Magyar NP-felismerők összehasonlítása

Miháltz Márton[1]

[1] MTA Nyelvtudományi Intézet, 1068 Budapest, Benczúr u. 33.
mmihaltz@gmail.com

Kivonat

Az előadásban szeretnénk bemutatni egy vizsgálat eredményét, melynek célja a cikk írásakor elérhető magyar nyelvű szintaktikai elemzőprogramok kiértékelése és összehasonlítása. Az elemzést a mondatokban található maximális főnévi csoportok határainak felismerésére korlátoztuk, összehasonlítási alapként a Szeged Treebank 2.0 [1] anyagát használtuk fel. A következő NP-felismerőket vetettük vizsgálat alá:

- MetaMorpho fordítóprogram szintaktikai elemzője [3]
- NooJ [5] magyar NP-nyelvtan
- Hunchunk gépi tanulásos NP-felismerő [4]

A MetaMorpho magyar-angol fordítóprogram forrásnyelvi szintaktikai elemző komponense kézzel írt szabályokkal működő jegystruktúrás környezetfüggetlen nyelvtant használ. A Nyelvtudományi Intézetben fejlesztett NP-nyelvtan a NooJ keretrendszerben készült véges állapotú automaták kaszkádja. A lexikai (morfológiai) elemzési szinthez több különböző megoldással is teszteltük. A Hunchunk rendszer a Szeged Treebanken tanított, maximum entrópiás Markov-modell NP-felismeréshez.

A Szeged Treebank 6 különböző témakörből (szépirodalom, iskolai fogalmazások, újságcikkek, számítástechnikai szövegek, jogi szövegek, gazdasági és pénzügyi rövidhírek) 1,2 millió szövegszót tartalmaz 82 ezer mondatban, részletes morfológiai és szintaktikai annotációval. A vizsgálathoz egyesítettük a mondatok halmazát, majd az ismétlődéseket kiszűrve 80,877 különböző mondathoz jutottunk. Minden mondatot külön, az eredeti szövegkörnyezete nélkül elemeztünk a vizsgált elemzőprogramokkal, a többször szereplő mondatokhoz az első előfordulásukhoz megadott annotációt használtuk fel (anélkül, hogy megvizsgáltuk volna, hogy a különböző előfordulások elemzései különböznek-e egymástól.)

A kiértékelés során minden mondatban megvizsgáltuk, hogy az egyes elemzők által megadott maximális NP-k közül hány szerepelt a treebankben (pontosság), illetve a treebank maximális NP-i közül hány található az elemző kimenetében (fedés), valamint megadtuk a két érték szokásos kombinációját is (F1-mérték). Egyezésnek csupán a teljesen megegyező kezdő- és záró terminálissal rendelkező NP-ket fogadtuk el, a részleges egyezéseket ebben a vizsgálatban ugyanúgy hibaként kezeltük, mint a teljesen rossz találatokat. A méréseket minden elemzővel elvégeztük külön-külön a 6 korpusz-témakör, illetve a 15 különböző forrás mindegyikére is.

Az 1. táblázatban közöljük a NooJ keretrendszerben írt szintaktikai elemző két különböző morfológiai elemzőt használó változatának összehasonlítását. Az 1. változat a Magyar Nemzeti Szövegtár [7] és a morphdb.hu [6] anyaga alapján készült morfo-

lógiai lexikont használja, míg a 2. változat egy, a NooJ rendszerben kézzel írt morfológiai elemző automatát. A 2. táblázatban a MetaMorpho és a NooJ elemző MNSZ-morphdb.hu-s változatának összehasonlítása látható.

1. táblázat: A NooJ elemző két változatának összehasonlítása a teljes treebank anyagán.

	NooJ 1.			NooJ 2.		
Témakör	P	R	F	P	R	F
Iskolai	43.61%	68.31%	53.23%	47.09%	67.52%	55.48%
Szám.tech.	34.19%	52.25%	41.34%	27.86%	43.18%	33.87%
Gazdasági	28.85%	48.80%	36.26%	23.92%	41.32%	30.30%
Szépirodalom	45.93%	68.19%	54.89%	43.87%	62.52%	51.56%
Hírek	35.16%	56.19%	43.25%	31.83%	50.43%	39.03%
Jogi	28.20%	51.34%	36.40%	22.58%	45.82%	30.25%
Teljes korpusz:	36.51%	58.72%	45.02%	33.34%	53.47%	41.07%

2. táblázat: A MetaMorpho és a NooJ elemzők összehasonlítása a teljes treebank anyagán.

	MetaMorpho			NooJ 1.		
Témakör	P	R	F	P	R	F
Iskolai	65.50%	71.92%	68.56%	43.61%	68.31%	53.23%
Szám.tech.	46.45%	56.72%	51.07%	34.19%	52.25%	41.34%
Gazdasági	43.78%	53.59%	48.19%	28.85%	48.80%	36.26%
Szépirodalom	63.91%	67.27%	65.55%	45.93%	68.19%	54.89%
Hírek	53.03%	58.43%	55.60%	35.16%	56.19%	43.25%
Jogi	35.21%	45.37%	39.65%	28.20%	51.34%	36.40%
Teljes korpusz:	52.14%	60.25%	55.90%	36.51%	58.72%	45.02%

A 3. táblázat a Hunchunk NP-felismerő és a másik két rendszer összehasonlítását foglalja össze. Mivel a Hunchunk rendszert a Szeged Treebank mondatainak egy részén tanították be, ehhez az összehasonlításhoz nem a teljes korpuszt, csak a tanításhoz fel nem használt, a szerzők által a kiértékelésre elkülönített 16.989 mondatot használtuk fel. Ezek közül kihagytunk 142 ismétlődő mondatot, illetve 494 mondatot a Hunchunk kimenetéből technikai okok miatt nem tudtunk az eredeti korpuszban azonosítani, így az összehasonlítás a maradék 16.353 mondat segítségével történt.

3. táblázat: A Hunchunk, a MetaMorpho és a NooJ elemzők összehasonlítása a treebank kiértékelésre elkülönített részén.

HunChunk			MetaMorpho			NooJ 1.		
P	R	F	P	R	F	P	R	F
78.67%	84.99%	81.71%	54.39%	61.52%	57.73%	37.57%	59.28%	45.99%

A NooJ elemző két változatának összehasonlításából egyértelműen kitűnik, hogy az MNSZ-morphdb.hu morfológiai anyagát használó változat teljesít jobban (1. táblá-

zat). A MetaMorpho elemző ennél a változatnál szignifikánsan jobban teljesít (2. táblázat). A Treebank szempontjából további érdekesség, hogy mindkét rendszer az iskolai fogalmazások és a szépirodalmi alkotások szövegein teljesít a legjobban és a jogi szövegeken a legrosszabbul.

A gépi tanulásos rendszer kiértékelő halmazán végzett mérések (3. táblázat) ugyanezt a sorrendet mutatják a két szabályalapú rendszer között, az élre viszont a Hunchunk rendszer kerül szignifikáns előnnyel. Mindenképpen szükséges azonban megemlíteni, hogy a gépi tanulásos rendszer teljesítménye szempontjából az alkalmazott technológián túl nem elhanyagolható szempont, hogy ez a rendszer a Szeged Treebank – a kiértékelő halmaz mondataihoz hasonló – mondatain tanulva a kiértékelő korpusz inherens sajátosságaira jobban rá volt hangolódva, mint a másik két, a korpusz anyagától függetlenül fejlesztett rendszer.

A bemutatott NP-felismerők kiértékelésében további lehetséges munka, ha a korrektebb összehasonlítás érdekében az elemzők és a Treebank különböző koncepciókkal készült nyelvtanai között megtaláljuk a legnagyobb közös részhalmazt, és az ezzel megadható elemzésekre redukálva ismételjük meg a kiértékelést. Néhány példa ilyen nyelvtani különbségekre: a MetaMorphoban a főnévi igeneves szerkezetek NP-knek számítanak, a Szeged Treebankben nem; a névutók a MetaMorphoban részei az NP-knek, a Treebankben nem; az olyan birtokos szerkezetek, ahol a birtok közvetlenül követi a birtokot, a Treebankben két NP-nek számítanak, a MetaMorpho és a NooJ nyelvtanában viszont van a kettőt egyesítő NP; a MetaMorphoban a főnévi fejhez kapcsolódó vonatkozó mellékmondat része a maximális NP-nek, a Treebankben nem stb. A részleges találatok súlyozott figyelembevétele és a hibatípusok vizsgálata szintén további lehetőségek.

Bibliográfia

1. Csendes D., Alexin Z., Csirik J., Kocsor A.: A Szeged Korpusz és Treebank verzióinak története. III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005) kiadványa, Szeged, december 8-9. (2005) 409–412
2. Oravecz, Cs., Dienes, P.: Efficient Stochastic Part-of-Speech tagging for Hungarian. In: Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas (2002) 710–717
3. Prószéky, G., Tihanyi; L., Ugray, G.: Moose: a robust high-performance parser and generator. In: Proceedings of the 9th Workshop of the European Association for Machine Translation, Foundation for International Studies. La Valletta, Malta (2004) 138–142
4. Recski G., Varga A., Zséder A., Kornai A.: Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban. In: VI. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2009)
5. Silberztein, M.: NooJ : an Object-Oriented Approach. In: Muller, C., Royauté, J., Silberztein M. (Eds): INTEX pour la Linguistique et le Traitement Automatique des Langues, Cahiers de la MSH. Presses Universitaires de Franche-Comté, Ledoux (2004) 359–369
6. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Simon, E., Vajda, P.: morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In: III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2005)
7. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation. Las Palmas (2002) 385–389