A tagmondati távolságszámítás módjainak hatása a névmási anaforafeloldásra

Kovács Viktória

Szegedi Tudományegyetem, Bölcsészettudományi Kar
viktoria.kovacs12@gmail.com

Kivonat: A névmási anaforafeloldás során a cél az egy szövegben található összes visszautaló névmás és az egyes visszautalásokhoz tartozó legközelebbi antecedensek azonosítása. A cikkben bemutatott gépi tanulási kísérletek segítségével a névmási visszautaló szó és az antecedense közötti távolság meghatározási módszereinek sikerességét vizsgálom. A vizsgálathoz a Szeged Koreferencia Korpusz (Vincze és mtsai, 2015) névmási visszautalásaiból építettem tanító és tesztfájlokat. Az osztályozók építéséhez a Szeged Korpuszban (Csendes és mtsai, 2005) található morfológiai és szintaktikai információkat használtam, a távolság meghatározásához állandó tényezőként pedig a Hobbs távolságot (Hobbs, 1978). A baseline modell ezeken kívül a két kifejezés közötti tagmondatzáró határátlépések számát vette figyelembe. A kísérletek során a tagmondatzáró határátlépések száma helyett olyan értékeket rendeltem a kifejezésekhez, amelyek kiszámításához figyelembe vettem a közbeékelődéseket (Gibson, 2000), az elérhetőségi elméletben megfogalmazott tagmondatok közötti kapcsolatokra vonatkozó elveket (Ariel, 2001), valamint azt a tényt, hogy a névmási visszautalások nagyrésze közelre történik (Grosz és mtsai, 1995).

1 Bevezetés

A koreferenciafeloldás egyik részfeladata a névmási anaforafeloldás, melynek során a cél a szöveg összes visszautaló névmásának és az egyes visszautalásokhoz tartozó legközelebbi antecedensek azonosítása. A kommunikáció során ezeknek a viszonyoknak a pontos felismeréséhez hozzájárulnak a morfológiai, a szintaktikai, a szemantikai és a pragmatikai információk is. A névmási anaforafeloldással kapcsolatos egyik leggyakoribb probléma a referenciális többértelműségen alapul, azaz a kifejezés több antecedensre is visszautalhat a szövegben: a jelen tanulmányban ismertetett vizsgálat a névmáshoz legközelebb eső antecedens azonosítását (Closest First) célozta meg. Egy másik gyakori probléma azon névmások kiszűrése, amelyek nem utalnak vissza. A magyar nyelvvel kapcsolatban (Lejtovicz és Kardkovács, 2006; Varasdi és mtsai, 2007; Miháltz, 2012) munkáit olvashatjuk, amelyek egyre jobb eredményeket érnek el a probléma megoldásának kapcsán.

Az anaforafeloldás során alapvető kérdés az antecedens keresési hatókörének meghatározása. Ezzel kapcsolatban két megközelítés lehetséges: vagy egy előre rögzített keresési tartományon belül vizsgáljuk meg a potenciális antecedensjelölteket és hasonlítjuk össze őket bizonyos grammatikai tulajdonságaik alapján; vagy

dinamikusan bővítjük a keresés hatókörét, és sorra vizsgáljuk a potenciális antecedensjelölteket addig, amíg nem találjuk meg a legvalószínűbb antecedensjelöltet. A két kifejezés közötti távolság tehát egy sarkalatos pontja mind a szabály alapú, mind a gépi tanuláson alapuló automatikus névmási anaforafeloldásnak. Ezen jellemző meghatározásához számos nyelvészeti elmélet is figyelembe vehető.

Jelen kísérlet egy korábbi munkán (Kovács, 2019) alapul, amelyben az elérhetőségi elmélet hatásának vizsgálata történt meg. A mostani kísérletben azt vizsgálom, hogy a szövegben is megjelenő névmáshoz tartozó antecedens azonosítása során a rögzített hatókör kiszámításának módja hogyan befolyásolja a gépi tanulás sikerességét. Ehhez különböző szintaktikai és kognitív nyelvészeti alapú elméletek távolságot befolyásoló tényezőit veszem figyelembe. Nem célom a jelen munkával az automatikus névmási anaforafeloldás problémáira megoldást találni, pusztán egyetlen jellemző meghatározási módjának hatását bemutatni egy gépi tanulási kísérleten keresztül. Az eredmények összefüggéseinek ismertetése, illetve nyelvészeti keretben való értelmezése további vizsgálatokat igényelnek.

1.1 Az anaforafeloldás kognitív nyelvészeti alapjai

Számos nyelvészeti modell, amely a koreferencia-, és ezen belül is az anaforafeloldást tűzte ki céljául, azon az elven alapul, hogy a beszélő vagy szövegalkotó a referáló kifejezés megválasztása során figyelembe veszi a hallgató vagy címzett mentális állapotát, hiszen az a célja, hogy a címzett felismerje azt, hogy mire utal, és ezáltal megértse a közölni kívánt információt (Ariel, 2001; Grósz és mtsai, 1995). Ez alapján a referáló kifejezés formájából és grammatikai tulajdonságaiból utólag is kikövetkeztethető a szövegben, hogy a szövegalkotó szerint mely entitások voltak az adott szituációban a címzett mentális állapotának középpontjában. A címzett mentális állapotának modellezése során azonban nem csak a kifejezések formája és grammatikai tulajdonságai lehetnek iránymutatók, hanem az adott kifejezés szövegben való elhelyezkedése is, hiszen minél nagyobb a távolság a visszautaló szó és az antecedense között, annál nagyobb erőfeszítésre van szüksége a címzettnek a kapcsolat feldolgozásához. A következő fejezetben azokat az elméleti megállapításokat fogom részletezni, amelyeket a két kifejezés közötti távolságra vonatkozóan a kísérletek során felhasználtam.

1.2 A visszautaló névmás és az antecedense közötti távolság

A két kifejezés közötti távolságból következtethetünk arra az erőfeszítésre, amelyet a címzettnek ki kell fejtenie ahhoz, hogy azonosítsa az anaforához tartozó antecedenst. Ennek megadásához általában két érték vehető figyelembe.

A Hobbs-távolság (Hobbs 1978) a két kifejezés közötti főnévi csoportok számát mutatja, azaz a lehetséges antecedensjelöltek számát, amelyekről a hallgatónak vagy címzettnek meg kell állapítania, hogy nem az anafora antecedensei. Ezt a jellemzőt minden egyes kísérletben figyelembe vettem kiindulási alapként.

A másik érték a két kifejezés közötti tagmondatok számából származik, ennek a jellemzőnek a hatását vizsgálja a három kísérlet. A tagmondatok számára hagyományos módon úgy tekintünk egyszerűen, mint egy hatókörre, ezt az elvet követi a baseline modellként szolgáló első kísérlet is. Azonban kognitív nyelvészeti kutatások alapján megállapítható, hogy nem pusztán a névmás és az antecedense közötti tagmondatok száma, de a tagmondatok egymáshoz való viszonya is meghatározza az anaforafeloldáshoz szükséges erőfeszítés mértékét.

A kognitív nyelvfeldolgozás során a feladat a szavak értelmezése és szerkezetbe való beépítése. Ezt a feldolgozást azonban nehezíti, amikor egy szerkezetbe egy újabb szerkezet közbeékelődik, és a korábbi szerkezetet a címzettnek hiányos állapotában tárolnia kell mindaddig, amíg a közbeékelt szerkezet teljessé nem válik. Minél több ilyen közbeékelődés található egy mondatban, annál nehezebb a címzettnek feldolgoznia az információt. Ebből az a következtetés vonható le, hogy az ilyen típusú mondatok értelmezése során a címzettnek nagyobb erőfeszítésre van szüksége az értelmezéshez, mint az egyszerű tagmondatok értelmezése során, és ez a különbség hatással van az antecedens azonosításához szükséges erőfeszítésre is (Gibson, 2000).

Szintén a hallgató és a címzett mentális állapotát helyezi a középpontba az elérhetőségi elmélet (Ariel, 2001), amely kitér az anafora és az antecedense közötti távolság anaforafeloldásra gyakorolt hatására is. Az elmélet szerint az alárendelő tagmondatból kisebb erőfeszítéssel érhető el az anaforához tartozó antecedens, mint a mellérendelő tagmondatból. Ezt a két jelenséget figyelembe véve végeztem el a második kísérletet.

A harmadik megállapítás, amelyet figyelembe vettem a kísérlet során, azt mondja ki, hogy a névmással való visszautalás azt feltételezi, hogy az antecedens könnyen elérhető, ezért a névmás antecedense a névmással azonos vagy szomszédos mondatban helyezkedik el (Grosz és mtsai, 1995). Azokat a névmásokat, amelyek ennél messzebbre utalnak vissza *long distance*, azaz nagy hatókörű anaforának nevezi a szakirodalom, és jellemzően olyan kifejezésre utalnak vissza, amely diskurzustopik, tehát a diskurzus fő témája, illetve annyira szaliens entitás, mint a szövegalkotó vagy a címzett maga. Ezzel az elvvel kiegészítve végeztem el a harmadik kísérletet.

2 Korpusz

A kísérletet a Szeged Korpusz (Csendes és mtsai, 2005) koreferencia-annotált alkorpuszán, a Szeged Koreferencia Korpuszon (Vincze és mtsai, 2015) végeztem el, ami iskolai fogalmazásokból és újsághírekből áll. Az összes visszautalás közül kizárólag a szövegben is megjelenő névmási visszautalásokat gyűjtöttem ki (tehát nem vettem figyelembe a zérókat), ez összesen 725 visszautalást jelent.

2.1 Módszer

A gépi tanuláshoz a Mention-Pair (Soon és mtsai, 2001) modellt használtam, amely során lehetséges visszautaló névmások és a hozzájuk tartozó lehetséges antecedensjelöltekből álló párokat nyertem ki a korpuszból. Ezek a párok és a hozzájuk rendelt morfológiai és szintaktikai jellemzők adták a tanító és tesztfájlokat.

A modell előnye, hogy a párokhoz kézzel hozzáadott jellemzők tanulásra gyakorolt hatása egymástól függetlenül vizsgálható, tehát a különböző elméletekben megfogalmazott elvek automatikus anaforafeloldásra gyakorolt hatásai ellenőrizhetők. A korpuszban nem csak a koreferens kapcsolatok találhatók meg, hanem a kifejezésekhez tartozó morfológiai és szintaktikai elemzések is. A konstituenselemzés segítségével a főnévi csoportok, valamint a hozzájuk tartozó morfológiai elemzések kinyerhetők a fájlokból. A párok első eleme olyan főnévi csoport, amely a korpuszban a morfológiai elemzés során PRON címkét kapta. Az antecedensjelöltek a névmásokat a szövegben megelőző főnévi csoportok (NP).

A tanító fájlok úgy jöttek létre, hogy minden egyes lehetséges visszautaló névmáshoz párként hozzárendeltem az őt megelőző NP-ket a kézzel is annotált valódi antecedensével bezárólag. Tehát minden esetben annyi pár jött létre, ahány NP található a névmás és az antecedense között, ezek a negatív példák, plusz egy, maga az antecedense, ez a pozitív példa. Mivel a távolság hatását szerettem volna vizsgálni, ezért minden annotált névmáshoz csak egy pozitív példa lett hozzárendelve, a szövegben hozzá legközelebbi. Azon névmások esetében, amelyekhez nem volt antecedens annotálva, tehát amelyek nem utaltak vissza, a névmást megelőző három főnévi csoporttal alkottam párokat. Ezzel azokhoz a névmásokhoz is generáltam negatív példákat, amelyek nem utaltak vissza a szövegben.

A tesztfájlokban viszont minden névmáshoz hozzárendeltem párként minden őt megelőző főnévi csoportot a szöveg első főnévi csoportjával bezárólag, hiszen elvben bármely főnévi csoport antecedense lehet bármely névmásnak. Így a tesztfájlokban a pozitív példák átlagosan az összes pár 0,39%-át tesztik ki.

A tanító fájlban való negatív példák szűrésére azért volt szükség, mert egy szövegben arányosan sokkal több a negatív példa, mint a pozitív, ez alapján pedig a legtöbb osztályozó a tesztfájban az összes párt negatívnak ítéli. A szűrés után a tanító fájlokban a pozitív példák száma az összes pár 10,35%-a lett.

2.2 A tanuláshoz használt jellemzők

A névmásokból és antecedensjelöltekből álló párokhoz hozzárendeltem a két kifejezésre vonatkozó morfológiai és szintaktikai információkat. A kísérlet célja az volt, hogy a párok tagjai közötti tagmondati távolság (*CPdist* jellemző) tanulásra gyakorolt hatását megvizsgáljam. Jelenleg a tanító fájlokban 14 tényező jellemzi a névmási anafora és antecedens párokat, ezek mind a Szeged Korpuszból származnak. A tanító fájlba maguk a kifejezések nem kerülnek bele, kizárólag az őket jellemző tulajdonságok.

A tanító fájlban a párok a következő módon vannak jellemezve (1):

(1) [CPdist, NPdist, antPOS, anaTyp, anaCas, antCas,
casAgr, anaNum, antNum, numAgr, anaPer, antPer, perAgr, anaphoric]

A jellemzők két fő csoportra oszthatók. Az első csoportba morfológiai és szintaktikai jellemzők kerültek, amelyeket közvetlenül a korpuszban a szavakhoz rendelt címkékből nyertem ki. A másik nagyobb csoportba azok a jellemzők tartoznak, amelyeket szintén a korpuszból, de nem a hozzárendelések segítségével

nyertem ki. Az utolsó jellemző pedig a koreferencia annotációból származó adat, amely azt mutatja, hogy a pár anaforikus-e vagy sem. A következő két fejezetben ezen jellemzők részletes ismertetése található, bővebben olvasható róluk Kovács 2019 munkájában (Kovács 2019).

2.2.1 A tanuláshoz felhasznált morfológiai és szintaktikai tulajdonságok

Az *antPOS* jellemző a magyarlánc elemzésében az antecedensjelölt fejéhez rendelt POS Taget írja ki értékként.

Az *anaTyp* jellemző a névmás típusát jelöli. A magyarlánc morfológiai elemzőjének PronType típusú címkéit veheti fel értékként.

Az antecedensjelölt esete *antCas*, száma *antNum*, és személye *antPer* három különböző jellemzőként jelenik meg a tanító fájlban, és szintén a morfológiai elemzésből származik. A főnévi csoportnál a csoport fejéhez rendelt Case típusú morfológiai címkével egyezik meg az érték.

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.

2.2.2 A tanuláshoz felhasznált távolságra vonatkozó tulajdonságok

A jellemzők meghatározása során azzal kísérleteztem, hogy a tagmondati távolságot a korábban említett elveket is figyelembe véve többféleképpen határoztam meg. Végül három esetet különböztettem meg. Az első esetben baseline-ként a *CPdist* jellemző egyszerűen a két kifejezés közötti tagmondatzáró határátlépéseket jelentette. Azokat az eseteket, ahol több záró határátlépés egybeesett, egy határátlépésként tekintettem. A baseline *CPdist* értéke az 2. és 3. példában szereplő visszautalás esetében így 5 lett.

2. [[Amíg vártuk **Petit**, [mert úgy hívják a kocsis haveromat], elmentünk fagyizni], [ott meg találkoztunk a barátom haverjaival.]] [Ők is ép fagyiztak.][[Velük elbeszélgettünk], [aztán jött **ő**]és [mentünk Tófaluba]].

3. [[Amíg vártuk **Petit**, [mert úgy hívják a kocsis haveromat], elmentünk fagyizni], [ott meg találkoztunk a barátom haverjaival.]] [Ők is ép fagyiztak.][[Velük elbeszélgettünk, [aztán jött **ő** is .]]

A második esetben figyelembe vettem közbeékelődéseket és az elérhetőségi elméletben az alá- és mellérendelő tagmondatok kapcsán megfogalmazott alapelveket. A szakirodalom alapján közbeékelődéseknek azokat az eseteket tekintettem, ahol a közbeékelt mondatnak sem a kezdete, sem a vége nem esik egybe az őt tartalmazó mondat kezdetével vagy végével. A 2. példában a *mert úgy hívják a kocsis haveromat* tagmondat megszakítja az *Amíg vártuk Petit (...), elmentünk fagyizni*. teljes tagmondatot. Tehát azt az egységet, hogy *Amíg vártuk Petit,* ebben a formában, hiányosan kell tárolnia a hallgatónak, mindaddig, amíg a közbeékelt mondat végéhez nem ér. Ezért a visszautaló névmástól számítva a tagmondati határátlépések számát

úgy vettem figyelembe, hogy a közbeékelődött mondat esetében egy belépési és egy kilépési értéket is számításba vettem.

Alárendelésnek tekintettem azokat az eseteket, ahol a beágyazott mondat kezdete vagy vége egybeesett az őt tartalmazó mondat kezdetével vagy végével, ezekben az esetekben a határátlépés egy pontot ért. Mellérendelésnek pedig azokat a szerkezeteket tekintettem, amelyeket más mondat tartalmazott, és ahol a megelőző mondat vége és a soron következő mondat eleje közé nem került főnévi csoport: itt a határátlépés két ponttal növelte a *CPdist* jellemző értékét. Ezeknél a szerkezeteknél is egy határátlépésnek számítottak az egybeeső mondatkezdő vagy egybeeső mondatzáró határok. A 2. példa értéke így 8 lett, a 3. példáé pedig 7.

A harmadik eset annyiban tért el a másodiktól, hogy a teljes mondat határátlépések, tehát azok a mondatok, amelyeket nem tartalmaz más mondat, nem egy, hanem három pontot értek, ezzel a nagy hatókörű anaforák esetét igyekeztem pontosítani. Ez esetben a 2. példa a 12 értéket vette fel, a 3. példa pedig 11-et.

3 A gépi tanulási kísérletek

A meghatározott jellemzők alapján a tanítófájlokon a Random Forest (Breiman, 2001) algoritmussal építettem osztályozót a Weka szoftver (Eibe és mtsai, 2016) segítségével. Mivel a cél egy adott jellemző meghatározási módszerének tanulásra gyakorolt hatásának vizsgálata volt, ezért az osztályozó építése során a Wekának az algoritmussal kapcsolatos alapértelmezett beállításain nem változtattam. A névmási anaforafeloldással kapcsolatban az osztályozóról olvashatók további eredmények a baszk (Arregi és mtsai, 2010) a maláj (Xian és mtsai, 2016) és az orosz (Ionov és Kutuzov, 2014) nyelveken végzett kísérletekről is.

3.1 Az osztályozó tesztelése

A három osztályozó teszteléséhez az alacsony számú visszautalás miatt a keresztvalidálás módszerét alkalmaztam. A korpuszt a szövegek alapján tíz részre osztottam: kilenc részből készült el a tanító fájl, egy részből pedig a tesztfájl. Ezt a módszert pedig tízszer megismételtem, a végleges kiértékeléshez pedig az egyes tesztek átlagát használtam fel. A tíz tesztfájlban található névmási visszautalások arányait a 1. táblázat mutatja.

A tesztfájlokon a kiértékelés során a fals pozitív példák szűréséhez a closest-first módszert (Soon és mtsai. 2001) alkalmaztam, mivel a távolság hatását vizsgáltam. Tehát a névmáshoz a szövegben legközelebb álló, az osztályozó által pozitívnak ítélt antecedensjelöltet tekintettem egyedül a névmáshoz rendelt antecedensnek. Ezzel egyre csökkentettem minden névmás tekintetében a pozitív példák számát, a legközelebbire.

1. táblázat: A tesztfájlok adatai (Dem= mutató névmás, Prs= személyes névmás, Rel= vonatkozó névmás, Other= egyéb névmási visszautalás)

	Dem	**Prs**	**Rel**	**Other**	**Összesen**
TEST1	7	22	44	1	74
TEST2	20	12	38	0	70
TEST3	12	22	42	0	76
TEST4	11	24	37	1	73
TEST5	12	25	41	0	78
TEST6	5	22	36	0	63
TEST7	10	17	39	1	67
TEST8	10	25	32	0	67
TEST9	12	22	40	0	74
TEST10	13	18	50	2	83
Összesen	112	209	399	5	725

2. táblázat: A tanulási kísérletek adatai (P = precision, R = recall, F = F-measure)

	Baseline			**Exp1**			**Exp2**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TEST1	22,41	35,14	27,37	22,31	36,49	27,69	23,53	37,84	29,02
TEST2	28,07	45,71	34,78	29,66	50,00	37,23	32,14	51,43	39,56
TEST3	29,20	43,42	34,92	28,57	42,11	34,04	30,63	44,74	36,36
TEST4	37,50	45,21	40,99	34,83	42,47	38,27	38,46	47,95	42,68
TEST5	40,19	55,13	46,49	39,62	53,85	45,65	41,18	53,85	46,67
TEST6	31,65	39,68	35,21	35,62	41,27	38,24	35,82	38,10	36,92
TEST7	36,61	61,19	45,81	41,84	61,19	49,70	39,60	59,70	47,62
TEST8	39,02	47,76	42,95	38,55	47,76	42,67	40,74	49,25	44,59
TEST9	30,85	39,19	34,52	34,04	43,24	38,10	34,02	44,59	38,6
TEST10	41,75	51,81	46,24	37,72	51,81	43,65	51,81	51,81	51,81
ÁTLAG	**33,73**	**46,42**	**38,93**	**34,28**	**47,02**	**39,52**	**36,79**	**47,92**	**41,38**

4 Eredmények

Ahhoz, hogy megtudjam, hogy az általam alkalmazott tagmondati távolságszámítás eredményes-e a gépi tanulás során, három tesztet végeztem el. A Baseline tesztelése

során nem tettem különbséget a tagmondatok között, ezek az eredmények láthatók a 2. táblázat Baseline oszlopában. Az első tesztelésnél már figyelembe vettem a közbeékelődéseket és az alá- és mellérendelő mondatok közötti különbségeket, ezt mutatja a táblázatban az Exp1 oszlop. A második teszt során már a nagy hatókörű anaforák alapján megfogalmazott elveket is figyelembe vettem, ezt mutatja az Exp2 oszlop.

Mivel az Exp2 eredményei konzisztensen jobbak lettek a Baseline eredményeinél, azt is megvizsgáltam, hogy az egyes visszautalási típusok tekintetében mekkora a változás. Mivel a visszaható névmási visszautalások száma a korpuszban kevés volt, ezért a mutató névmási (Dem), személyes névmási (Prs) illetve vonatkozó névmási (Rel) kategóriákat vizsgáltam meg. Ezeknek az eredményeit a 3. és a 4. táblázat mutatja.

3. táblázat: A tanulási kísérletek adatai a visszautaló névmások típusai szerint a Baseline esetében (P = precision, R = recall, F = F-measure)

	Dem			**Prs**			**Rel**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TEST1	01,96	14,29	03,45	50,00	04,54	08,33	39,34	54,55	45,71
TEST2	06,82	15,00	09,38	0	0	0	44,62	76,32	56,31
TEST3	04,08	16,67	06,56	33,33	13,64	19,35	51,85	66,67	58,33
TEST4	0	0	0	40,00	25,00	30,77	63,41	70,27	66,67
TEST5	11,90	41,67	18,52	20,00	04,00	06,66	61,67	90,24	73,27
TEST6	0	0	0	20,00	04,54	07,40	60,00	66,67	63,16
TEST7	06,67	30,00	10,91	66,67	23,53	34,78	57,63	87,18	69,39
TEST8	11,54	30,00	16,67	40,00	16,00	22,86	55,56	78,13	64,94
TEST9	03,70	08,33	05,13	33,33	04,54	08,00	42,86	67,50	52,43
TEST10	04,17	07,69	05,41	33,33	05,55	09,52	54,79	80,00	65,04
ÁTLAG	**5,08**	**16,36**	**7,60**	**33,67**	**10,14**	**14,77**	**53,17**	**73,75**	**61,52**

5 Kiértékelés, hibaelemzés

Az összes visszautalást tekintve az Exp1 kísérletben a tíz teszt átlaga alapján javított az eredményeken az új *CPdist* számolási módszer a Baseline-hoz képest, a pontosságon 0,55%-kal, a fedésen 0,59%-kal, az F mértékek átlagait tekintve pedig 0,6%-kal. Ennek ellenére nem vonható le egyértelműen az a konklúzió, hogy az új jellemző eredményesebb, mivel 5 teszt javított, 5 pedig rontott a Baseline *CPdist* számolási módszeréhez képest. Az Exp2 kísérletben minden egyes tesztről elmondható, hogy jobb eredménye lett, mint a Baseline tesztjeinek. Az összes teszt átlaga alapján a fals pozitív esetek szűrésén javított az új jellemző a legtöbbet, tehát a pontosságon 3,07%-kal. A fedésen 1,5%-kal, az F mértéken pedig 2,45%-kal javított az új jellemző, tehát egyértelműen ez a legeredményesebb a három számítási módszer közül.

4. táblázat: A tanulási kísérletek adatai a visszautaló névmások típusai szerint az Exp2 esetében (P = precision, R = recall, F = F-measure)

	Dem			**Prs**			**Rel**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TEST1	03,85	28,57	06,78	20,00	04,54	07,41	41,67	56,82	48,08
TEST2	13,95	30,00	19,05	25,00	08,33	12,50	46,77	76,32	58,00
TEST3	04,08	16,67	06,56	42,86	13,64	20,69	53,70	69,05	60,42
TEST4	03,03	09,09	04,55	53,85	29,17	37,84	63,41	70,27	66,67
TEST5	12,20	41,67	18,87	25,00	04,00	06,89	63,16	87,80	73,47
TEST6	0	0	0	33,33	04,54	08,00	63,89	63,89	63,89
TEST7	07,89	30,00	12,50	66,67	23,53	34,78	61,11	84,62	70,97
TEST8	13,04	30,00	18,18	45,45	20,00	27,78	54,35	78,13	64,10
TEST9	04,00	08,33	05,41	33,33	09,09	14,29	46,15	75,00	57,14
TEST10	02,94	07,69	04,26	50,00	05,56	10,00	56,34	80,00	66,12
ÁTLAG	**6,50**	**20,20**	**9,61**	**39,55**	**12,24**	**18,02**	**55,06**	**74,19**	**62,88**

Az egyes visszautalási típusok szerint is elmondható, hogy az Exp2 *CPdist* jellemzője átlagosan javított az eredményeken. A legnagyobb javulás a személyes névmási visszautalásnál látható, ahol a fedésen 2,11%-kal, a pontosságon 5,88%-kal, az F mértéken pedig 3,25%-kal javított a jellemző. A mutató névmási visszautalás esetében a fedésen 3,84%-ot, a pontosságon 1,41%-ot az F mértéken 2,01%-ot javított. A legkisebb változás a vonatkozó névmási visszautalásnál történt, ebben az esetben a fedés 0,44%-kal, a pontosság 1,88%-kal, az F mérték pedig 1,36%-kal nőtt. Ennek oka, hogy a vonatkozó névmási visszautalás antecedense mindig a megelőző tagmondatban található, ezért az új jellemző nem változtatott számottevően az értékeken. Azonban azokban az esetekben, ahol a visszautalás messzebbre is történhet, a távolság számítás módszerének finomítása javított az eredményeken, a személyes névmási visszautalás esetében a fals pozitív, a mutató névmási visszautalás esetében pedig a fals negatív előfordulások szűrésében.

Általánosságban elmondható, hogy a tesztek között nagy eltérések vannak, ennek oka, hogy az egyes tesztekben eltér a visszautalások száma. Emellett az is szembetűnő, hogy a vonatkozó névmási visszautalás felismerése sokkal eredményesebb, mint a személyes- vagy mutató névmási visszautalásé. Ennek egyik oka, hogy a névmás és antecedense közötti nagyobb távolság esetén az osztályozó gyakran tévesen egy közelebbi főnévi csoportot jelöl meg antecedensnek. A másik ok pedig maguknak a visszautalásoknak az aránya a szövegekben. A vonatkozó névmási visszautalások sokkal gyakoribbak a szövegekben, ezért mind a tanító, mind a tesztfájlokban nagyobb arányban fordulnak elő.

Általános problémát jelent azoknak a névmásoknak a kezelése, amelyeknek nem szükséges a szövegben antecedenst keresni. Mivel jelen tanulmány célja kizárólag a

távolságszámítás hatásának vizsgálata volt, ezt a problémát nem kezeltem. Jelenleg az osztályozó az összes PRON címkével jellemzett kifejezést kigyűjti, és sorra vizsgálja hozzájuk a lehetséges antecedensjelölteket. Az ebből a problémából fakadó hibákat úgy lehetne orvosolni, hogy a tanító fájlhoz olyan negatív példákat adok, amelyek deiktikus névmásokat tartalmaznak, azonban ezzel tovább növelném a negatív és pozitív példák közti arányok különbségét. Ahhoz hogy a későbbiekben az egyes elméletekből fakadó jellemzők tanításra gyakorolt hatásáról pontosabb képet kapjak a tesztfájlokban szereplő névmások, illetve a negatív példák előszűrése lesz szükséges.

6 Konklúzió

A gépi tanulási kísérlet célja az volt, hogy megvizsgáljam a visszautaló névmás és az antecedense közötti távolság számítási módszereinek hatását egy osztályozó eredményességére. Az Exp2 kísérlet eredményei alapján elmondható, hogy a két kifejezés közötti mondatszintű távolság számítása során javít az osztályozó eredményességén az, ha a különböző mondattípusokat más-más súllyal vesszük számításba, leginkább azokban az esetekben, ahol az antecedens nem azonos vagy szomszédos tagmondatban található.

Hivatkozások

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

Arregi, O., Ceberio, K., Díaz de Illarraza, A., Goenaga, I., Sierra, B., Zelaia, A.: A First Machine Learning Approach to Pronominal Anaphora Resolution in Basque. In Kuri-Morales, A., Simari, G. R. (szerk.) Advances in Artificial Intelligence – IBERAMIA 2010. Vol. 6433 Springer Berlin Heidelberg. 234–243. (2010.)

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

Csendes D., Csirik J., Gyimóthy T., Kocsor A.: The Szeged Treebank. In: Matoušek, V., Mautner, P., Pavelka, T. (szerk) Text, Speech and Dialogue. TSD 2005. Lecture Notes in Computer Science, vol 3658. Springer, Berlin, Heidelberg (2005)

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

Gibson, E.: The dependency locality theory: A distance-based theory of linguistic complexity. In Miyashita, Y., Marantz, A., O'Neil, W. (szerk.) Image, language, brain, Cambridge, MA: MIT Press, 95–126. (2000)

Grosz, B. J., Joshi, A. K., Weinstein, S.: Centering: A Framework for Modeling the Local Coherence of Discourse Computational Linguistics, Volume 21, Number 2, 203–225. (1995)

Hobbs, J. R.: Resolving pronoun references, Lingua, Volume 44, Issue 4, 311-338. (1978)

Ionov, M., Kutuzov, A. The impact of morphology processing quality on automated anaphora resolution for Russian. In Selegey, V.P., Baytin, A.V., Belikov, V.I., Boguslavsky, I.M., Dobrov, B.V., Dobrovolsky, D.O., Zakharov, L.M., Iomdin, L.L., Kobozeva, I.M., Kozerenko, E.B., Krongauz, M.A., Laufer, N.I., Lukashevich, N.V., McCarthy, D., Nivre,

J., Osipov, G.S., Raskin, V., Hovy, E., Sharoff, S.A. (szerk.) Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" 232-240. (2014)

Kovács, V.: Az elérhetőségi elmélet névmási anaforafeloldásra gyakorolt hatása. In Váradi, T. (sorozatszerkesztő), Ludányi, Zs., Gráczi, T. E. (szerk.) Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019. XIII. Alkalmazott Nyelvészeti Doktoranduszkonferencia. Budapest: MTA Nyelvtudományi Intézet. 114–123. (2019)

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

Miháltz, M.: Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. Általános Nyelvészeti Tanulmányok, 24, 151–166. (2012)

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

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

Vincze, V., Hegedűs, K., Farkas, R.: SzegedKoref: kézzel annotált magyar nyelvű koreferenciakorpusz. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged, 312–319 (2015)

Xian, B. C. M., Saloot, M. A., Ghazali, A. S. M., Bouzekri, K., Mahmud, R. Lukose, D. Benchmarking Mi-AR: Malay anaphora resolution. In *2016 International Conference on Optoelectronics and Image Processing (ICOIP)*, IEEE, Warsaw 59-69 (2016)