Szegedi Tudományegyetem
Nyelvtudományi Doktori Iskola
Elméleti nyelvészet program

Névmási anaforafeloldási kísérletek a magyar nyelvben

TÉZISEK

Kovács Viktória

Témavezető: Dr. Szécsényi Tibor

Szeged

2021

1. Bevezetés: Az értekezés témája, célja és felépítése

A disszertáció célja gépi tanulási kísérleteken keresztül megvizsgálni a jelenleg bevett, automatikus anaforafeloldást célzó statisztikai alapú felügyelt gépi tanulási kísérleti módszerek eredményeit a névmási anaforafeloldás tekintetében a magyar nyelvben, ezen belül is nagy hangsúlyt fektetve a tanulás alapjául szolgáló jellemzőkészlet összeállításának lehetőségeire. Az anaforafeloldással kapcsolatban az elsődleges kérdés, hogy valójában mit is szeretnénk automatikusan kinyerni a szövegből, és itt nem feltétlenül az anafora definíciójára kell gondolnunk.

A számítógépes nyelvészet az anaforafeloldást hagyományosan az automatikus tartalomkinyerés vagy a gépi fordítás problémájának szemszögéből közelíti meg, ezért a kinyerés során kevéssé veszi figyelembe a szövegalkotót és annak mentális állapotát. A hangsúly a helyes visszautalások kinyerésén van, azokon, amelyek megfelelnek a grammatikai és szintaktikai szabályoknak, éppen ezért gyakran ezek a szabályok az alapjai az automatikus feloldást célzó rendszereknek. A visszautalások azonban nem mindig ilyen egyszerűen azonosíthatók. Számos esetet képzelhetünk el, amelyekben a visszautalás annak ellenére sikeresnek mondható, tehát a befogadó helyesen azonosítja a visszautaló szó antecedensét, hogy a visszautalás a nyelvtani szabályoknak megfelelt volna. Abban az esetben, ha a nyelvtani szabályokra és a helyes visszautalásokra megállapított szintaktikai vagy szemantikai szabályokra nem támaszkodhatunk a szövegtípus vagy maga a célnyelv felszíni szerkezete miatt, a funkcionális és kognitív nyelvészet mentális állapotra és figyelmi állapotra tett megállapításai mérvadók lehetnek. Éppen ezért érdemes megvizsgálni, hogy az egyes anaforafeloldással kapcsolatos mentális állapotra vonatkozó megállapítások hogyan és mekkora pontossággal implementálhatók számítógépes környezetbe, és hogy milyen hatással vannak az automatikus anaforafeloldás eredményességére.

A dolgozat elején meghatározom a kutatás célkitűzéseit, és az előzetes hipotéziseimet a gépi tanulási kísérletekkel kapcsolatban. Ezután egy szakirodalmi áttekintésen keresztül megvizsgálom a koreferencia és az anafora meghatározási lehetőségeit a nyelvészeti és a számítógépes nyelvészeti szakirodalomban, ennek során kitérek az automatikus feloldási lehetőségekre is. A dolgozat második felében először ismertetem az általam felhasznált korpuszok felépítését és a kísérletek végrehajtásához szükséges megelőző lépéseket, majd részletes bemutatom a tanulási kísérletek során felhasznált jellemzőket. Végül az általam elvégzett gépi tanulási kísérletek részletezése és az eredményekből származó konklúzió levonása található.

2. Az értekezés hipotézisei és kutatási kérdései

A nullhipotézisem az, hogy lehetséges az automatikus névmási anaforafeloldás a magyar nyelvben szemantikai információk nélkül is, pusztán morfológiai, szintaktikai és egyéb, a felszíni szerkezetből kinyerhető, kognitív nyelvészeti alapú jellemzők segítségével. Ennek bizonyításához több kísérletet végeztem el, amelyekben nem vettem figyelembe szemantikai információkat. A kísérletekhez először meg kell vizsgálni a nyelvészeti és számítógépes nyelvészeti szakirodalom jelenlegi álláspontját a névmási anafora definíciójáról és a névmáshoz tartozó antecedens azonosítási lehetőségeiről, valamint annak nehézségeiről. Ezután a két, névmási visszautalások tekintetében manuálisan annotált magyar nyelvű korpuszt, a Szeged Korpusz (Csendes és mtsai., 2004, 2005; Vincze és mtsai., 2010) koreferenciaannotált alkorpuszát, a SzegedKoref korpuszt (Vincze és mtsai., 2018) és a KorKorpuszt (Vadász, 2020) kell megvizsgálni, hogy azonos típusú információk legyenek kinyerhetők belőlük, majd meg kell határozni a gépi tanulási kísérletek során milyen és mennyi pozitív és negatív példát, illetve milyen jellemzőket veszek figyelembe.

A tanító és tesztelő fájlok felépítésének szempontjából a Mention-pair technikát (Aone & Benett, 1995) alkalmazom minden esetben. A tesztelés során a tesztfájlokban megtalálható az adott korpuszrészletben előforduló összes névmás, és hozzá párként hozzárendelve az összes névmást megelőző főnévi csoport, mint lehetséges antecedensjelölt.

A gépi tanulás célja, hogy az épített modell felismerjen legalább egy antecedenst, amellyel a visszautaló névmás anaforikus kapcsolatban áll. Az anaforafeloldás tekintetében az egyik probléma a negatív és pozitív tanítópéldák kiegyensúlyozatlan eloszlása, ennek a problémának a kiküszöbölésére több módszer is létezik. A kísérletek alapjául szolgáló módszerből kifolyólag a tanulás során a negatív példák olyan szópárok lesznek, amelyek nem állnak anaforikus kapcsolatban, pozitív példák pedig olyan szópárok lesznek, amelyek anaforikus kapcsolatban állnak (kézzel annotált esetek a korpuszokban). Ennek következtében egy szövegből lényegesen több negatív, mint pozitív példa állítható elő, ami befolyásolja annak a valószínűségét, hogy az osztályozó mennyire sikeresen ismeri fel a két csoport (anaforikus, nem anaforikus) tagjait. Az első hipotézisem szerint azok a modellek érik el a legjobb eredményeket a tesztelés során, amelyekben a pozitív és negatív példák eloszlása azonos a tesztfájlokban várható pozitív és negatív esetek eloszlásával, tehát sem a pozitív, sem a negatív példák számát nem csökkentjük a tanítófájlokban manuálisan. Fontos kiemelni ebben az esetben, hogy az általam összehasonlított, a tanítófájlokban megtalálható pozitív és negatív párok megoszlására vonatkozó módszerek pusztán elméleti jellegű kísérletek, hiszen

egy valós, számítógépes nyelvészeti alkalmazás esetében nem határozható meg előre milyen lesz az adott szöveg, amelyen a feladatot el kell végezni, lehet akár egy egész regény vagy épp csak egy mondat, így a bennük található pozitív és negatív párok arányára sem lehet előjelzést tenni.

Mivel a kísérletek célja a névmáshoz tartozó egyetlen antecedens kiválasztása, a modell azonban névmás-antecedensjelölt párokat osztályoz, a második hipotézis arra vonatkozik, hogyan választhatunk a pozitívnak ítélt párok segítségével egyetlen antecedensjelöltet. Két módszert hasonlítok össze a tanulási kísérletek során, ezek a Best-first (Ng & Cardie, 2002) és a Closest-first (Soon és mtsai., 2001) módszerek. A Best-first módszer az osztályozó által legmagasabb valószínűségi értékkel ellátott névmás-antecedens párt jelöli meg a névmás antecedensének, a Closest-first módszer a szövegben a névmáshoz legközelebb eső, az osztályozó által antecedensnek ítélt főnévi csoportot jelöli a névmás antecedensének. A második hipotézisem szerint a legnagyobb valószínűségi értékkel ellátott névmás-antecedens pár kiválasztása nagyobb hatékonyságot eredményez, hiszen a névmások gyakran utalnak a szövegben messzebbre, így pusztán a lehetséges antecedensjelölt közelségének figyelembevétele fals pozitív eredményt okozhat.

A kísérletek harmadik szempontja a kifejezéspárokhoz rendelhető jellemzők vizsgálata. A disszertáció célja megvizsgálni, hogy kizárólag morfológiai és szintaktikai jellemzők, valamint egyéb felszíni szerkezetből kinyerhető, kognitív alapú jellemzők (Ariel, 2014; Gibson, 2000; Grosz és mtsai., 1995; Hobbs, 1978) segítségével is lehetséges automatikus névmási anaforafeloldást végezni a magyar nyelvben. A kísérletek során a harmadik kutatási kérdésem, hogy ezek közül a jellemzők közül melyek a leghatékonyabbak a modellépítés szempontjából. Először megvizsgálom, hogy a két kifejezés közötti távolság kiszámítására milyen lehetőségek merülnek fel, és ezek közül melyik a legeredményesebb, másrészt a korpuszból kinyerhető morfológiai és szintaktikai jellemzőket négy kognitív alapon megfogalmazott jellemzőcsomaggal egészítem ki egyesével, hogy megvizsgáljam, az általam megfogalmazott jellemzők hogyan módosítják a modellépítés sikerességét. A harmadik hipotézisem, hogy a tanulási kísérlethez hozzáadott nem nyelvi jellemzők javítanak a modellépítés sikerességén.

A gépi tanulási kísérleteket külön végzem el az egyes névmási visszautalási típusok tekintetében (személyes névmás, mutató névmás, vonatkozó névmás), feltételezve, hogy egymástól eltérően viselkedhetnek a fent említett hipotézisek szempontjából. Az egyes névmástípusokkal elvégzett kísérletek végén megvizsgálom, hogy melyik a legsikeresebb

módszer a modellépítés szempontjából, mind a pozitív és negatív példák aránya, mind az alkalmazott jellemzők szempontját figyelembe véve.

3. A kutatás korpusza és módszerei

A kísérlet során két korpuszt használok fel: a SzegedKoref Korpuszt (Vincze és mtsai., 2018), amely a Szeged Korpusz (Csendes és mtsai., 2005) koreferenciaannotált alkorpusza, valamint összehasonlításként a KorKorpuszt (Vadász, 2020).

A SzegedKoref Korpusz a Szeged Korpusz egy alkorpusza, ezért részletes morfológiai és szintaktikai annotációt is tartalmaz. Az eredeti Szeged Korpuszból a 8. és 10. osztályosok fogalmazásait, valamint hvg-s cikkeket tartalmazza. Mivel ezek a szövegek megtalálhatók a Szeged Korpusz egy másik alkorpuszában, a Szeged Dependencia Korpuszban (Vincze és mtsai., 2010) is, így a függőségi elemzések is a rendelkezésemre álltak a szövegekhez. A szavak eredeti alakja mellett megtalálható a korpuszokban a szófaj, a morfológiai elemzés, valamint a függőségi elemzés kimenete: élek, él címke, szófaji címke, továbbá a konstituens elemzés kimenete. A korpuszban minden szóhoz tartozik egy azonosítószám, ami azt mutatja, hogy az adott szó a mondat hányadik szava, ezt az értéket használja fel a függőségi elemzés is az élek megállapításához. A mondatokat üres sor választja el egymástól. Mivel a korpuszban az összes elemzett fogalmazás és cikk egy fájlban található, van egy plusz azonosító is, ami azt mutatja, hogy az adott sor melyik szöveghez tartozik.

A SzegedKoref koreferencia korpuszban az eredeti Szeged Korpuszban található információkon túl egy további oszlop is található, amely azt mutatja meg, hogy az adott szó, illetve az a frázis, amelynek a szó része, a szöveg melyik koreferencialáncába tartozik. A koreferens szavak, frázisok a korpuszban ugyanazt az azonosítót kapják, vagyis nem az anafora–antecedens párok vannak jelölve, hanem ekvivalencia osztályok.

A KorKorpuszban, hasonlóan a SzegedKoref koreferencia korpuszhoz, minden szónak van egy azonosítója, ami a mondatban elfoglalt pozícióját mutatja. Szintén megtalálható a szó eredeti formája, a lemma, a szófaji címke, morfológiai elemzés és függőségi elemzés kimenete. A koreferenciaannotáció két oszlopban található, az első oszlop azt mutatja, hogy az adott szónak hol található az antecedense, ez két szám érték: hányadik mondat, hányadik szó, kettősponttal elválasztva.

A gépi tanulási kísérletekhez a Mention-pair (Soon és mtsai., 2001) modellt használtam, amelyhez a lehetséges visszautaló névmások és a hozzájuk tartozó lehetséges antecedensjelöltekből álló párokat kellett kinyerni a korpuszból. Mivel a két korpuszban nem csak a koreferens kapcsolatok találhatók meg, hanem a kifejezésekhez tartozó morfológiai és

szintaktikai elemzések is, így ezek a párok és a hozzájuk rendelt morfológiai és szintaktikai jellemzők adták a tanító és tesztfájlokat. A modell előnye az, hogy segítségével azoknak a jellemzőknek a tanulásra gyakorolt hatása egymástól függetlenül vizsgálható, amelyek a korábban ismertetett különböző elméletekben megfogalmazott elvek szerint hatással vannak az anaforafeloldásra. Ezek a jellemzők expliciten nem szerepelnek a korpuszokban, viszont automatikus eszközökkel meghatározhatók. A konstituens elemzés segítségével a főnévi csoportok, valamint a frázisok fejéhez tartozó morfológiai elemzések kinyerhetők a fájlokból. A párok első eleme olyan főnévi csoport, amely a korpuszban a morfológiai elemzés oszlopban PronType attribútummal rendelkezik, ehhez pedig a prs, dem vagy rel címkét kapta. Az antecedensjelöltek pedig a névmásokat a szövegben megelőző főnévi csoportok (NP).

A gépi tanulás során a lehetséges anafora-antecedens párok közül kell kiválasztani a korpuszban ténylegesen koreferensként jelölt párokat. Ez a kiválasztás vonatkozhat az összes koreferens pár azonosítására, vagy az anaforához legalább egy kézzel is annotált antecedens azonosítására, amely lehet a legközelebbi vagy az osztályozó által a legnagyobb valószínűségi értékkel ellátott antecedens. Kísérleteim során a névmáshoz egyetlen kézzel is annotált antecedensjelölt azonosítását tűztem ki célul, ennek során pedig megvizsgáltam az osztályozó által a legközelebbi és a legnagyobb valószínűségi értékkel ellátott antecedensként értékelt főnévi csoportot is. A gépi tanulási kísérleteket a Weka szoftver (Eibe és mtsai., 2016) segítségével végeztem el, és a Random forest algoritmust (Breiman, 2001) alkalmaztam. Az általam végzett kísérletekben párokról hoz döntéseket az osztályozó, mégpedig azt, hogy anaforikus kapcsolatban állnak egymással vagy sem, így a MUC feladatban alkalmazott, a kifejezések közötti kapcsolatokra összpontosító kiértékelésből indulhatunk ki. A MUC feladat a kiértékelés során (Vilain és mtsai., 1995) a standard IR metrikákat alkalmazza (F=F-mérték, P=Pontosság, R=Fedés), ezek kiszámításához pedig a gold standard korpuszban jelölt, valamint a modell által azonosított koreferens kapcsolatok számát használja.

4. Az értekezés eredményei

Az elvégzett gépi tanulási kísérletek egy nullhipotézisen és három további hipotézisen alapultak.

A nullhipotézisem az volt, hogy gépi tanulás segítségével lehetséges szemantikai információk nélkül is anaforikus párokat azonosítani a szövegekben. Ez a hipotézisem helytállónak bizonyult, hiszen a kísérletek során minden esetben azonosított az osztályozó helyesen

anaforikus párokat. A gépi tanulási kísérletek során a tanító és tesztfájlok létrehozásához a korpuszokban megtalálható szintaktikai elemzést, a tanulás alapját adó jellemzőkhöz pedig szintén kizárólag a korpuszokban megtalálható morfológiai és szintaktikai nyelvi jellemzőket, valamint általam megfogalmazott, kognitív alapú jellemzőket vettem figyelembe. A teljes automatikus névmási anaforafeloldás következtetést és a szövegkörnyezet, a kontextus ismeretét igényli. Ennek fényében az általam kidolgozott módszer a lehetséges antecedensjelöltek előszűrésére tűnik alkalmasnak.

Az első hipotézisem a tanítófájlokban megtalálható pozitív és negatív példák egymáshoz viszonyított arányára vonatkozik. Az általam megfogalmazott jellemzők mellett a tanítóadathalmazban megjelenő pozitív és negatív példák arányaival is kísérleteket végeztem. Négy módszer segítségével építettem fel a tanítóadathalmazokat a gépi tanulási kísérlethez: Az elsőben a párok generálásához figyelembe vettem minden a szövegben megtalálható névmást és hozzárendeltem párként a névmást megelőző összes főnévi csoportot, tehát sem a pozitív tanítópéldák, sem a negatív tanítópéldák számát nem csökkentettem (Exp1). A második esetben kizárólag a visszautaló névmásokból generáltam tanítópéldákat, ezekhez pedig párként hozzárendeltem az őket megelőző főnévi csoportokat az első kézzel is annotált főnévi csoporttal bezárólag (Exp2). Ezzel a módszerrel mind a negatív, mind a pozitív tanítópéldák számát csökkentettem. A harmadik esetben a kisebbségi csoport, azaz a pozitív példák számát növeltem úgy, hogy a második esetet kiegészítve pozitív párként a névmástól számított távolabbi kézzel is annotált főnévi csoportokat is a tanítóadatbázishoz rendeltem (Exp3). A negyedik esetben a pozitív és negatív példák arányait kiegyenlítettem egymással, úgy, hogy a negatív példák közül annyit adtam a tanítóadatbázishoz véletlenszerűen, amennyi pozitív példát tartalmazott (Exp4). Az alap elképzelés az volt, hogy az osztályozó akkor fog a legjobban teljesíteni, ha a példák mennyisége szempontjából hasonló összetételű fájlon fog tanulni, mint amin a modell végül tesztelésre kerül. A kísérlet eredményeként elmondható, hogy két esetben ez az elképzelés teljesült. Az összes személyes névmás és a mutatónévmás esetében ez a módszer mutatkozott a legeredményesebbnek. Itt azonban ki kell térnem arra, hogy a mutató névmás esetében a névmásoknak mindössze 18,5%-a volt visszautaló, tehát a negatív párok magasabb aránya a tanítófájlban ebből kifolyólag a tesztelés során is javított az eredményen. Az egyes szám harmadik személyű és vonatkozónévmási visszautalás esetében már nem ez mutatkozik a legsikeresebb módszernek. Ez részben az algoritmus működéséből is adódik, hiszen a Random Forest algoritmus a tanítópéldákból is véletlenszerűen választ, így annak az osztálynak a tagjai, amelyből több van a tanítófájlban, valószínűbben vesznek részt a

modellépítésben és ezáltal hatékonyabban ismeri fel az osztályozó a tesztelés során őket. Ennek következtében a tanítófájlok létrehozása során törekedni kell arra, hogy a pozitív példák minél nagyobb arányban legyenek jelen, közel azonosban, mint a negatív példák, hogy egyenlő eséllyel kerüljenek be a modellépítés alapjául szolgáló példák közé, a vonatkozó névmási visszautalás esetében például a legközelebbi kézzel is annotált antecedens mindig igen közel található, tehát a negatív példák szűrése nem feltétlenül javít összességében az osztályozó eredményességén. Ezt a gondolatot támasztja alá az is, hogy a kizárólag az egyes szám harmadik személyű személyes névmásokat vizsgáló teszt esetében az EXP3, a vonatkozó névmások esetében pedig az EXP4 mutatkozott jobb stratégiának. Más osztályozó esetében természetesen lehetséges, hogy eltérő eredményeket kaphatunk, így a végső konklúzió levonása további kutatást igényel.

A második hipotézisem a névmáshoz antecedensként azonosított főnévi csoportok szűrésére vonatkozott. Mivel az általam alkalmazott Mention-pair technika több főnévi csoportot is megenged, mint a névmás antecedense, a névmási anaforafeloldás során azonban egy antecedens azonosítása a cél, két szűrési módszert hasonlítottam össze. A hipotézisem az volt, hogy a Best-first módszer alkalmasabb lesz a helyes antecedens azonosítására, hiszen a névmáshoz tartozó legközelebbi antecedens is gyakran több tagmondatnyi távolságra található a névmástól. A kísérletek alapján ez a hipotézis részben volt helytálló. A két módszer közötti különbségeket nagyban befolyásolja a tanítófájlban szereplő példák aránya. Abban az esetben, ha a pozitív példák nagyobb arányban vannak jelen a tanítófájlokban, tehát távolabbi pozitív esetek is bele kerülnek, akkor a tesztelés során is előfordul, hogy távolabbi antecedenst azonosít az osztályozó és a döntés során a Best-first alkalmasabbnak bizonyul, mint pusztán a közelség (Closest-first) alapján való döntés. Ha a legjobb eredményeket elért kísérleteket vesszük figyelembe, akkor az összes személyes névmás és a mutatónévmás esetében, ahol a tanítófájlban a példák aránya megegyezett a tesztfájlban található példák arányával nem mutatkozott különbség a két módszer között. Az egyes szám harmadik személyű személyes névmás esetében a Closest-first, míg a vonatkozó névmás esetében a Best-first módszer ért el jobb eredményt.

A harmadik és egyben utolsó hipotézisem az volt, hogy a morfológiai és szintaktikai jellemzőket kognitív alapon megfogalmazott jellemzőkkel kiegészítve jobb eredményeket fog elérni az osztályozó, mint önmagában a korpuszokban megtalálható nyelvi jellemzők segítségével. Ez a hipotézisem összetettebb volt a többinél, hiszen öt dolgot vizsgáltam ezen a témakörön belül. Egyrészt meghatároztam kognitív nyelvészeti alapon egy távolságszámítási

módszert, amelynek során figyelembe vettem a tagmondatok egymással való viszonyát is, majd ezt hozzáadtam a gépi tanulási kísérletekhez és összehasonlítottam azzal a távolságszámítási módszerrel, amelynek során minden a névmás és az antecedense közötti tagmondati határátlépés azonos módon növeli a távolság értékét. Az általam meghatározott távolságszámítás növelte az osztályozó eredményességét a vonatkozó névmási visszautaláson kívül minden esetben, így a hipotézisemnek ez a része helytálló volt. Ezen kívül négy jellemzőcsoportot határoztam meg, az első az antecedensjelölt határozottságát (Def), a második a hosszát (Length), a harmadik a pozícióját (Pos) vizsgálta a negyedik pedig azt vizsgálta, hogy az antecedensjelölt névmás-e (Pron). Ugyan az összes személyes névmás és a vonatkozó névmás esetében a Pron, a mutató névmás esetében pedig a Length jellemző hozzáadásával épített modell érte el a legjobb eredményeket, ezek a jellemzőcsoportok összességében nem, vagy nagyon keveset javítottak az osztályozó eredményességén az egy antecedensjelöltre leszűrt változatokban. Így a hipotézisemnek ez a része nem mondható helytállónak. Ennek oka lehet, hogy a személyes névmás esetében az osztályozó nagyobb mértékben tud a morfológiai tulajdonságok egyeztetésére támaszkodni, a vonatkozó névmás esetében pedig, ahol az antecedens mindig közel található a távolság lesz az egyik legfontosabb jellemző.

Az általam végzett kísérletek alapján összességében elmondható:

1 Érdemes gépi tanulási kísérleteket végezni a koreferencia és anafora szempontjából kézi annotációt tartalmazó magyar nyelvű korpuszokon, hiszen szemantikai információ nélkül is azonosít az osztályozó anaforikus párokat.

2 Rámutattam arra, hogy a gépi tanulási kísérletek során magának az annotációnak a jellege is fontos, hiszen ez befolyásolja a pozitív és negatív példák mennyiségét és minőségét. Szintén fontos tapasztalat az adott feladat szempontjából a szövegtípus kérdése, hiszen a SzegedKoref korpuszon végzett kísérletek eredményei nagyban eltérnek a KorKorpuszon végzett kísérletek eredményeitől. Ezek a tapasztalatok megerősítik, hogy a névmási anaforafeloldás erőteljesen szövegfüggő feladat.

3 A szöveg és annotáció típusán felül maga a névmástípus is fontos szempont, hiszen az általam végzett kísérletek eredményein jól látszik, hogy az egyes névmástípusok eltérően viselkednek a vizsgált szempontok alapján egymástól.

4 A tanítófájlokban megtalálható párok aránya, a jellemzők, valamint az egy jelölt azonosítására vonatkozó módszerek tovább tesztelhetők más algoritmusokkal, így ezek az eredmények összehasonlíthatóvá válnak a már más nyelvekre elvégzett hasonló kísérletekkel.

5 Bizonyítást nyert, hogy a névmási anaforafeloldás során mindenképpen érdemes figyelembe venni a két kifejezés közötti tagmondatok egymáshoz való viszonyát, hiszen a Cp2 távolságszámítási módszer majd minden esetben javított az eredményeken a szimpla tagmondati határátlépések egyesével való számolásához viszonyítva.

6. Kísérleteim további eredménye az a megállapítás is, hogy az általam vizsgált jellemzők hatása eltérhet abban az esetben, ha nem kizárólag egy antecedens azonosítása a feladat célja. Azokban a tesztekben, ahol minden pozitív pár azonosítása cél volt, már nagyobb eltérések láthatók, az egyes névmások alapján az általam megfogalmazott jellemzők között: a Pron jellemző a személyes névmás, a Length jellemző a mutatónévmás a Def jellemző pedig a vonatkozó névmás esetében érte el a legjobb eredményt. Az általam definiált kognitív alapú jellemzők ezek alapján nagyobb hatást gyakorolhatnak a koreferenciafeloldás során.

Hivatkozások

Aone, C., & Benett, S. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-1995)* (o. 122–129).

Ariel, M. (2014). *Accessing noun-phrase antecedents.* Routledge.

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

Csendes, D., Csirik, J., & Gyimóthy, T. (2004). 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)* (o. 19–23).

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

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.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium.* (o. 95–126).

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

Hobbs, J. (1978). Resolving pronoun references. *Lingua, 44,* 311–338.

Kocsány P. (2016). A mondatközi anafora és az ő névmás szerepei. *Jelentés és Nyelvhasználat, 3,* 117–150. https://doi.org/10.14232/JENY.2016.1.6

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)* (o. 104–111).

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

Vadász, N. (2020). KorKorpusz: Kézzel annotált, többrétegű pilotkorpusz építése. In G. Berend, G. Gosztolya, & V. Vincze (Szerk.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia* (o. 141–154).

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. *Proceedings of the 6th conference on Message understanding - MUC6 '95*, 45.

Vincze, V., Hegedűs, K., Sliz-Nagy, A., & Farkas, R. (2018). SzegedKoref: A Hungarian Coreference Corpus. In *11th edition of the Language Resources and Evaluation Conference*. European Language Resources Association.

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., & Csirik, J. (2010). Hungarian Dependency Treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*.

Az értekezés témakörében megjelent publikációim

Kovács, V. (2017). Koreferenciaviszonyok vizsgálata enyhe kognitív zavarban szenvedők beszédátirataiban. In Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017 (pp. 120–130).

Kovács, V. (2019). Az elérhetőségi elmélet névmási anaforafeloldásra gyakorolt hatása. In Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019 (pp. 113–121).

Kovács, V. (2020). A tagmondati távolságszámítás módjainak hatása a névmási anaforafeloldásra. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (pp. 129–139).

További az értekezés témaköréhez kapcsolódó társszerzős publikációk

Kovács, V., Simkó, K. I., & Szécsényi, T. (2016). Szabályalapú szintaktikai elemző szintaktikai szabályok nélkül. In XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016) (pp. 251–259). http://doi.org/10.13140/RG.2.1.2705.9284

Katalin, I. S., Viktória, K., & Veronika, V. (2017). USzeged: Identifying Verbal Multiword Expressions with POS Tagging and Parsing Techniques. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) (pp. 48–53).

Simkó, K. I., Kovács, V., & Vincze, V. (2017). Szintaktikai címkekészletek hatása az elemzés eredményességére. In XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017) (pp. 316–322).

Simkó, K. I., Viktória, K., & Veronika, V. (2018). Identifying verbal multiword expressions with POS tagging and parsing techniques. In Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop (pp. 227–243).

Simkó, K. I., Kovács, V., & Vincze, V. (2018). Igei többszavas kifejezések felismerése nyelvfüggetlen módszerekkel. In XIV. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2018 (pp. 381–392).

Szécsényi, T., Kovács, V., & Bús, I. V. (2019). [mi[ti[ők]]]. JELENTÉS ÉS NYELVHASZNÁLAT, 6(2), 165–191. http://doi.org/10.14232/JENY.2019.2.12

Szécsényi, T., & Kovács, V. (2020). A topikalizálhatóságot befolyásoló tényezők statisztikai vizsgálata. In Általános nyelvészeti tanulmányok XXXII. (pp. 237–247).