Nulla vagy semmi? Esetegyértelműsítés az ablakban

Ligeti-Nagy Noémi[1,3], Vadász Noémi[1,3], Dömötör Andrea[1,3], Indig Balázs[2,3]

[1]Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar
2087 Piliscsaba, Egyetem u. 1.
[2]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter u. 50/A
[3]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/A
VEZETEKNEV.KERESZTNEV@itk.ppke.hu

Kivonat A testes esetragot magukon nem viselő névszók mondatbeli szerepének azonosítása egy gépi mondatelemző számára nem triviális feladat. A magyarban az ilyen elem többféle szerepet is betölthet: lehet a mondat alanya; egy jelöletlen birtokos szerkezet birtokosa; egy névutós szerkezetben a névutó vonzata; egy, a főnévi fejet módosító elem; vagy névszói állítmány. A cikkben egy eljárást ismertetünk, amely pusztán a kérdéses névszót követő két tokenből nagy százalékban képes megoldani az esetegyértelműsítést.
Kulcsszavak: esetegyértelműsítés, korpusznyelvészet, mondatelemző

1. Bevezetés

A cikk az AnaGramma nyelvi elemzőrendszer [1,2] számára kínál egy esetegyértelműsítő eljárást, amely a testes esetragot nem viselő névszók mondatbeli szerepét képes tisztázni a névszó szűk környezete alapján. A nyelvi elemzőrendszer működési alapelveinek ismertetése után a testes esetrag nélküli névszók lehetséges mondatbeli szerepeit vesszük sorra, majd bemutatjuk magát az esetegyértelműsítő algoritmust.

Rögtön az elején szükséges tisztázni két terminológiai kérdést. A főnevek (köznevek és tulajdonnevek), a melléknevek, a számnevek és a melléknévi igenevek alkotják azt a csoportot, amelynek elemeit mi most tárgyalni kívánjuk, ezekre a továbbiakban az egyszerűség kedvéért együttesen névszóként fogunk hivatkozni.

Mondatbeli szerep alatt jelen tanulmányban egyrészt a hagyományos értelemben vett mondatbeli szerepet értjük, azaz azt, hogy az adott szó vonzat vagy határozó, másrészt mindazokat a szerepeket, amelyek főnévi csoporton belüli viszonyokban tölthetők be. Tehát a mondatbeli szerep azonosítása annak meghatározását jelenti, hogy az adott névszó vonzat (vagy határozó)-e a mondatban, és ha nem, akkor milyen egyéb funkciója van egy főnévi csoporton belül.

2. Előzmények

A testes esetrag nélküli névszók mondatbeli szerepét tisztázó algoritmusunk az AnaGramma nyelvi elemzőrendszer keretrendszerébe illeszkedik, amely az emberi szövegfeldolgozást modellálja, ezért az elemzendő szöveget balról jobbra és szavanként elemzi. Az aktuálisan vizsgált szó elemzését az azt közvetlenül követő néhány szó is befolyásolhatja, amihez az AnaGramma egy +2 token méretű előretekintő elemzési ablakot használ, melyben a kétfázisú mondatelemzés [3] első fázisa valósul meg. Az elemzés első fázisában a mondatot alkotó összetevők előkészítése történik, amelyek a második fázisban megkapják a mondatbeli szerepüket. A névszók esetében az első fázisban történik a testes esetrag nélküli elemek lehetséges mondatbeli szerepének tisztázása[1].

2.1. A testes esetrag nélküli névszók

A magukon testes esetragot nem viselő névszók szerepe a mondatban megfigyeléseink alapján a következő lehet: lehet a mondat alanya, jelöletlen birtokos szerkezet birtokosa, egy névutós szerkezetben a névutó vonzata[2], főnevet módosító elem vagy névszói állítmány (része). A különböző mondatbeli szerepek alapján megállapíthatjuk, hogy a testes esetragot nem viselő névszó végén vagy egy testetlen esetrag van (ha a névszó alanyesetben vagy jelöletlen birtokos esetben van), vagy nincs semmi (ha főnévi fejet módosít vagy egy névutó vonzata). A névszói állítmány szerepében álló névszók esetében (elsősorban technikai megfontolásból) szintén egy testetlen esetragot feltételezünk, amely a névszó „állítmány-esetét" jelöli. A cikkben a testetlen esetragokat és az esetragnélküliséget következetesen így jelöljük:

- az alanyeset és a jelöletlen birtokos testetlen esetragja: α
- az esetragnélküliség jele: 0
- az állítmány esetének testetlen esetragja: β

A 3. fejezetben ismertetésre kerülő korpuszvizsgálataink során szembesültünk azzal, hogy a fentieken kívül még két további olyan eset van, amikor a testetlenség jelöl egy esetet vagy egy főnévi csoporton belüli viszonyt. Az egyik a vokatívusz (ld. az 1-es példában a *báróné* tokent, amely egy testes esetrag nélküli főnév, melynek a mondatbeli szerepe itt a „megszólítás-eset"). A másik esetet nehéz a nevén nevezni, a 2a és a 2b alatti példákkal igyekszünk megvilágítani (a mondatok az MNSZ 2.0.4-es verziójából [5] származnak). A példákban félkövérrel szedtük a kérdéses elemet. Minden esetben többtagú nevekről van szó; vagy egy tulajdonnév és egy köznév kapcsolatáról, ahol a köznév a tulajdonnév birtokosának foglalkozását, társadalmi szerepét, nem élő dolgok esetében például

[1] Az igék esetében pedig a lehetséges vonzatkeretek megszorítása történik [4].

[2] Az, hogy a testes esetrag nélküli, névutó előtti névszót a névutó vonzatának hívjuk, csupán egy praktikus megfontolásból alkalmazott egyszerűsítés, nem elméleti nyelvészeti megalapozottságú terminológiahasználat. A névutó és az esetrag nélküli névszó együttjárását kívánjuk vele érzékeltetni: nincs névutó névszó nélkül.

funkcióját (*Opel személygépjármű*) nevezi meg, vagy több köznév kapcsolatáról (*elnök úr*), ahol hasonló az elemek közötti viszony, mint a tulajdonnév-köznév kapcsolat esetén. Ezekben a helyzetekben az „előtagon", a félkövérrel szedett elemeken nem feltételezhetjük egyik eddig említett esetnek a jelenlétét sem. Ezt az esetet mi most *többtagúnév-esetnek* hívjuk, de elemzése mélyebb kutatást igényel, így ezen tanulmány keretein belül nem vizsgáljuk. Sem a vokatívusz, sem a többtagúnév-eset egyértelműsítése nem célja a 3. fejezetben bemutatott algoritmusunknak, ezekkel később, a kutatás egy következő fázisában kell foglalkozni.

(1) A hintóm rendelkezésére áll, **báróné**!

(2) a. **Tóth** kisasszony a lövést követően állítólag a park fái mögé bújt.

b. Ölelgetnétek ti még engem, hogy így, aranyos **elnök** elvtárs, meg úgy aranyos **elnök** úr ...

A jelen cikkben ismertetett `Nom-or-What` eljárás az AnaGramma kereteiben működő `Nom-or-Gen` eljárás [6] továbbfejlesztett és kiegészített változata, melynek a célkitűzése az, hogy egy testes esetrag nélküli elem esetében annak szűk kontextusa alapján eldöntse, hogy egy jelöletlen birtokos szerkezet birtokosa-e. A `Nom-or-Gen` eljárás a főnevet módosító elemeket egységesen `NPMod` kategóriával látta el. Emellett a főnevet módosító elemek rendelkeznek csak a saját szófaji kategóriájukra jellemző jegyekkel is (pl. a számnevek a csak a számnevekre jellemző jegyekkel). Mivel a korpuszokban a testes esetrag nélküli névszók esetcímkéje kivétel nélkül `NOM`, és az esetegyértelműsítő eljárás feladata tisztázni ezeknek az elemeknek az aktuális szerepét, ezért az eljárás a `Nom-or-What` nevet kapta.

A `Nom-or-Gen` eljárást tehát úgy feljesztettük tovább, hogy immár az alanyesettel, a főnevet módosító elemekkel és a névutók vonzataival kapcsolatban is képes legyen döntést hozni. Az állítmányi szerepű névszók esetében csak akkor tudna dönteni, amikor a névszó szűk környezetében szerepel olyan információ, amely ezt lehetővé teszi. Ez azonban ritkán fordul elő, az állítmányi szerep felismerésében ugyanis jellemzően inkább a névszót megelőző mondatrész nyújt nagyobb segítséget. Az ablak alapján a névszói állítmányi eset csak akkor azonosítható egyértelműen, ha az ablakban létige található, sőt, még ezt is le kell szűkíteni 1-2. személyű létigére, egyébként nem lehetünk biztosak benne, hogy a létige valóban kopula. Míg a 3a példában az ablak alapján egyértelműsíthető a névszói állítmányi eset, addig a 3b és a 3c esetragjainak elkülönítéséhez az előzmények ismeretére is szükség van. Még nehezebb a helyzet, ha egyáltalán nincs a mondatban testes kopula, ilyenkor két fő ismérvre támaszkodhatunk: 1) a mondatban nincs finit ige, 2) a mondatban már van nominatívusz. Mindkét jellemző csak a nagyobb kontextusból, jellemzően az állítmányi szerepű névszót megelőző mondatrészből állapítható meg.

(3) a. *Negyedik **gyerek** voltam a családban.*
gyerek/gyerek/FN.β

b. *Erdélyi Dániel maga is iskolás **gyerek** volt a film ábrázolta korszakban.*
gyerek/gyerek/FN.β

c. *Kevés zsidó **gyerek** volt a falumban.*
gyerek/gyerek/FN.NOM

Mindezek alapján azt feltételezzük, hogy az alanyesetet, a birtokos esetet, a főnévi módosító szerepet és a névutó vonzatát a kétfázisú mondatelemzés első fázisában tisztázzuk, amikor a mondat elemeit előkészítjük arra, hogy megkapják szerepüket a mondatban. A névszói állítmány megtalálása a kétfázisú mondatelemzés második fázisában történik, hiszen felismeréséhez szélesebb kontextus szükséges. Ebben a cikkben csak az elemzés első fázisával foglalkozunk.

3. Módszer

Kutatásunk első lépéseként a Szeged Treebank 2.0 [7] főnévi csoportként jelölt egységein belüli, testes esetrag nélküli névszói (és melléknévi igenévi) tokenek viselkedését vizsgáltuk. Már a kutatás korai fázisában egyértelművé vált, hogy külön szabályrendszert kell alkotni a főnevekre, a melléknevekre, a számnevekre és a melléknévi igenevekre, mivel jellemzően máshogy viselkednek, ami a testes esetrag nélküli tokeneiket illeti. Mindegyik szófaji kategóriához két-két listát készítettünk: egyet azokról az esetekről, amikor az adott szófaji kategóriához tartozó, NOM-nak címkézett elem egy NP belsejében található, és azokról az esetekről, amikor az adott szófaji kategóriához tartozó, NOM-nak címkézett elem egy NP legutolsó eleme. Ezen listák alapján állapítottuk meg, melyek azok az elemek, amelyek előtt biztosan eldől egy NOM jelentése, és melyek azok, amelyek előtt nem. Megfigyeléseinket, és az algoritmus működését döntési fák segítségével szemléltetjük (1., 2. és 3. ábrák).

Mivel az algoritmus az ablakban szereplő tokenekre támaszkodva dönt, értelemszerűen lesznek olyan esetek, amikor nem egyértelműsítheti a testetlenséget. Ezekre az esetekre *alapértelmezett (default)* értéket kellett meghatároznunk: a főnevek (mind a köznevek, mind a tulajdonnevek) esetében ez az *alapértelmezett* érték az α, az alanyeset és a jelöletlen birtokos eset testetlen esetragja; a melléknevek, számnevek és a melléknévi igenevek esetében pedig a *default_*0, ami abban különbözik a 0-tól, hogy nem mond egyértelmű ítéletet, az adott NPMod kategóriájú elem még NOM vagy GEN címkét is kaphat később, a mondat egészének ismerete alapján. A 4a példában az *ember* token mondatbeli szerepe az ablakban látható elemek alapján nem állapítható meg biztosan, ezért α esetragot kap. A 4b példában az *egyik* esete nem egyértelműsíthető az ablak alapján, ezért *default_*0 esetet kap.

(4) a. *Elindul reggel hazulról az* ***ember*** *ingujjasan vagy pulóverben.*
ember/ember/FN.α

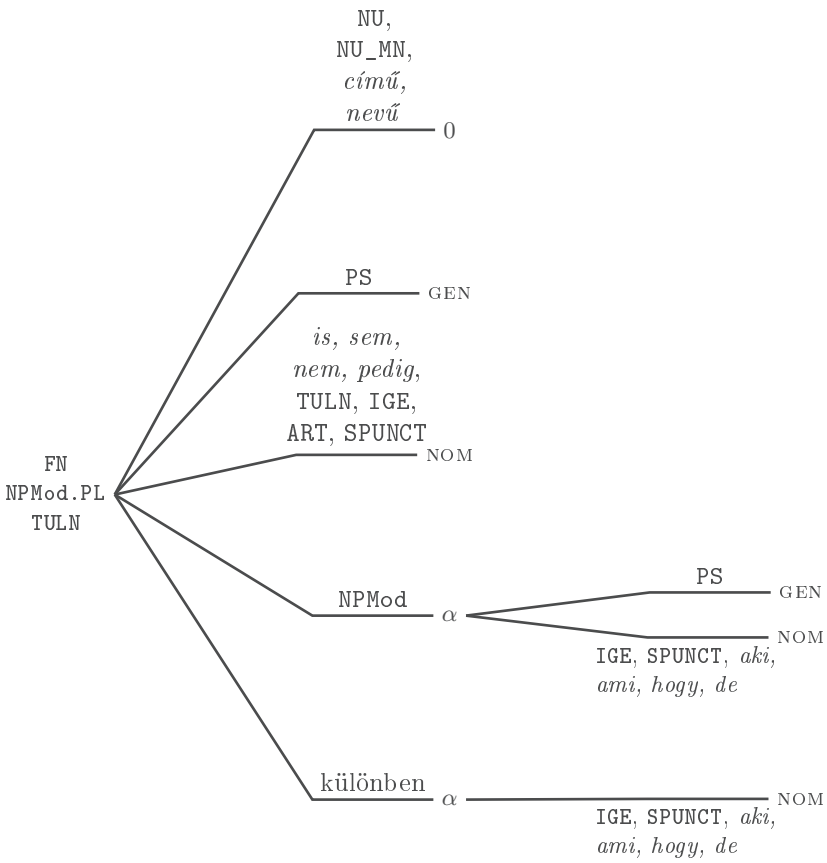
b. *A Nagy_ Szent_ Bazil-rendnek két kolostora is működött, az* ***egyik*** *Veszprémben, a másik Dunapentelén.*
egyik/egyik/MN_NM.default_0

Az 1. ábrán látható a főnevekre (köznevekre és tulajdonnevekre), illetve többesszámú melléknévnek, számnévnek, melléknévi igenévnek annotált elemekre vonatkozó szabályok összefoglalása döntési fában. A fa gyökere az aktuális elem szófaji címkéje. A fa első szintjének élein az ablakban látható első elemen található információk szerepelnek. Például az ablak első elemén látható NU címke következtében az algoritmus a 0 esetet ítéli meg az aktuálisan vizsgált token NOM esetragjának helyére. A fa második szintjén lévő élek az ablak második elemén látható információkat tartalmazzák: ezek csak akkor aktivizálódnak, ha az első elem alapján az algoritmus nem tudott dönteni, és a *default* α esetet illesztette a NOM helyére. Ekkor az ablak második elemén látható bizonyos címkeelemek alapján még van lehetősége egyértelműsíteni az esetet. Az ábrán nem tüntettük fel, de a 0 egyértelműsítése után (a fa második ágának végrehajtása előtt) történik egy lépés, amely független az ablak első elemétől. Ha az aktuálisan vizsgált token semmilyen esetben sem lehet jelöletlen birtokos szerkezet birtokosa, akkor az algoritmus a NOM esetet ítéli meg a szónak, és megtörtént az egyértelműsítés. Ilyen tokenek az *az, ez, mindaz, mindez, aki, ami*. Fontos kitétel, hogy az ablak második elemén látható birtokos személyjel (PS) csak abban az esetben vezet az aktuális elem NOM címkéjének GEN esetragra cseréléséhez, ha az aktuális elemen nincsen birtokos személyjel. Ezzel a *Magyarország* ***kormánya*** *mostani megbízásából* szerkezetek előfordulását zárjuk ki. A *Magyarország* ***kormánya*** *megbízásából*-típusúak az algoritmus számára is megítélhetőek, az ablak első elemén látható birtokos személyjel alapján. A példákban félkövérrel szedett tokenek az aktuálisan vizsgált elemek.

A legfelső élen, a NU társaságában található a NU_MN kategória. Bár ilyen címkét az MNSz2-ben nem találunk, mi fontosnak tartjuk a megnevezését: az *alatti, általi, mögötti* stb. szavak tartoznak ide. Ezek, a névutókhoz hasonlóan, azonnal egyértelműsítik az őket megelőző névszói elem végén a 0 esetet. Hozzájuk hasonló a *című, nevű* tokenek szerepe, ezért azokat is feltüntettük itt[3].

A 2. ábrán látható a(z egyesszámú) melléknevekre és melléknévi igenevekre vonatkozó szabályok összefoglalása döntési fában. Ennél a szabálycsoportnál is fontos kitétel, hogy az ablak második elemén látható birtokos személyjel (PS) csak abban az esetben vezet az aktuális elem NOM címkéjének GEN esetragra cseréléséhez, ha az aktuális elemen nincsen birtokos személyjel. Végül a 3. ábra a számnevekre vonatkozó szabályok összefoglalása.

[3] Megjegyzendő, hogy a *című* és a *nevű* előtti eset inkább a többtagú nevek esetével azonos, de ez további kutatást igényel.

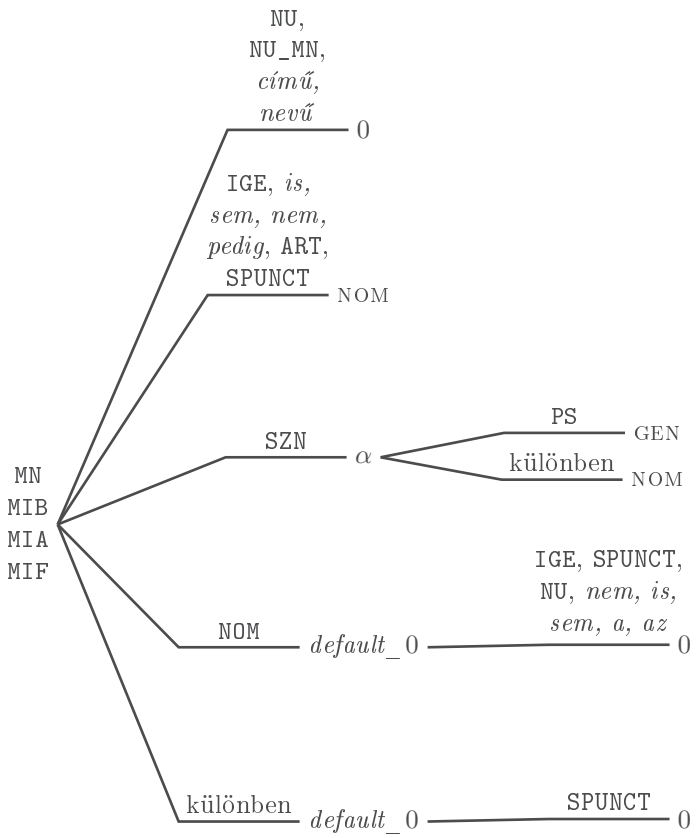


1. ábra: A főnevekre (köznevekre és tulajdonnevekre), illetve többesszámú melléknévnek, számnévnek, melléknévi igenévnek annotált elemekre vonatkozó szabályok összefoglalása döntési fában. A fa gyökere az aktuális elem szófaji címkéje. A fa első szintjének élein az ablakban látható első elemen található információk szerepelnek. A fa második szintjén lévő élek az ablak második elemén látható információkat tartalmazzák.

4. Eredmények

Az algoritmus teljesítményét 1 000 darab tesztmondaton mértük ki. A tesztmondatokat az MNSZ2 szolgáltatta, és azzal a megszorítással éltünk, hogy a mondatban legyen legalább egy finit ige. A kapott 1 000 darab véletlenszerűen kiválasztott mondaton három változtatást eszközöltünk:

– Mivel a tulajdonneveknek jól körülhatárolható, fontos funkciója van az őket megelőző elem esetének egyértelműsítése során, ezért a tesztmondatainkon kézzel annotáltunk minden tulajdonnevet. A többelemű tulajdonneveket _ jellel összekapcsoltuk, és `FN` címkéjűket `TULN` címkére cseréltük.

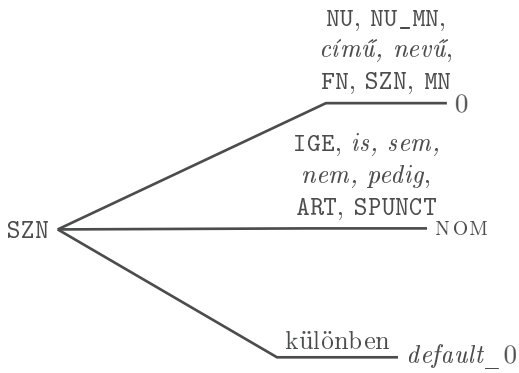


2. ábra: A(z egyesszámú) melléknevekre és melléknévi igenevekre vonatkozó szabályok összefoglalása döntési fában. A fa gyökere az aktuális elem szófaji címkéje. A fa első szintjének élein az ablakban látható első elemen található információk szerepelnek. A fa második szintjén lévő élek az ablak második elemén látható információkat tartalmazzák.

– A kopulás, vagy finit igét nem – legfeljebb létigét – tartalmazó tagmondatokat töröltük. Ha egy egész mondatot kellett a kopula miatt törölni, akkor ahelyett újat kértünk a korpuszból.
– Az esetlegesen bennragadt, „szemétnek” minősülő elemeket, úgy mint a sorszám a mondat elején, vagy a végén, töröltük a mondattokenekből.

Az így megtisztított 1 000 darab mondat minden egyes, eredetileg NOM címkével rendelkező eleme három elemzést kapott: egyet az algoritmustól, egyet a kézi annotáció során az ablak elemeire támaszkodva, és egyet a kézi annotáció során mint végleges elemzés.

A kétféle kézi annotációval a célunk az, hogy összetettebben mondhassunk ítéletet mind az algoritmus megvalósításáról, mind a mögötte húzódó elméleti



3. ábra: A számnevekre vonatkozó szabályok összefoglalása döntési fában. A fa gyökere az aktuális elem szófaji címkéje. A fa első szintjének élein az ablakban látható első elemen található információk szerepelnek.

megfontolásokról. Az ablak alapján hozott manuális ítélet (az 5-ös példában az α) lényege, hogy kiértékelhessük, mennyiben valósítja meg az algoritmus a tőle elvárt viselkedést. Ha az algoritmus jól teljesít, az azt jelenti hogy az ablakban látható elemek alapján a lehető legpontosabban meghatározza, hogy az adott testes esetrag nélküli elem milyen szerep(ek)et tölthet be ebben a mondatban, és mindezt helyesen végzi. A teljes mondat alapján definiált annotáció (az 5-ös példában a NOM pedig az adott tokenek végső szerepét mondja meg (jelen esetben a *patak* a mondat alanya). Ezzel az algoritmus mögött meghúzódó elméleti elgondolásokról szeretnénk ítéletet mondani: ha ez a kézi elemzés egyezik azzal, amit a kézi annotáció során az ablak alapján mondtunk, akkor jól használható és pontos az eljárásunk; az ablak alapján nagy bizonyossággal meg lehet határozni egy névszó mondatbeli szerepét, anélkül, hogy alaposabban körül kéne néznünk (más szavakkal: a kétfázisú mondatelemzés első fázisában ezek a szerepek jól eldönthetők).

(5) *A **patak** tőlük keletre húzódott.*
az ablak: *patak tőlük keletre*
az algoritmus ítélete: *patak FN.**nom***
kézi annotáció, az ablak alapján: *patak FN.*α
kézi annotáció, a teljes mondat alapján: *patak FN.**nom***

A kézi annotáció során, első körben, tehát az ablak alapján való döntésnél, a következő címkéket kaphatták a testes esetrag nélküli tokenek:

- NOM: nominatívusz
- GEN: birtokos
- 0: esetrag nélküli (névutó előtti névszó[4], vagy más névszó módosítója)

[4] A névutó előtti névszó végén lévő testetlenség pontos jelentése kérdéses: jogos elképzelés lenne egy NOM jelenlétét feltételezni, az esetragos névszót vonzó névutók

- α: nem eldönthető; a főnevek *alapértelmezett* értékét kapja (NOM-má vagy GEN-né egyértelműsödhet később)
- *default_0*: nem eldönthető; az `NPMod` elemek *default* értékét kapja
- VOK: vokatívusz esetű
- *postag_hiba*: valamelyik vizsgált elem hibás szófaji címkét kapott, ezért rossz az elemzés (például melléknévi igenévnek címkézett ige esetében)
- többtagú nevek esete (ld. pl. 2a)

A kézi annotáció második körében, a teljes mondat alapján hozott ítéleteknél a következő címkét kaphatták az esetegyértelműsítésre váró névszók:

- NOM
- GEN
- 0
- VOK
- többelemű nevek esete
- α vagy *default_0*, abban az esetben, ha a teljes mondat kétértelmű

A tesztmondatokban összesen 125 olyan token volt, amelyeknél vagy az adott token annotációja volt hibás (például `NOM` esetragú melléknévi igenévnek volt címkézve egy ige), vagy az ablakban lévő egyik vagy másik szóalak (például az őt követő melléknévi igenév volt igének címkézve). Ezeket az eseteket nem javítottuk, nem számítottuk az értékelésnél. Szintén nem került bele a kiértékelésbe a 34 vokatívuszi esetű névszó, illetve a 45 darab *többtagúnév-esetű* token.

A kiértékelés szabályait és a kategóriákat az 1. táblázat tartalmazza. *True positive* (TP), *false positive* (FP) és *false negative* (FN) kategóriákat állapítunk meg. Az egyes oszlopokat a következőképpen kell értelmezni: ha a „kiértékelendő eredmény" oszlopban látható értékre a „sztenderd" oszlop adott értéke illeszkedik, akkor ez egy TP, FP vagy FN találat, attól függően, melyik sorban található ez a párosítás. Ez a megfeleltetés igaz akkor is, mikor az algoritmus eredményét hasonlítjuk a kézi, csak az ablakot figyelembe vevő annotációhoz, illetve akkor is, mikor a kézi, csak az ablakot figyelembe vevő annotációt a teljes mondatot figyelembe vevő kézi annotációhoz hasonlítjuk. A TP eredmények a teljes egyezések. FP eredménynek a túlspecifikálást tekintjük: ha például az algoritmus egy elemről azt állítja, hogy nominatívusz, de a kézi annotáció szerint az ablak alapján még nem mondhatna ilyet, csak egy *default* α-t, akkor az FP eredmény. Viszont ha alulspecifikál, tehát *default* értéket ad egy elemnek, pedig az ablakból eldönthető lenne pontosabban is, az FN.

4.1. Az ablak kiértékelése

A 2. táblázatban látható eredmények azt mutatják, hogy elméleti alapvetésünk, azaz a kételemű ablak alapján történő egyértelműsítés milyen pontosságot és fe-

példáinak analógiájára, pl. *a kerítésen kívül – a kerítés mellett*; ugyanakkor a névutós és esetragos névszók analógiája ennek ellenkezőjét sejteti, azaz hogy a névszó végén valóban nincsen semmi, ahogy az esetrag előtt sem feltételezünk semmit a szótőn, pl. *az asztalon – az asztal alatt*. Mi most ez utóbbi analógiát tekintjük mérvadónak, és a névutó előtti névszók végén nem feltételezünk NOM vagy más esetragot.

kategória	a kiértékelendő eredmény	a sztenderd
TP	NOM	NOM
	GEN	GEN
	α	α
	*default*_0	*default*_0
FP	NOM	α
	GEN	α
	0	*default*_0
FN	α	NOM
	α	GEN
	*default*_0	0

1. táblázat. A kiértékelés szabályai.

dést eredményez, ha a testes esetrag nélküli névszói elemek teljes mondat alapján történő esetegyértelműsítéséhez hasonlítjuk. Jól látszik, hogy a pontosság 97,73%-kal igen magas. Alapvető célunk az volt, hogy precízen döntsünk, ne kelljen a mondatelemzés egy későbbi fázisában korrigálni korábbi, később tévesnek bizonyuló ítéleteinket. Ennek a feltételnek sikerült megfelelnünk, ezt a magas pontosság szemlélteti.

TP	FP	FN	pontosság	fedés	F-mérték
1 590	37	761	97,73%	67,63%	79,94%

2. táblázat. Az ablak alapján történő kézi annotálás eredményeinek összehasonlítása a teljes mondatot figyelembe vevő kézi annotálás eredményeivel.

A fedés ugyanakkor csak 67,63%: ez legfőképpen a *false negative* találatok magas száma miatt van, ami bővebb magyarázatot igényel. Azt az esetet tekintettük FN találatnak, ha az ablak szerinti kézi annotációnál az *alapértelmezett* értéket kapta egy elem, a teljes mondat alapján azonban már specifikusabb eredményre jutottunk. Tehát az alulspecifikáltságot tekintjük fals negatív eredménynek. A 3. táblázatban látható a FN találatok részletezése.

Ugyanakkor látni kell azt is, hogy ezek nem feltétlenül hibák - a főnevek *alapértelmezett* esetének számító α pontosan a NOM és a GEN esetragot fogja össze: ezekben az esetekben az történik, hogy az ablak alapján még nem állapítható meg egyértelműen, hogy az α esetragú névszó alanya vagy egy birtokos szerepű tagja a mondatnak, ezt csak a tágabb kontextus segítségével lehet eldönteni. Izgalmasabb azonban a *default*_0-k esete: kiugróan magas azoknak a melléknéveknek és melléknévi igeneveknek a száma, melyeknél az *alapértelmezett* esetet ítéltük meg az ablak alapján, de az esetük a mondat alapján 0. Sőt, összesen 6

hibatípus	hibaszám
FN	761
NOM helyett α	186
GEN helyett α	56
0 helyett *default_*0	519

3. táblázat. A kétféle kézi annotáció összehasonlításakor megfigyelt *false negative* eredmények. Az egyes sorok azt mutatják, milyen eset helyett milyen esetet egyértelműsített (alulspecifikálva) a csak az ablak alapján történő kézi annotáció.

olyan eset fordult elő, hogy az ablak alapján *default_*0-nak ítélt esetet a mondat teljes egésze NOM-ként egyértelműsítette, minden más alkalommal 0 lett ezekből. Mit értünk ez alatt? A 6a példában szemléltetjük, hogy az ablak alapján még elképzelhetőnek tartjuk, hogy az adott token a mondat alanya lesz, vagy egy jelöletlen birtokos szerkezetben a birtokos, de a teljes kontextus (6b) egyértelművé teszi, hogy ez egy főnevet módosító elem.

(6) a. *telepített mintegy negyven*
ablak alapján: telepített telepít IGE._MIB.*default_*0

b. *Kétéves koromban elvesztettem anyai nagyszüleimet, s velük együtt a szülőfalumból Magyarországra* ***telepített*** *mintegy negyven családban szinte minden rokonomat.*
mondat alapján: telepített telepít IGE._MIB.0

Ezeknek az eredményeknek az összefoglalásaképpen elmondhatjuk, hogy a magas pontosság megfelel az eljárással kapcsolatosan támasztott elvárásainknak. A fedés javítható lenne (67,63%-ról 89,7%-ra) azzal, ha a melléknevek és melléknévi igenevek az *alapértelmezett* érték helyett automatikusan a 0 esetet kapnák meg az ablak alapján.

4.2. Az algoritmus kiértékelése

A 4. táblázatban az algoritmus teljesítményének kiértékelése látható. Ebben az esetben a gépi esetegyértelműsítést hasonlítottuk össze a csupán az ablakot figyelembe vevő kézi annotációval.

TP	FP	FN	pontosság	fedés	F-mérték
2 220	63	125	97,24%	94,67%	95,94%

4. táblázat. Az algoritmus teljesítményének kiértékelése a csak az ablakot figyelembe vevő kézi annotáción.

Mind a pontosság (97,24%), mind a fedés (94,67%) kiemelkedően magas: az algoritmus teljesítménye megbízható, amit az ablak alapján egyértelműsíteni lehet, azt egyértelműsíti, de nem specifikál olyan eseteket, amiket még nem lehetne. Az egyetlen elemzésre számot tartó érték talán az 5. táblázat utolsó sora, ahol az algoritmus túlzottan alulspecifikál; *default* értéket illeszt a melléknevekre, melléknévi igenevekre, holott 0-t kéne. Ez az aleset párhuzamba vonható a 2. táblázat kapcsán már említett, és a 6a és a 6b példákban bemutatott jelenséggel. Az 5. táblázatban szereplő 112 eset tovább erősíti azt a sejtést, hogy érdemes lenne az `NPMod` kategóriájú elemekre *default*_0 helyett mindig 0-t illeszteni.

hibatípus	hibaszám
FN	125
NOM helyett α	9
GEN helyett α	4
0helyett *default*_0	112

5. táblázat. Az algoritmus teljesítményének kiértékelésekor megfigyelt *false negative* eredmények. Az egyes sorok azt mutatják, milyen eset helyett milyen esetet egyértelműsített (alulspecifikálva) az algoritmus.

5. Összegzés

A cikkben bemutattunk egy algoritmust, mely a testes esetragot magukon nem viselő, ezért automatikusan `NOM` címkével ellátott főnevek, melléknevek, számnevek és melléknévi igenevek esetegyértelműsítését végzi kizárólag az aktuálisan vizsgált token és az azt követő két szó, az ablak információi alapján[5]. A korpuszmegfigyeléseink alapján körvonalazódó szabályrendszert implementáltuk, majd kiértékeltük. Eredményeink szerint igen magas pontossággal és fedéssel teljesített. További feladataink közé tartozik a vokatívusz eset feltérképezése és egyértelműsítő algoritmusának megírása, illetve a többtagú nevek belsejében szereplő elemek esetének megvizsgálása. Ezeket az ablakban szereplő információk mellett a korábban már feldolgozott elemekre támaszkodva szükséges egyértelműsíteni. Mindezt pedig követheti az állítmány-esetű névszók detektálása a tágabb kontextus alapján.

Hivatkozások

1. Prószéky, G., Indig, B.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. Alkalmazott Nyelvtudomány **15**(1-2) (2015) 29–44

[5] Az eljárás implementciója, a tesztfájl és az annotált fájl elérhető itt: `https://github.com/ppke-nlpg/nom-or-what`.

2. Prószéky, G., Indig, B., Vadász, N.: Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In Kas, B., ed.: „Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére. MTA NYTI, Budapest (2016) 223–232
3. Frazier, L., Fodor, J.D.: The Sausage Machine: A New Two-Stage Parsing Model. Cognition **6**(4) (1978) 291–325
4. Vadász, N., Kalivoda, Á., Indig, B.: Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével. In Vincze, V., ed.: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017), Szeged, Szegedi Tudományegyetem Informatikai Tanszékcsoport (2017) 3–12
5. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Calzolari, N., et al., eds.: Proceedings of the 9th International Conference on Language Resources and Evaluation, May 26–31, 2014, Reykjavik, Iceland, ELRA (2014) 1719–1723
6. Vadász, N., Indig, B.: A birtokos esete az ablakkal. In Scheibl, G., ed.: LingDok: nyelvész-doktoranduszok dolgozatai. Szegedi Tudományegyetem (2017) Megjelenés alatt.
7. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In Matoušek, V., Mautner, P., Pavelka, T., eds.: Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings, Berlin, Heidelberg, Springer Berlin Heidelberg (2005) 123–131