A Szeged Korpusz és Treebank verzióinak története

Csendes Dóra[1], Alexin Zoltán[1], Csirik János[1], Kocsor András[2]

[1] Szegedi Tudományegyetem Informatikai Tanszékcsoport
6720 Szeged, Árpád tér
{dcsendes, alexin, csirik}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Aradi vértanúk tere 1.
{kocsor}@inf.u-szeged.hu

1 Bevezetés

A Szegedi Tudományegyetem Informatikai Tanszékcsoportján 1998 óta folytatott természetesnyelvi kutatások és fejlesztések egyik fő célja egy nagyméretű, kézileg annotált szöveges adatbázis kialakítása volt. Tettük ezt azért, hogy további számítógépes nyelvészeti kutatásokhoz jó minőségű alapot biztosítsunk, ill. számítógépes tanuló algoritmusok számára egy nagy megbízhatóságú adatbázist hozzunk létre. A munkálatok eredményeként mára a szegedi szövegállomány négy különböző verziója készült el Szeged Korpusz 1.0, Szeged Korpusz 2.0, Szeged Treebank 1.0, ill. Szeged Treebank 2.0 néven. Az összes verzió kialakításánál egy automatikus előelemzési fázist egy részletes kézi ellenőrzés és javítás követett. Az alábbiakban röviden beszámolunk a szövegállomány fejlesztési munkálatairól. A négy verzióban szereplő fájlok XML formátumúak, belső szerkezetüket a TEIXLite, ill. a TEI P4 DTD séma írja le.

2 Rövid történeti áttekintés

2.1 Szeged Korpusz 1.0

A Szeged Korpusz [1] szövegeinek gyűjtése 1999-ben kezdődött meg. A vállalkozásra a MUTEXT-EAST projekt [4] ösztönözte az akkori konzorciumot[79], amelynek keretében megszületett a TELRI korpusz magyar változata, és megtörténtek az első kísérletek a szövegek automatikus elemzésére. Ennek továbbfejlesztésére vettük célba egy bővebb szövegállomány összeállítását, hogy a további számítógépes nyelvészeti kísérleteket már egy reprezentatívabb korpuszon végezhessük. A szövegek kiválasztása során a legfőbb szempont az volt, hogy tematikailag a lehető legkülönfélébbek legyenek. Végül öt különböző témakörből választottuk ki a szövegeket, nevezetesen: szépirodalmi írásokból, 14-16 éves tanulók fogalmazásaiból, napilapokban és folyóiratokban megjelent újságcikkekből, jogi szövegekből, és számítástechnikai szövegekből, témakörönként kb. 200 ezer szó terjedelemben. Az így összegyűjtött korpusz 1 millió szövegszót és további 200 ezer írásjelet tartalmaz. Természetesen ez a mennyiség nem elegendő ahhoz, hogy a teljes mai magyar nyelv szókészletét és nyelvtani

[79] Szegedi Tudományegyetem Informatikai Tanszékcsoport, MorphoLogic Kft.

struktúráját lefedje, de elég reprezentatívnak bizonyul abból a szempontból, hogy számítógépes nyelvészeti kutatásokat lehessen rá alapozni.

A Szeged Korpusz első verziója egy morfo-szintaktikailag elemzett és kézzel egyértelműsített természetesnyelvi szöveges adatbázis, amely 139.000 különböző szóalakot tartalmaz. A szövegek morfo-szintaktikai annotálásához a nemzetközileg elfogadott MSD (Morpho-Syntactic Description) kódrendszert használtuk. A korpusz jelen verziója a többjelentésű szavak esetében csak a kiválasztott morfo-szintaktikai kódokat tartalmazza, a lehetségeseket nem tünteti fel.

2.2 Szeged Korpusz 2.0

A Szeged Korpusz 2.0 verziója az 1.0 verzió kibővítésével keletkezett. A meglévő 1 millió szavas szövegállományt az akkori konzorcium[80] egy 200 ezer szavas rövidhír részkorpusszal bővítette ki, amely elsősorban gazdasági és pénzügyi rövidhíreket tartalmaz. Így a korpusz 1,2 millió szövegszavasra nőtt, amely 155.500 különböző szóalakot tartalmaz, és további 250 ezer írásjelet is magában foglal. A korpusz második verziója az elsőhöz hasonlóan egy morfo-szintaktikailag elemzett és kézzel egyértelműsített természetesnyelvi szöveges adatbázis. A méretnövekedésen kívül az első verziótól abban tér el, hogy a kontextusnak megfelelően kiválasztott morfo-szintaktikai kódok mellett a lehetséges kódok is szerepelnek az adatbázisban, így hatékonyan alkalmazható automatikus szófaji annotáló módszerek tesztelésére.

2.3 Szeged Treebank 1.0

A természetesnyelvi feldolgozás fontos lépése a szintaktikai elemzés és annotálás, azaz a különböző szintaktikai egységek, pl. főnévi vagy melléknévi csoportok, névutós szerkezetek bejelölése. Mivel a mondatok többségében az egész mondat jelentése szempontjából a főnévi csoportok (NP-k) kulcsfontosságú szerepet játszanak, ezért a Szeged Treebank 1.0 [2] verziójában ezeknek a szerkezeteknek a bejelölése volt az elsődleges cél. Ezen kívül, ugyancsak a mondatok tartalmának értelmezhetősége szempontjából, fontos szerepe van a tagmondatok (CP-k) elkülönülésének és egymáshoz való viszonyának (alárendelés, mellérendelés), ezért ezeket is jelöltük a szövegeken. A főnévi csoportok és tagmondatok bejelölését a Szeged Korpusz 2.0 állományán, 82.000 mondaton végeztük.

Számos olyan alkalmazásról tudunk, ahol elegendő a szövegek részleges szintaktikai elemzése (shallow parsing). Ilyen pl. az automatikus információkinyerés (information extraction) vagy kivonatolás (text summarisation) is. Az itt leírt Szeged Treebank 1.0 verzió ilyen alkalmazásokban került felhasználásra, ill. további elemzéshez szolgál kiindulópontként.

2.4 Szeged Treebank 2.0

A Szeged Treebank 2.0 [3] az első verziónál gazdagabb elemzést és annotációt tartalmaz. Jelen verzió magában foglalja az összes előző verzió eredményeit (morfo-szintaktikai, NP- és CP-annotálást), és ezt kiegészíti további szintaktikai elemzéssel, amely a melléknévi, határozószói csoportok, névutós szerkezetek, igék, stb. bejelölését foglalja magában. A treebank kialakításakor a már ismert forrásmunkákra és meglévő elméletekre támaszkodtunk. Ezek tanulmányozásával és összevetésével nyelvész

[80] SZTE Informatikai Tanszékcsoport, MorphoLogic Kft, MTA Nyelvtudományi Intézete

szakértőink egy konzisztens szintaktikai szabályrendszert dolgoztak ki a generatív szintaxis szabályainak megfelelően. A használt szintaktikai címkék a nemzetközi szabványnak megfelelőek, és lehetővé teszik az adott szintaktikai szerkezetre vonatkozó attribútumok tárolását is. A Szeged Treebank 2.0-ra vonatkozó statisztikai adatokat az alábbi két összefoglaló táblázat mutatja.

1. Táblázat: A szintaxisfák mélység szerinti eloszlása a treebank mondataiban

Szintaxisfa-mélység	**1**	**2**	**3**	**4**	**5**	**6-7**	**8-10**	**11-20**
Fogalmazások	141	2922	7898	8388	3942	1380	62	0
Jogi szövegek	2	110	687	1554	2127	3346	1337	115
Újságcikkek	29	577	1466	2469	2545	2567	534	24
Üzleti hírek	0	75	864	2396	2844	2933	455	10
Szépirodalom	493	4649	5230	4170	2373	1495	152	2
Számítástechnika	9	541	1133	2413	2654	2638	373	7
Összes	**674**	**8874**	**17278**	**21390**	**16485**	**14359**	**2913**	**158**

2. Táblázat: A szintaxisfák szélesség szerinti eloszlása a treebank mondataiban

Szintaxisfa-szélesség	**1**	**2**	**3**	**4**	**5**	**6-7**	**8-10**	**11-20**	**21-50**	**50-**
Fogalmazások	25	126	319	578	1109	2811	4738	11309	3667	51
Jogi szövegek	20	56	60	72	48	147	429	2640	5153	653
Újságcikkek	1	83	97	120	156	438	1000	3693	4401	222
Üzleti hírek	1	0	2	11	158	114	502	3741	5006	42
Szépirodalom	15	434	1099	1336	1397	2691	3095	5487	2864	146
Számítástechnika	104	142	108	80	130	266	681	3643	4430	184
Összes	**166**	**841**	**1685**	**2197**	**2998**	**6467**	**10445**	**30513**	**25521**	**1298**

3 Gépi tanulási alkalmazások

A részletes kézi annotálásnak köszönhetően a Szeged Korpusz és Szeged Treebank különböző verziói megbízható tanulási és tesztelési adatbázisként szolgálnak számítógépes tanulóalgoritmusok számára. A Nyelvtechnológiai csoport a következő területeken kísérletezett tanuló algoritmusok alkalmazásával:

- morfo-szintaktikai elemzés és egyértelműsítés,
- részleges és teljes szintaktikai elemzés,
- információkinyerés részleges szemantikai információk (szemantikai keretek) segítségével.

3.1 Morfo-szintaktikai egyértelműsítés (POS tagging)

Nagy sikerrel alkalmaztunk különböző tanulóalgoritmusokat és azok kombinációját a szófaji elemzés és egyértelműsítés területén [6]. A magas találati arányok (97% feletti találati pontosság) figyelemre méltó, hiszen a magyar nyelv gazdag és változatos morfológiája meglehetősen nagy kihívást jelent az automatizált módszerek számára.

Ezen kívül azért is jelentősek az elért eredmények, mert a felhasznált morfo-szintaktikai kódrendszer (MSD) nagyon részletes, így a többértelmű szavak aránya eléri az 50%-ot.

3.2 Szintaktikai elemzés

A Szeged Treebank első verzióját numerikus, szabály-alapú és statisztikai tanulóalgoritmusok, ill ezek kombinációjából összeálló módszerek tanítására használtuk [5]. A cél az volt, hogy automatikus információkinyerés támogatása céljából a tanulóalgoritmusok minél pontosabban tudják beazonosítani a mondatokban szereplő főnévi csoportokat, ill. minél jobban el tudják találni a tagmondatok határait. A legjobb pontossági eredmények 85-92% között vannak, de az algoritmusok teljesítménye nagyban függ a feldolgozott szöveg típusától (pl. jogi vs. újságnyelvi szöveg) és a mondatszerkezetek bonyolultságától.

Folyó kutatásaink egy automatikus szintaktikai elemző kifejlesztésére irányulnak a Szeged Treebank 2.0 verziója alapján, vagyis részletes szintaktikai annotálás felhasználásával. Már a legkorábbi eredmények elérték a 80-85% körüli találati pontosságot, amit bíztató előjelnek tekintünk a további fejlesztések szempontjából. Egy teljes szintaktikai szerkezeteket felismerni és elemezni képes program nemcsak az automatikus információkinyerést segítené nagyban, hanem gépi fordítórendszerek is jól hasznosíthatják.

Bibliográfia

1. Alexin Z., Csirik J., Gyimóthy T., Bibok K., Hatvani Cs., Prószéky G., Tihanyi L.: *Manually Annotated Hungarian Corpus*, in Proc. of the Research Note Sessions of EACL'03, Budapest, Hungary, pp. 53-56 (2003)
2. Csendes, D., Csirik, J., Gyimóthy, T.: *The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus* in Proc. of TSD 2004, Brno, Czech Republic and LNAI vol. 3206, pp. 41-49 (2004)
3. Csendes D., Csirik J., Gyimóthy T., Kocsor A.: *The Szeged Treebank*, in Proc. of TSD 2005, Karlovy Vary, Czech Republic and LNAI vol. 3658, pp. 123-132 (2005)
4. Erjavec, T., Monachini, M.: *Specification and Notation for Lexicon Encoding*, Copernicus Project 106 „MULTEX-EAST", Work Package 1 – Task 1.1, Deliverable D1.1F (1997)
5. Hócza, A., Felföldi, L., Kocsor, A.: *Learning Syntactic Patterns Using Boosting and Other Classifier Combination Schemas* in Proc. of TSD 2005, Karlovy Vary, Czech Republic and LNAI vol. 3658, pp. 69–76 (2005)
6. Kuba, A., Csirik, J., Hócza, A.: *POS tagging of Hungarian with combined statistical and rule-based methods* in Proceedings of TSD 2004, Brno, Czech Republic and LNAI vol. 3206 (2004)