A sekély mondattani elemzés további lépései

Recski Gábor

MTA SZTAKI
Nyelvtechnológiai Kutatócsoport
e-mail: recski@sztaki.hu

1. Bevezetés

A sekély mondattani elemzés (shallow parsing), mely a mondatok fő összetevőinek azonosítását jelenti a mély mondatszerkezet feltérképezése nélkül, számos nyelvtechnológiai eljárás kulcsfontosságú lépése. A legnagyobb mondattani egységek pontos azonosítása nélkülözhetetlen lehet a gépi megértésben, a gépi fordításban, de az információkinyerésben és -visszakeresésben is. Cikkünkben elsőként bemutatjuk, hogyan képes az eredetileg főnévi csoportok azonosítására kifejlesztett `hunchunk` rendszer a megfelelő tanulóadat birtokában tetszőleges kategóriájú frázisok azonosítására. A 2.1 fejezetben röviden összefoglaljuk a tanulóadat előállításának és a rendszer tanításának menetét, a 2.2. részben a `hunchunk` felépítéséről ejtünk néhány szót, végül a 2.3 fejezetben értékeljük a rendszer teljesítményét.

A mondat sekély szerkezetének megismeréséhez elengedhetetlen, hogy azonosítani tudjuk a több, gyakran nem szomszédos szóból álló igei szerkezeteket. A 3.1 fejezetben egy olyan eszközt ismertetünk, mely azonosítja egy ige és a tőle különálló igekötő kapcsolatát – felhasználva ehhez a rendelkezésre álló morfológiai elemzést, valamint az egyes igekötős igék gyakoriságáról meglévő ismereteinket is. Ugyancsak a mondatszerkezet hatékonyabb feltérképezését segíti elő, ha képesek vagyunk észlelni az igéből és annak infinitívuszi bővítményéből álló szerkezeteket - a 3.2. fejezetben erre teszünk kísérletet.

2. Mondattani egységek azonosítása

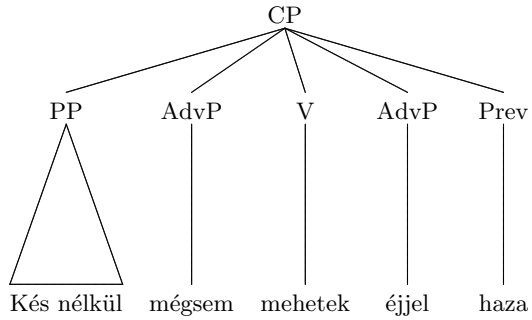
A `hunchunk` rendszer [1] magyar főnévi csoportok azonosítására készült, azonban megfelelő tanulóadat birtokában tetszőleges olyan nyelvfeldolgozási feladatra alkalmas, mely szószintű címkézésként is megfogalmazható. A Szeged Treebank [2] segítségével a főnévtől különböző mondattani kategóriákra is készíthetünk tanulóadatot, így lehetővé téve, hogy a `hunchunk` a legmagasabb szintű mondattani egységeket azonosítsa.

2.1. Tanítás

A Szeged Treebank egy vegyes műfajú, több mint 80000, szintaktikailag teljesen annotált mondatot tartalmazó korpusz. A tanítóadat előállításához a mondat-

tani elemzés legfelső két szintjét használjuk – a legfelső szinten a tagmondatok (CP) különülnek el, az ezek alatti legmagasabb szintű egységek azok, melyeket azonosítani szeretnénk. A korpuszból ugyancsak kinyerhető az egyes szavakra vonatkozó morfológiai információ MSD-kódolásban, ezt a korpusz készítésekor átalakítottuk a KR-formalizmusnak megfelelő alakra [3], mivel az általunk használt `hunmorph` morfológiai elemző [4] is ezt a formátumot követi.

Az egyes frázisokhoz tartozást a szavakhoz rendelt címkék jelzik. A címkézés során a Start/End konvenciót alkalmazzuk [5], mely az elterjedtebb IO és IOB konvencióknál [6] több címkét igényel, ugyanakkor lehetővé teszi többféle frázisbeli pozíció megkülönböztetését: míg az előbbi megoldások vagy egy címkével (I-NP) jelölik a frázishoz tartozó szavakat, esetleg a frázist kezdő szót jelölik külön szimbólummal (B-NP), addig az általunk használt jelölés a chunkhoz nem tartozó szavakon (O) kívül négy címkét használ (B-NP, I-NP, E-NP, 1-NP), melyek rendre a frázis elején, közepén és végén álló, valamint az önmagában frázist alkotó szavakat jelölik. Így a korpuszban található, 1. ábra szerinti elemzéssel bíró mondat az újonnan létrejött korpuszban a 1. táblázat szerinti címkézést kapja.



1. ábra. Mondattani elemzés

1. táblázat. Címkézés

Kés	nélkül	mégsem	mehetek	éjjel	haza	.
B-PP	E-PP	1-ADVP	O	1-ADVP	O	O

Az egyes mondattani kategóriák nagyon különböző gyakorisággal fordulnak elő maximális frázisként a korpuszban (l. 2. táblázat). Mint látható, melléknévi frázis csak elvétve fordul elő tagmondat közvetlen összetevőjeként, akkor is általában hibás annotáció következményeként (vö. *A köd mint* $[_{AdjP}$ melegvizes] *rongy feküdt az arcomon*).

2. táblázat. Kategóriák megoszlása a korpuszban

NP	268726	73.58%
ADVP	79536	21.78%
PP	16925	4.63%
ADJP	34	0.00%
Összesen	365221	100%

2.2. A `hunchunk` rendszer

A `hunchunk` egy felügyelt tanulásra épülő, szószintű címkézési feladatokat ellátó eszköz, melyet sikerrel alkalmaztunk főnévi csoportok azonosítására és tulajdonnév-felismerésre [1,7]. A rendszer a maximum entrópia módszerrel tanul [8], majd egy-egy mondat legvalószínűbb címkézését rejtett Markov-modellekkel [9], az egyes címkék közötti átmenetvalószínűségek figyelembevételével keresi meg. Az újfajta modell tanítása során változtatás nélkül alkalmaztuk azt a jegykészletet és azon beállításokat, melyek a maximális főnévi csoportok azonosítása során a legsikeresebbnek bizonyultak. Változást a folyamatban csupán az jelentett, hogy a sokszorosára bővült címkekészlet (5 helyett 21 különböző címke) jelentősen növeli mind a tanítás, mind a címkézés idejét.

2.3. Kiértékelés

A tanítást a korpusz 90 százalékán végeztük, a fennmaradó 10 százalékon mértük az eszköz teljesítményét. A rendszer teljesítményét két adat, a pontosság és a fedés jellemzi, a helyesen megtalált frázisok arányát előbbi az összes azonosított frázis arányában, utóbbi a tényleges frázisok arányában mutatja. A szakirodalomban megszokott módon a két érték harmonikus közepeként előálló ún. F-pontszámmal jellemezzük a rendszer általános teljesítményét. A `hunchunk` eredményei az egyes mondattani kategóriákon, valamint összesítve, a 3. táblázatban láthatók. Az `AdjP` kategóriát, mivel a tanulóadatban is nagyon ritkán és szabálytalanul voltak jelen, a címkéző is csak elvétve és látszólag „ok nékül” választotta, ennek hatása azonban elhanyagolható a rendszer összteljesítménye szempontjából.

3. táblázat.

	Pontosság	Fedés	F1
NP	89.36%	88.80%	89.08
ADVP	92.68%	92.99%	92.83
PP	88.70%	88.02%	88.36
ADJP	0.00%	0.00%	0.00
összesen	**90.06%**	**89.68%**	**89.87**

3. Igék

A sekély mondattani elemzés lehetővé teszi, hogy egy-egy mondaton belül azonosítsuk a főbb argumentumokat. Az állítmány azonosításához azonban olyan eszközre is szükségünk lesz, mely felfedezi az elvált igekötőket és a több szóból álló igei komplexumokat. A Szeged Treebank mindkét fajta függőségi viszonyt kódolja, így az elkészült eszközök teljesítményét módunkban áll kiértékelni.

3.1. Igekötők

A Szeged Treebankben található morfológiai elemzésből – csakúgy, mint a `hunmorph` morfológiai elemző kimenetéből – egyértelműen azonosíthatók az önmagukban álló igekötők. Célunk, hogy minél pontosabban tudjuk azonosítani, mely igéhez tartoznak. A kezdeti legegyszerűbb eljárásunk minden igekötőhöz a hozzá a mondatban legközelebb álló igét párosítja; ez a módszer az igekötő-ige párokat csupán 82% körüli F-pontszámmal azonosítja. A pontosságot kis mértékben javítja, ha az igét csak az igekötőhöz legközelebb álló írásjelek között keressük.

A legjelentősebb hibaosztályt az infinitívuszi konstrukciók okozzák (vö. *fel akar mászni*) – ha az infinitívusz mellett álló segédige kiváltja az igekötő elválását, akkor a segédige közelebb kerül az igekötőhöz, mint az infinitívusz alakban álló ige. Kálmán C. és mtsai [10] felsorolják azon segédigéket, melyek leggyakrabban az igekötő és ige közé kerülnek: *akar, bír, fog, kell, kezd, kíván, lehet, mer, óhajt, próbál, szabad, szándékozik, szeret, szokik, talál, tetszik, tud* (pp. 81-82)[1]; jelentős javulást érünk el, ha ezen igéket kizárjuk a keresésből. Célszerű volt továbbá kizárni a létigét, mivel különböző alakjaiban ugyancsak gyakran kerül egy ige és annak igekötője közé (vö. *meg lehet szokni*, *meg van csinálva*). A különböző eljárásokkal elért eredményeket a 4. táblázat összesíti.

4. táblázat. Igekötő-ige párok azonosítása

	Pontosság	Fedés	F1
baseline	82.81%	82.37%	82.59
+írásjelek között	84.41%	82.55%	83.47
+segédige szűrés	97.06%	93.41%	95.20
+létige szűrés	**97.52%**	**95.32%**	**96.41**

A hibák szemrevételezéséből kiderül, hogy azok túlnyomó többségét már a korpusz valamilyen apró hibája okozza. Így például nem járhat sikerrel az eljárás, ha bárhol is téves vagy hiányos az igék és igekötők morfológiai elemzése, vagy éppen a kiértékelés alapjául szolgáló mondattani annotációba csúszik apróbb hiba. Végül a hibaforrás sok esetben a korpuszban szereplő kétféle annotáció

[1] A segédigék beférkőzési hajlandóságáról tett megállapításokat [11] korpuszalapú vizsgálattal is megerősítette.

következetlensége egyes nem egyértelmű esetekben. Pl. az alábbi mondatban: *Vaksötét volt a fenékben, csak tapogatva jutott előre* az *előre* szó morfológiai elemzése szerint igekötő, a szintaktikai annotáció alapján azonban bővítmény. A jelenség fordítottja is előfordul: az *ide figyeljen* mondatban hiába jelez igekötő-ige viszonyt a korpusz, az algoritmusunk nem tudja azonosítani, mivel az *ide* szó a morfológiai elemzés szerint nem igekötő, hanem határozó. Ezen szavak grammatikai státuszának vizsgálata nyilvánvalóan túlmutat jelen cikk határain, az azonban kijelenthető, hogy az általunk eltévesztett párosítások jelentős része olyan szerkezeteket érint, amelyekről a kézi annotátorok sem hoztak következetes döntéseket.

3.2. Komplex igék

A több szóból álló igei szerkezetek egy másik gyakori, ámde könnyen azonosítható típusát adják a már említett, egy finit és egy *-ni* végű igéből álló szerkezetek. Magas pontosság érhető el a fentihez hasonló baseline módszer néhány triviális javításával. A módszer itt is csupán annyi, hogy a morfológia elemzés szerint infinitívuszi jeggyel bíró igéket a hozzájuk legközelebbi finit igéhez kapcsoljuk, nem lépve át közben írásjelet. A módszer pontosságát az 5. táblázat mutatja.

5. táblázat. Infinitívuszok és finit igék párosítása

Pontosság	Fedés	F1
97.02%	**96.35%**	**96.69**

Ez a baseline módszer az infinitívuszok két gyakori előfordulását is rosszul ismeri fel, ezek adják a hibák legnagyobb részét. Egyrészt nem kezeljük két infinitívusz függőségi viszonyát (vö. *Sürgősen igyekeznem kell Almirába jutni*), így a példamondatban a *jutni* szót nem az *igyekeznem* szóval kapcsoljuk össze. Ha azonban csak annyit módosítunk az algoritmuson, hogy nem követeljük meg a választott ige finitségét, akkor a módszer rosszul kezelné az olyan mondatokat, melyben egy finit igéhez több, egymást követő infinitívusz is társul, pl: *A madzagnagyiparos hűlni és zsibbadni kezdett.*

A másik nagy hibaosztályt a koordinált és vesszővel elválasztott infinitívuszok adják. Mivel a fenti eljárást nem egész mondatokon, hanem két írásjel közé eső szósorozatokon végezzük, így ha egy infinitívuszt mégis írásjel választ el a hozzá tartozó finit igétől, akkor ezt a párosítást biztosan nem találjuk meg (vö. *a szakadt ing mögött mégiscsak olyan szív dobog, amelyik tudott szeretni, fájni és aggódni is valamikor.*) Ha azonban általánosságban megengedjük az írásjeleken átívelő függőséget, akkor ez számos téves párosításhoz és így a pontosság jelentős romlásához vezet a fedés kismértékű növekedése mellett.

Mindkét problémára legalább részben megoldást jelentene, ha egy előfeldolgozási lépésben felismernénk a koordinált szerkezeteket. Ez egyúttal újabb hasznos eljárás lenne az alapvető mondatszerkezet feltérképezésére, így remélhetőleg a jövőben erre is sort keríthetünk.

4. Összefoglalás

Cikkünkben három, a magyar mondatok sekély szerkezetének feltérképezését szolgáló eljárást mutattunk be, melyeket a Szeged Treebank korpusz segítségével értékelünk ki. Megmutattuk, hogy a tagmondatok közvetlen összetevőit alkotó maximális frázisok a főnévi csoportokhoz hasonló hatékonysággal azonosíthatóak a felügyelt tanulásra alapuló `hunchunk` eszközzel. A cikk második felében két egyszerű eljárást írtunk le, melyek képesek morfológiailag elemzett szövegből kinyerni az elvált igekötőjű igéket és az ige+infinitívusz szerkezeteket. Mindkét eljárás 96 százalék feletti F-pontszámot ér el. Az igekötők és igék párosításakor a hibák legnagyobb részéért a korpuszban fellelhető ellentmondások felelnek, míg az infinitívuszok esetében a pontosság valószínűleg jelentősen javítható, amennyiben a több egymást követő infinitívuszi alakot tartalmazó mondatok szerkezetéről előzetesen több információt nyernénk ki.

Hivatkozások

1. Recski, G., Varga, D., Zséder, A., Kornai, A.: Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban [Identifying noun phrases in a parallel corpus of English and Hungarian]. VI. Magyar Számítógépes Nyelvészeti Konferencia [6th Hungarian Conference on Computational Linguistics] (2009)
2. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Lecture Notes in Computer Science: Text, Speech and Dialogue. (2005) 123–131
3. Rebrus, P., Vajda, P., Halácsy, P., Rung, A., Trón, V.: Általános célú morfológiai elemző kimeneti formalizmusa [Output formalism of a general-purpose morphological analyzer]. II. Magyar Számítógépes Nyelvészeti Konferencia [6th Hungarian Conference on Computational Linguistics] (2004)
4. Trón, V., Kornai, A., Gyepesi, G., Németh, L., Halácsy, P., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the Workshop on Software, Association for Computational Linguistics (2005) 77–85
5. Uchimoto, K., Ma, Q., Murata, M., Ozaku, H., Isahara, H.: Named entity extraction based on a maximum entropy model and transformation rules. In: ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2000) 326–335
6. Sang, E.F.T.K., Veenstra, J.: Representing text chunks. In: EACL. (1999) 173–179
7. Varga, D., Simon, E.: Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica **16** (2006) 293–301
8. Ratnaparkhi, A., et al.: A maximum entropy model for part-of-speech tagging. In: Proceedings of the conference on empirical methods in natural language processing. Volume 1. (1996) 133–142
9. Rabiner, R.L.: A tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proc. IEEE. Volume 77. (1989) 257–286
10. Kálmán C., G., Kálmán, L., Ádám Nádasdy, Prószéky, G.: A magyar segédigék rendszere. Általános Nyelvészeti Tanulmányok (1989) 49–103
11. Modrián-Horváth, B.: Gesichtspunkte zu einer funktionalen Typologie der Ungarischen Infinitiv regierenden Hilfsverben. Acta Linguistica Hungarica **56**(4) (2009) 405–439