Igei vonzatkeretek az MNSZ tagmondataiban

Sass Bálint

MTA Nyelvtudományi Intézet, Nyelvtechnológiai Osztály,
PPKE, Információtechnológiai Kar, MMT Doktori Iskola
joker@nytud.hu

Kivonat: A vonzatkeretekkel foglalkozó mulkálatok célja a magyar vonzatkeretrendszer korpuszalapú feltérképezése. A Magyar Nemzeti Szövegtárból származó egyszerű mondatok vizsgálata után most alkalmassá vált a rendszer tetszőleges morfoszintaktikailag elemzett szöveg feldolgozására az új tagmondatra bontó modul segítségével. A tagmondatok egy igét és annak bővítményeit tartalmazzák, így alkalmas bemenetét képezik a vonzatkeret-illesztő modulnak. A magyar vonzatok és bővítmények tanulmányozására létrejött egy internetes alkalmazás, mely a fenti emailcímen igényelt jelszóval elérhető a http://corpus.nytud.hu/mazsola címen.

1 Bevezetés

A már több mint egy éve folyó munkálatok távlati célja a Magyar Nemzeti Szövegtárban lévő igei vonzatkeretek feltérképezése. A III. MSZNY konferencián a vonzatkeret-táblázatban összegyűjtött, kézzel kódolt, ismert keretek azonosításáról számoltam be [6].

A készülő magyar-angol gépi fordítási projekt számára szükség volt további vonzatkeretekre, a vonzatkeret-táblázat kiegészítésére. Kidolgoztam egy módszert az eddig ismeretlen keretek azonosítására, melynek segítségével korpuszból nyerhetünk ki új vonzatkereteteket [5]. *A vonzatok: nem kompozicionális bővítmények* elvére alapozva az igék és a mellettük álló NP-k alkotta keretjelöltek idiómaságát mértem a *DF (distributed frequency)* idiómasági mértékkel [7]. Eszerint a mérték szerint az a keret az idiomatikusabb, melynek bővítményei az adott formában kevés (szélső esetben egyetlen) igével fordulnak elő (pl. *fittyet hány*). A magasabb idiómasági érték azt jelenti, hogy a keretjelölt valódi vonzatkeret. Az így nyert új keretek kézi ellenőrzést követően a magyar-angol gépi fordító rendszer lexikai erőforrásába kerültek.

Maga a vonzatkeret-felismerés a következőkben fejlődött: az igei vonzatkeretek feldolgozása a cél, ezért csak olyan tagmondatokkal foglalkozom, melyekben található ige vagy főnévi igenév. Természetesen utóbbi is elegendő, mert a főnévi igenév mindig hordozza a neki megfelelő ige vonzatkeretét.

Továbbfejlesztettem az igekötő- és igetőazonosítást, az elváló igekötőket az igetőhöz ragasztottam, a gyakori képzőket (pl. *-hAt*) levágtam, mivel nem befolyásolják a vonzatkeretet.

2 Tagmondatra bontás

2.1 Motiváció

Az az optimális, ha a vonzatkeret-illesztő modul számára mindig pontosan egy egy vonzatkeretet tartalmazó szövegdarabot tudunk bemenetként adni. A korábbiakban az MNSZ egy preparált részkorpuszával dolgoztam: egyszerű heurisztikával kiválasztottam a korpuszból a rövid (maximum tíz szavas), írásjel nélküli mondatokat. Ezek a mondatok nagy valószínűséggel egy keretet tartalmaznak, azonban nyilvánvalóan nem reprezentálják a valós nyelvhasználatot, így a belőlük nyert gyakorisági adatok nem kellően megalapozottak [6]. Egyetlen előny a könnyű feldolgozhatóság.

Jelen cikk lényegi előrelépése, hogy elkészült egy előfeldolgozó, tagmondatra bontó modul, így a rendszer alkalmassá vált tetszőleges szöveg feldolgozására. Most a szöveg tagmondatai képviselik a nagy valószínűséggel egy vonzatkeretet tartalmazó egységet. A tagmondat kifejezést tehát ebben az értelemben használom: a mondat egy vonzatkeretet tartalmazó része, így lényeges követelmény lesz annak garantálása, hogy a tagmondat egy igét tartalmazzon. Sok helyen találkozhatunk a mondatok bizonyos szempontból könnyebben elemezhető, kisebb részekre darabolásával [4], itt is erről van szó.

2.2 Korábbi megoldások

Nincs tudomásom magyar nyelvre alkalmazható, részletesen ismertetett, reprodukálható tagmondatra bontó módszerről. Létezik egy az INTEX/NooJ nyelvfeldolgozó rendszerben implementált eljárás, melynek vázlatos leírása a [8] cikkben olvasható. Ezenkívül két kéziratos leíráshoz jutottam hozzá, melyeket itt most röviden ismertetek: a fenti eljárás részleit tartalmazó kézirat [3], illetve egy másik megközelítést tartalmazó kézirat [9].

A [3]-ban leírt *tagmondathatár-azonosító* rendszer az alábbi tizenegy darab szabályból áll. A szabályokat reguláris kifejezésre emlékeztető szintaxissal írom le, adott szabály illeszkedése esetén a '@' jel helyére kerül be egy tagmondathatár.

1. [,|-] @ [kötőszó|határozószó]? [vonatkozó névmás]
2. [-] @ [kötőszó|határozószó]? [vonatkozó névmás] [bármi]+ [-] [,]? @
3. [,|-] @ [bármi|NP|AdjP]? [*pedig*|*akár*|*azonban*|*viszont*|*ellenben*|*mihelyt*|*tehát*|*ugyanis*]
4. [,|-] @ [NP]? [*meg*]
5. [,|-] @ [határozószó]? [*nehogy*|*mintha*]
6. [,] @ [kötőszó, kivéve: *de*|*illetve*|*illetőleg*|*mintegy*]
7. [,|-] @ [múlt idejű, egyes szám harmadik személyű ige]
8. @ [kötőszó] [kötőszó]
9. @ [kötőszó] [kérdőszó]
10. [,] @ [határozószó|NP]? [kérdőszó]
11. [,] @ [*az* szótőként] [határozói igenév] [,] [meglévő tagmondathatár] [*hogy*]

Az eljáráshoz tartozik még egy a szabályalkalmazások után futó program, mely lehetséges tagmondathatárként megjelöli az összes kötőszót, mely két olyan finit ige között helyezkedik, melyek között még nincs tagmondathatár. Látjuk, hogy a szabályrendszerben részletesen benne foglaltatik, hogy az egyes kötőszók hányadik pozícióban szoktak állni a tagmondathatárhoz képest, és milyen típusú elemek előzhetik meg őket. A cikk közvetlen tagmondatkezdő kötőszókat tartalmazó listáját már sikerrel alkalmaztuk korábban is [1].

A [9]-ben leírt, de nem implementált eljárás igazi célja, hogy megállapítsa a szöveg *kötőszavairól*, hogy szerkezeteket koordinálnak vagy esetleg tagmondatokat kötnek össze, így mintegy melléktermékként kapjuk meg a tagmondatokat.

Az eljárás a következő előfeldolgozó lépésekkel kezdődik:

– az első finit ige előtti és az utolsó finit ige utáni kötőszavakat jelöljük meg koordináló kötőszavaknak;
– ha két finit ige között egyetlen kötőszót találunk, akkor az jelöljük meg tagmondathatárnak.

Ez után hajtandó végre a következő utasítássorozat a szöveg összes kötőszaván:

d) ha a kötőszó két oldalán lévő két frázis különböző típusú, a kötőszó tagmondathatárt képvisel;
di) különben AdjP és AdvP esetén koordinációról van szó;
dii) NP esetén, ha egyezik az eset és az NP-k monotonicitása, akkor koordináció, ha nem egyezik, akkor tagmondathatár;
diii) VP esetén, ha egyezik a szám és a személy, akkor koordináció, ha nem egyezik, akkor tagmondathatár;
div) egyébként tagmondathatár.

A kéziratban [2,9] megfogalmazott fontos elv szerint: a finit ige vonzatai az igét tartalmazó tagmondaton belül vannak. A kézirat említést tesz a magyar névszói állítmányról, mégis a továbbiakban már mindig csak finit igékre hivatkozik, így lehetőséget ad arra, hogy esetleg hibásan bevegyük a finit ige vonzatai közé a szomszédos névszói prédikátum vonzatait is. Mivel ez az eljárás a kötőszavak alapján hoz döntést, nem ad számot azokról a tagmondathatárokról, ahol nincs jelen kötőszó. A kiértékeléskor használt korpuszból származó példamondatok minkét problémára rávilágítanak:

Meglehet: mindkettőre igen a válasz.

Nagy tanulság, hogy győzelmem Kwasniewski ellen azt jelentette volna, Lengyelországban még forradalom zajlik.

2.3 Az inkrementális nyelvtanfejlesztés

Módszeremben lényegében a fent ismertetett első eljárásra építek [3]. A végső szabályrendszer kialakítása során az *inkrementális nyelvtanfejlesztésnek* nevezett módszert alkalmaztam, az alábbiak szerint.

Minden elemi fejlesztési lépés (újabb szabály hozzávétele vagy strukurális változtatás) után lefuttattam a programot egy 600 mondatos tesztkorpuszon, és megvizsgáltam, hogy mit változtatott a kimeneten ez az egy lépés. Manuálisan értékeltem az eredményt. Ha nagyrészt megfelelőnek ítéltem – néhányszor szinte hibátlan működésre is volt példa – akkor megtartottam, ha sok esetben rossz eredményt hozott, akkor elvetettem az adott fejlesztési lépést.

Ehhez az informális értékeléshez nagyon fontos, hogy mindig megfelelő számú olyan eset álljon rendelkezésre, amikor az adott jelenség, amit a fejlesztési lépés kezel, előfordul. Ahelyett, hogy külön az egyes szabályokhoz preparált kisméretű korpuszokat készítenék, megpdóbálok mindig akkora korpuszt venni, amelyben legalább kb. 20 példányban előfordul a jelenség. Esetemben, ha nem volt elég példa, vagy nem volt egyértelmű az értékelés, akkor egy 4500 mondatos tesztkorpuszon végeztem el újból a vizsgálatot. Ha a nagyobb korpuszban sem volt példa a jelenségre, akkor túl ritka lévén, elhagytam az adott lépést. Így folyamatos korpuszkontroll mellett és viszonylag alacsony időráfordítással tudtam a fejlesztést végezni.

Ahhoz, hogy ezt a fejlesztési módszert alkalmazni tudjam, szükséges, hogy a szabályok függetlenek legyenek egymástól. Erre törekedtem is, ezért nincs is a rendszerben olyan, hogy egy már meglévő tagmondathatárra hivatkozom, ahogy az a fenti 11. szabályban történik.

2.4 A tagmondatra bontó eljárás

A fejlesztés során a következő megfigyeléseket tettem:

1. kettőspontnál illetve pontosvesszőnél minden mástól függetlenül meg lehet jelölni a tagmondathatárt;
2. a *meg* kötőszóra épülő (4.) szabályt elhagytam, mert a szó szófaji egyértelműsítés itt bizonytalan, az esetek nagy részében ténylegesen igekötő a kötőszónak jelölt *meg*;
3. a két egymást követő kötőszót váró 8. és kötőszó + kérdőszót váró 9. szabály az esetek nagy részében rossz helyen jelölt ki tagmondathatárt, így ezt a két szabályt elhagytam;
4. kérdőszó elé ékelődő határozószóra nem volt példa, a szabályt elhagytam (10. szabály első fele);
5. kérdőszó elé ékelődő frázis esetén ismét a szófaji egyértelműsítés korlátjába ütköztem, itt legtöbbször a kérdőszónak jelölt *ki* igekötő kapcsán működött a szabály (10. szabály második fele).

Ez a tagmondatra bontó modul előfeldolgozóként a részleges szintaktikai elemzés előtt fut, ezért a szabályokban szereplő frázisokat opcionális 1-2-3 darab tetszőleges szóval helyettesítettem.

Tudjuk, hogy az ige vonzatai vele egy tagmondatban vannak [2,9]. Ezt kiegészíthatjük azzal, hogy csak a tagmondat igéjének a vonzatai vannak a tagmondatban. Ebből következik, hogy az ige-koordinációt nem engedjük meg, tehát két finit ige között mindig van tagmondathatár. Ilyenkor megjelölhetjük az összes közbeeső kötőszót, mint *lehetséges* tagmondathatárt [8]. Egyszerű továbblépés, hogy ha egyetlen közbeeső kötőszó van, akkor az lesz a tagmondathatár [9].

Azt figyeltem meg, hogy nemcsak kötőszó, hanem legalább ugyanolyan gyakran közbeeső központozás (vessző, pontosvessző, kötőjel) is lehet tagmondathatár. Tehát két finit ige között (finit igét közvetlenül követő *volna* nem számít annak) megjelölöm ezen írásjelek *utáni* és a kötőszavak *előtti* összes pozíciót, mint lehetséges tagmondathatárt. Majd ezek közül – heurisztikus döntéssel – mindig a leginkább jobbra esőt választom ki, bízva abban, hogy így a legkisebb az esélye annak, hogy egy felsorolás közepére helyezem el a tagmondathatárt.

3 Kiértékelés

A fentiek szerint kialakított magyar tagmondatra bontó eljárás tesztelésére és kiértékelésére az MNSZ részét képező *Magyar Nemzet* napilap anyagából választottam ki véletlenszerűen 200 mondatot. A következő nagyon egyszerű taggelési útmutató szerint végeztem a tagmondatok manuális bejelölését:

1. Jelöljük be a szövegben a tagmondatokat.
2. Minden finit ige külön tagmondatba kerüljön.
3. A tagmondatvégi központozás minden esetben a megelőző tagmondathoz tartozzon.

Sok esetben nehezen tudtam eldönteni, hogy adott ponton valóban van-e tagmondathatár, vagy nincs. Az alábbi szövegrészben a vessző után például végül nem jelöltem tagmondathatárt:

> *Ami a szabad demokratákat illeti,*
> *Pető Iván lemondása tipikusan jelzésértékű ...*

Az eredmények a következők lettek: a 171 bejelölt tagmondathatárból a program 148-at talált meg (23-at hagyott ki), helytelen tagmondathatárt 29-et jelölt meg, azaz:

pontosság = **83,6%** lefedés = **86,5%**

Ezen mérőszámokat befolyásoló tényező lehet az, hogy a szöveg egy viszonylag bonyolult jogi nyelvezeten írt részletet: egy rendeletszöveget tartalmazott. Az eredeti korpuszban sokszor helytelen volt a mondatok határainak megállapítása. Egyszerűbb szerkezetű szöveg esetén, valamint jobb mondatrabontás alkalmazásával minden bizonnyal még növelhetők ezek az értékek.

Amint várható volt, a hibák főleg olyan pontokon jelentkeznek, ahol szinte semmi konkrét jel nem utal arra, hogy ott egy tagmondat kezdődik, nincs kötőszó (sőt esetleg központozás sem), illetve névszói állítmány van valamelyik tagmondatban. Ilyen példa:

> *A kérdés második felére azt felelném,*
> *minden lehetséges s minden az erőviszonyoktól függ.*

4 Lekérdező

A munkálatok egyik eredményeként létrehoztam egy internetes felületen hozzáférhető nyelvészeti kutatóeszközt, melynek segítségével a magyar igei vonzatkereteket, az igék bővítményszerkezetét tudjuk kvantitatívan tanumányozni. A fenti emailcímen igényelt jelszóval érhető el az alábbi címen:

http://corpus.nytud.hu/mazsola

Adott igetőhöz megadhatunk két darab esetraggal (vagy névutóval) meghatározott bővítményt, az ezekhez tartozó konkrét szótövet is megköthetjük, illetve megadhatunk szóközzel elválasztott szótőlistát. Lehetőség van esetek és szótövek kizárására, az a tagadásra. A felület lehetőséget ad kiegészítő szabadszavas keresésre is, itt tetszőleges kiterjesztett reguláris kifejezést használhatunk. Kérhetjük, hogy a találatok az egyik szempont (az igető vagy valamelyik bővítmény) szerint csoportosítva, gyakoriság szerinti sorrendben jelenjenek meg.

Az eredményoldalon a találatok száma alatt a gyakorisági lista található, alatta pedig a korpuszból származó megfelelő példamondatok sorakoznak, egy kattintással könnyen elérhető elrendezésben.

A korábban kidolgozott idiómasági mérést [5] nagy műveletigénye miatt egyelőre nem integráltam az eszközbe, helyette az adott bővítménynek (*y,* pl. *megdöbbenésnek*) a vonzatkeret többi részéhez (*x*, pl. *ad hangot*) viszonyított MI-értéke jelenik meg.

$$\mathrm{MI}(x,y) = \log_2 N\, \mathrm{f}(x,y) / \mathrm{f}(x)\mathrm{f}(y)$$

Mivel egy lekérdezésben x állandó, ez felírható:

$$= \log_2 C\, \mathrm{f}(x,y) / \mathrm{f}(y)$$

alakban, ez pedig, mivel nem érdekesek a konkrét MI értékek, csak össze akarjuk hasonlítani őket, felírható így:

$$= \log_2 \mathrm{f}(x,y) / \mathrm{f}(y)$$

Ebből az értékből számítódik ki a megjelenítéskor a betűméret: nagyobb betűméret, nagyobb MI-értéket, szokatlanabb vonzatkeretet jelent.

Példák:

vesz + ACC
vesz + ACC(*rész*) + INE
hány + ACC
néz + nemACC

Az eszközt aktívan használjuk a magyar-angol gépi fordító projektben az egyes szabad keretek különféle fix lemmákkal való leköthetőségének vizsgálatakor.

5 Fejlesztési lehetőségek

Tervezem a modul olyan továbbfejlesztését, hogy részleges szintaktikai elemzést követően, vagy azt követően *is* lehessen vele tagmondatokat azonosítani.

A tipikus hibák kiküszöbölésére egyik lehetőség egy általános állítmány-azonosító eljárás kifejlesztése, mely a névszói állítmányokat is felöleli. Ekkor valóban teljesülhetne az az elv, miszerint két állítmány között mindig van tagmondathatár.

Köszönöm Gábor Katának és Varasdi Károlynak, hogy rendelkezésemre bocsátották kéziratait.

Bibliográfia

1. Bottyán, G., Sass, B.: Conjugated infinitives in the Hungarian National Corpus. In Garabik, Radovan (ed.): Computer Treatment of Slavic and East European Languages, 3rd International Seminar (SLOVKO2005), Szlovák Tudományos Akadémia, Pozsony (2005) 27-30
2. Gábor K., Héja E., Mészáros Á.: Kötőszók korpusz-alapú vizsgálata. In Alexin Z., Csendes D. (szerk.): MSZNY2003, SZTE, Szeged (2003) 305-306
3. Gábor, K.: Tagmondathatár-kijelölő rendszer. Kézirat. MTA, Nyelvtudományi Intézet.
4. Kim Ch., Hong M.: A Korean Syntactic Parser Customized for Korean-English Patent MT System. In: Salakoski T. et al. (eds.): Advances in Natural Language Processing. LNCS, Vol. 4139. Springer-Verlag, Berlin Heidelberg New York (2006) 44-55
5. Sass B.: Extracting Idiomatic Hungarian Verb Frames. In: Salakoski T. et al. (eds.): Advances in Natural Language Processing. LNCS, Vol. 4139. Springer-Verlag, Berlin Heidelberg New York (2006) 303–309
6. Sass B.: Vonzatkeretek a Magyar Nemzeti Szövegtárban. In: Alexin Z., Csendes D. (szerk.): MSZNY2005, SZTE, Szeged (2005) 257-264
7. Tapanainen, P., Piitulainen J., Järvinen T.: Idiomatic Object Usage and Support Verbs, In: Proceedings of the 17th COLING – 36th ACL, Montreal, Canada (1998) 1289-1293
8. Váradi T., Gábor K.: A magyar INTEX fejlesztésről. In Alexin Z., Csendes D. (szerk.): MSZNY2004, SZTE, Szeged (2004) 3-10
9. Varasdi K.: Coordination. Kézirat. MTA, Nyelvtudományi Intézet.