen (recall) 1,18%-ot, az F-mértéket tekintve pedig 2,51%-ot javítottak. A tesztek eredményeit a 1. táblázat mutatja.

Az elérhetőségi elmélet alapján megfogalmazott jellemzők gépi tanításra gyakorolt hatását a Szeged korpusz koreferenciaannotált részében (Csendes et al. 2005) található névmási visszautalások tekintetében is teszteltem, mivel ez a korpusz több névmási visszautalást tartalmaz. A korpuszban nincsenek jelölve a tagmondatra történő visszautalások, ezért az osztályozás előtt kizárólag a főnévi csoportokat rendeztem párba a névmásokkal. Az eredményeket a 2. táblázat mutatja.

A tíz teszt átlagát tekintve tehát az elérhetőségi elmélet számítógéppel automatikusan is kezelhető elvei a névmáshoz tartozó antecedens azonosítását tekintve a Szeged

korpuszon a pontosságon (precision) javítottak 4,92%-ot, a fedésen (recall) rontottak 2,1%-ot, az F-mértéket tekintve pedig javítottak 1,92%-ot.

	firstmatch			*w/oA*			*trainFull*		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TEST1	4,04	26,67	7,02	31,58	13,33	18,75	38,89	15,56	22,23
TEST2	2,99	16,67	5,07	52,38	18,33	27,16	68,75	18,33	28,94
TEST3	3,42	14,58	5,54	55,00	22,92	32,36	63,13	25,00	35,82
TEST4	3,11	14,29	5,11	53,58	25,00	34,09	63,63	32,56	43,08
TEST5*	4,84	13,04	7,06	75,00	39,13	51,43	75,00	39,13	51,43
TEST6	0,99	12,50	1,83	50,00	27,50	35,48	61,11	27,50	37,93
TEST7**	0,85	9,52	1,56	70,59	28,57	40,68	68,75	26,19	37,93
TEST8	4,17	14,71	6,50	57,90	32,35	41,51	71,43	29,41	41,67
TEST9	0,57	6,45	1,05	55,56	32,26	40,82	57,90	35,48	44,00
TEST10	0,90	9,80	1,65	54,55	23,53	32,88	68,42	25,49	37,14
ÁTLAG	**2,59**	**13,82**	**4,24**	**55,61**	**26,29**	**35,51**	**63,70**	**27,47**	**38,02**

1. táblázat. A tanítási kísérletek adatai (P = precision, R = recall, F = F-measure)

	w/oA			**trainFull**		
	P	**R**	**F**	**P**	**R**	**F**
TEST1	21,93	34,25	26,74	24,77	36,99	29,67
TEST2	27,73	47,14	34,92	36,96	48,57	41,98
TEST3	28,33	44,74	34,69	35,42	44,74	39,54
TEST4	40,74	45,21	42,86	45,21	45,21	45,21
TEST5	37,17	53,85	43,98	42,55	51,28	46,51
TEST6	31,65	39,68	35,21	30,12	39,68	34,25
TEST7	40,00	59,70	47,90	58,33	41,79	48,69
TEST8	41,77	49,25	45,20	40,79	46,27	43,36
TEST9	32,80	37,84	35,14	34,74	44,60	39,06
TEST10	39,64	53,00	45,36	42,10	44,58	43,30
ÁTLAG	**34,18**	**46,47**	**39,20**	**39,10**	**44,37**	**41,16**

2. táblázat. A Szeged korpuszon végzett tanítási kísérletek adatai (P = precision, R = recall, F = F-mérték)

3.1 Kiértékelés, hibaelemzés

Az internetes blogkorpusznál tíz tesztből nyolc esetében javítottak az elérhetőségi elmélet alapján megfogalmazott jellemzők, egy esetben nem változtattak az eredményen (*), egy esetben pedig rontottak (**). Az egyes tesztek értékei nagy eltéréseket mutatnak egymástól. Ennek az az oka, hogy a tesztfájlokban egymástól eltérő mennyiségű visszautalás volt, de egyik teszt sem tartalmazott 60-nál több visszautalást, tehát már egy anaforikus kapcsolat is több százalékos eltérést jelent a kiértékelésben.

Egyedül a TEST7 esetében mondható el, hogy az elérhetőségi elmélet alapján megfogalmazott elvek rontottak a tanítás sikerességén. A tesztfájlban 42 db névmási visszautalás volt, ebből 18 személyes, 9 mutató és 15 vonatkozó névmási. A személyes

névmási visszautalások közül egyiket sem azonosította az osztályozó, és a mutatónévmási visszautalások közül is mindössze kettőt. Ennek oka egyrészt az, hogy magában a tanítófájlban is kevesebb az ilyen típusú visszautalás, ezért a tesztelés során is átlagosan kisebb a felismert visszautalások aránya. Másrészt a személyes névmási visszautalás esetében a visszautalás távolsága átlagosan nagyobb, így az osztályozó tévesen egy közelebbi főnévi csoportot jelöl meg antecedensnek. Abban az esetben, ha az osztályozó nem csak az első antecedensnek azonosított főnévi csoportig vagy tagmondatig fut, azonosítja a névmás valódi antecedensét is.

Általános problémát jelent azoknak a névmásoknak a kezelése is, amelyeknek nem szükséges a szövegben antecedenst keresni. Jelenleg az osztályozó a deiktikus névmásokhoz is sorra vizsgálja a lehetséges jelölteket. Az ebből a problémából fakadó hibákat úgy lehetne orvosolni, hogy a tanítófájlhoz olyan negatív példákat adok, amelyek deiktikus névmásokat tartalmaznak, illetve előszűröm a névmásokat a tesztfájlban.

Az osztályozó jelenleg csak a szövegben megjelenő névmásokat és a hozzájuk rendelhető antecedensjelölteket vizsgálta meg. Abban az esetben, ha a zérónévmásokhoz is lehetne antecedensjelöltet rendelni szintén egy megelőző lépés segítségével, kevesebb antecedensjelöltet kellene az osztályozónak átvizsgálnia, ez lehetséges, hogy javítana az eredményen.

A Szeged korpuszon végzett teszt eredményeinek értékelése során figyelembe kell venni, hogy a Szeged korpusz koreferenciaannotált változatában csak a főnévi csoportra történő visszautalások vannak annotálva, tehát a visszautalások száma alacsonyabb, ezáltal a pozitív tanítópéldák száma is kevesebb. Ennek ellenére összességében az F-mértéket tekintve így is javítottak az elérhetőségi elmélet alapján megfogalmazott elvek a tanítás sikerességén.

4 Konklúzió

A gépi tanítási kísérlet célja egyrészt az volt, hogy megvizsgáljam mennyire eredményes egy olyan osztályozó, amely a lehető legkevesebb előelemző lépéssel és manuális annotációból származó információk nélkül működik, másrészt pedig, hogy az elérhetőségi elmélet automatikusan is kezelhető elveinek tanításra gyakorolt hatását ellenőrizzem.

Az elérhetőségi elméletben megfogalmazott, automatikusan is kezelhető elvek növelték a tanítás hatékonyságát, leginkább a fals pozitív esetek szűrésében hatékonyak, tehát érdemes további kognitív nyelvészeti és szemantikai, pragmatikai elveket is figyelembe venni a gépi tanítás során.

A tesztek alapján az a következtetés vonható le, hogy a fals negatív példák száma okozza a legnagyobb problémát. Ezen elsősorban a tanítófájl növelésével lehet javítani. Egy másik megoldási lehetőség, hogy az osztályozás során nem csak az első antecedensnek értékelt főnévi csoportot, illetve tagmondatot vesszük figyelembe, hanem többet, és egy második lépésben ezeken a jelölteken újabb szűrést hajtunk végre.

Azonban ezzel együtt is további előelemző lépésekre van szükség, amelyek pontosan azonosítják azokat a névmásokat a szövegben, amelyekhez ténylegesen szükséges antecedenst keresni.

Források

W1 = [https://prohardver.hu/fooldal/rovat/fujitsu_blog]
W2 = [http://webisztan.blog]
W3 = [http://konyvkritikak.blog.hu]
W4 = [http://filmvilag.blog]
W5 = [https://varosikonyha.blog.hu]
W6 = [https://www.egyedikutya.hu/egyedi-kutya-blog]
W7 = [https://jateknaplo.blog.hu]
W8 = [https://neszeszer.blog.hu]
W9 = [https://otthonedes.blog.hu]

Irodalom

Ariel, M. 2001. Accessibility theory. An overview. In: Sanders, T. – Schilperoord, J. – Wilbert S. (szerk.) *Text Representation: Linguistic and Psycholinguistic Aspects.* Amsterdam: John Benjamins Publishing Company. 29–87.

Ariel, M. 2014. *Accessing noun-phrase antecedents*. London: Routledge.

Arregi, O. – Ceberio, K. – Díaz de Illarraza, A. – Goenaga, I. – Sierra, B. – Zelaia, A. 2010. A First Machine Learning Approach to Pronominal Anaphora Resolution in Basque. In: Kuri Morales, A. – Simari, G. R. (szerk.) *Advances in Artificial Intelligence.* Berlin, Heidelberg: Springer. 234–243.

Breiman, L. 2001. Random Forest. *Machine Learning* 45(1): 5–32.

Csendes, D. – Csirik, J. – Gyimóthy, T. – Kocsor, A. 2005. The Szeged Treebank. In: Matoušek, V. – Mautner, P. – Pavelka, T. (szerk.) *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)*. Karlovy Vary, Czech Republic: Springer. 123–131.

Eibe, F. – Hall, M. A. – Witten, I. H. 2016. *The WEKA Workbench. Online Appendix for „Data Mining: Practical Machine Learning Tools and Techniques"*. Fourth Edition. Morgan Kaufmann.

Lejtovicz K. E. – Kardkovács Z. T. 2006. Anaforafeloldás magyar nyelvű szövegekben. In: Alexin Z. – Csendes D. (szerk.) *IV. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2006.* Szeged: Szegedi Tudományegyetem. 362–363.

Miháltz M. 2012. Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. In: Kenesei I. – Prószéky G. – Váradi T. (szerk.) *Általános Nyelvészeti Tanulmányok* 24: 151–166.

Müller, C. – Strube, M. 2006. Multilevel annotation of linguistic data with MMAX2. In: Braun, S. – Kohn, K. – Mukherjee, J. (szerk.) *Corpus Technology and Language Pedagogy*. *New Resources, New Tools, New Methods*. (English Corpus Linguistics 3) Frankfurt: Peter Lang. 197–214.

Soon, W. M. – Ng, H. T. – Lim, D. C. Y. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* 27(4): 521–544.

Varasdi K. – Vajda P. – Miháltz M. – Naszódi M. 2007. NP-koreferenciák feloldása magyar szövegekben a Magyar WordNet ontológia segítségével. In: Tanács A. – Csendes D. (szerk.) *V. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem. 138–146

Zsibrita J. – Vincze V. – Farkas R. 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Angelova, G. – Bontcheva, K. – Mitkov, R. (szerk.) *Proceedings of RANLP 2013.* Hissar, Bulgaria: Shoumen. 763–771.