Szintaktikai elemzők eredményeinek összehasonlítása

Hócza András[1], Kovács Kornél[2], Kocsor András[2]

[1] Szegedi Tudományegyetem, Informatika Tanszék
6720 Szeged, Árpád tér 2.
hocza@inf.u-szeged.hu

[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Aradi védtanuk tere 1.
{kkornel, kocsor}@inf.u-szeged.hu

Kivonat: A mondatok szintaktikai elemzése alapvető részét képezi további természetesnyelvi feladatoknak. Ezért fontos feladat, hogy a magyar nyelvre készült szintaktikai elemzők részére is kidolgozzunk egy olyan hátteret, amelyet az angol nyelv esetén a Penn Treebank biztosít. A dolgozat beszámol a Szeged Treebank adattárára épülő adatbázis kialakításáról, amely magyar nyelvű szintaktikai elemzőkhöz készült modellépítés és az egységes összehasonlíthatóság érdekében. Ezen az adatbázison alkalmaztunk néhány rendelkezésre álló módszert, hogy összehasonlítási alapot teremtsünk a későbbi megközelítésekhez, valamint az egyik módszert a Penn Treebank angol szövegeire is alkalmaztuk, hogy képet kapjunk szintaxiselemzési szempontból a két nyelv összetettségéről.

Kulcsszavak: teljes szintaxis, gépi tanulás, szabály alapú módszerek

1 Bevezetés

Egy mondat teljes szintaxisának felismerése olyan folyamat, amely során meg kell határozni, hogy milyen egymás után következő szavak csoportosíthatók egybe, mint például főnévi, melléknévi, igei szerkezetek. Ezek a szócsoportok egymásba ágyazottak, fastruktúrát alkotnak. Egy mondat teljes szintaxisa olyan összefüggő fa, melynek levelei a mondat szavai, illetve írásjelei. A mondat szintaxisának feltárása számos természetesnyelvi feladathoz szolgáltathat alapvetően fontos információkat. Ilyen terület a feltárt mondatszintaxis felhasználására például az adatbányászat, információkinyerés, gépi fordítás. Ezért fontos kifejleszteni egy tetszőleges magyar szövegen jó hatásfokkal működő automatikus szintaktikai elemzőt.

Az angol nyelv szintaktikai elemzésének meglehetősen nagy szakirodalma van, mivel lényegesen korábban kialakítottak szintaktikailag annotált szöveges adatbázisokat. Ilyen például a Penn Treebank (Marcus et al., 1993), amelyen a tudományos cikkekben mérni szokták az elemzési módszerek pontosságát. Magyar nyelvre a Szeged Treebank (Csendes et al., 2005) munkálatai nemrég fejeződtek be, ami után megnyílt a lehetőség gépi tanulási technikákon alapuló magyar szintaktikai elemzők kifejlesztésére és tesztelésére.

A korábbi években már készültek magyar nyelvre szintaktikai elemzők és most, hogy lehetőség adódott rá, fontos feladat ezek megbízhatóságának a feltérképezése és összehasonlítása egymással, továbbá más, a szakirodalomban leírt módszerekkel. Azonban két módszer összehasonlítása csak teljesen azonos feltételek mellett, azonos szövegekre alkalmazva ad pontos képet a módszerek hatékonyságáról. Ezért dolgozatunkban beszámolunk arról, hogyan próbáljuk megteremteni az egységes összehasonlíthatóság feltételeit egy Szeged Treebank adattárára épülő adatbázis kialakításával. Ezen az adatbázison alkalmaztunk néhány rendelkezésre álló módszert, hogy meglegyen a kiinduló összehasonlítási alap a jövőben készülő hatékonyabb magyar szintaktikai elemzők számára. Az is érdekes kérdés, hogy az angol és a magyar nyelv szintaktikai elemzése mennyire összetett feladat, mik a különbségek, és mik a hasonlóságok. Ezért egy magyarra kifejlesztett elemzőt kipróbáltunk a Penn Treebank adataiból vett angol szövegeken is.

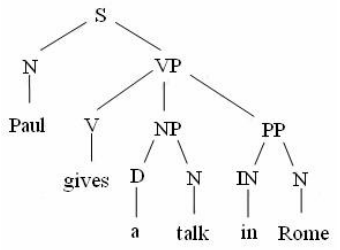
A dolgozat a következő módon épül fel: a 2. rész általánosan mutatja be a mondatszintaxis kutatását angol és magyar nyelvre, a 3. részben a magyar nyelvű szintaktikai elemzők egységes összehasonlításához készült adatbázisról lesz szó, a 4. rész a Szeged Treebank magyar szövegeinek adatbázisán elért eredményeket írja le, valamint az a Penn Treebank angol szövegin végzett próbát, és végül a 5. rész összefoglalja az elért eredményeket.

2 A mondatszintaxis felismerése

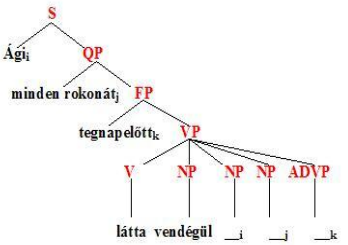
A mondatszintaxis felismerésének célja, hogy egy mondat különféle szócsoportokból álló elemzési fája automatikus módszerrel, minél pontosabban előálljon. A mondatok szintaxisának felismerése, valamint az eredmények mérése és az alkalmazott módszerek összehasonlítása számos problémát vet fel. A feladat nehézsége és az alkalmazott módszerek megvalósítása nagyban függ attól, hogy milyen nyelvről van szó.

2.1 A magyar nyelv szintaktikai elemzésének nehézségei

A magyar nyelv számos olyan nyelvi sajátossággal rendelkezik, ami megnehezíti a szintaxisfelismerést az indoeurópai nyelvekhez (pl. angolhoz) képest. Az egyik jelentős különbség a viszonylag szabad szórend (1. ábra), azaz egy mondat szintaktikai egységei többféleképpen átrendezhetők úgy, hogy a kapott mondatok nyelvtanilag szintén szabályosak lesznek. Azonban az így kapott mondatok jelentései módosulhatnak a kiinduló mondatéhoz képest. A mondatrészi szerepet a magyar nyelv ragozással és névutók alkalmazásával oldja meg. Ebből adódik a másik probléma, a nagyfokú morfológiai változatosság. Az említett sajátosságok összességében jelentősen megnövelik a lehetséges minták, nyelvi sémák számát, melyek rontják a statisztikai alapú gépi tanulás hatékonyságát.



Paul gives a talk in Rome.
– (nincs más lehetséges szórend)



Ági minden rokonát vendégül látta tegnapelőtt.
Minden rokonát vendégül látta Ági tegnapelőtt.

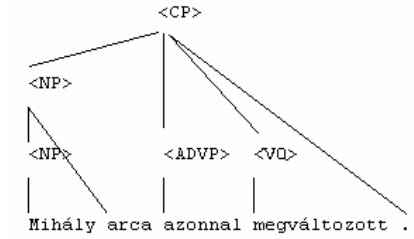
1. ábra: A magyarban a viszonylag szabad szórend miatt egy adott VP sokféle elrendezésben előfordulhat, sőt nem is mindig alkot összefüggő szerkezetet.

2.2 A szintaktikai elemzők pontosságának mérése

Az eredményeket az összehasonlíthatóság érdekében közös mérőszámokkal kell jellemezni. Az elemzett mondatokhoz rendelkezésre kell állnia egy nyelvész szakértő által készített kézi elemzésnek, ami etalonként szolgál. Minél jobban hasonlít az automatikusan előállított elemzési fa az etalonhoz, annál precízebb eredményről beszélhetünk. Fák hasonlóságának a vizsgálata úgy történik, hogy az összehasonlító algoritmus kigyűjti az összes előforduló szócsoportot (2. ábra) mindkét fából és ezeket veti össze a következő, az irodalomban gyakran alkalmazott képletek szerint:

- ***Pontosság:*** a helyesen felismert szócsoportok száma / az összes felismert szócsoportok száma.
- ***Fedés:*** a helyesen felismert szócsoportok száma / az etalonban ténylegesen szereplő szócsoportok száma.
- ***Középarány*** ($F_{\beta=1}$): 2* Pontosság* Fedés/(Pontosság+ Fedés)

Szintaxisfa:



XML leírás:

```
        <CP>
                <NP>
                        <NP>
1:                              Mihály
                        </NP>
2:                      arca
                </NP>
                <ADVP>
3:                      azonnal
                </ADVP>
                <V0>
4:                      megváltozott
                </V0>
5:              .
        </CP>
```

A szintaxisfából képzett szócsoport lista:
NP(1-1), NP(1-2), ADVP(3-3), V0(4-4), CP(1-5)

2. ábra: Példa arra, hogyan lehet egy egyszerű szintaxisfából szócsoportlistát képezni. A listaelemek egyértelműen azonosítják a szócsoportokat, ezért lehet a szócsoportlistákat fák hasonlóságának vizsgálatához felhasználni.

2.3 Az angol nyelv szintaktikai elemzésének szakirodalma

Az angol nyelvre számos eredmény létezik a mondatszintaxis felismerésének témakörében. A Penn Treebank (Marcus et al., 1993) annotált szövegeiben elkülönítettek egy részt, amely a megjelenése óta összehasonlítási alapot képez a témához kapcsolódó publikációk eredményeihez. Ezután többféle módszert alkalmaztak, hogy még jobb eredményt érjenek el, ezekből néhányat az 1. táblázat foglal össze.

Az első publikációban (Abney, 1991) nyelvtani kódok alapján osztályozta a szavakat, hogy azok kezdő, vég- vagy belső elemei-e egy adott típusú frázisnak. (Ramshaw és Marcus, 1995) transzformáción alapuló tanulást valósított meg. (Argamon, 1998) egyszerre végezte főnévi és igei szerkezetek felismerését. (Tjong Kim Sang és Veenstra, 1999) bevezette a több fokozatban (kaszkád) alkalmazott felismerést. A legújabb módszerek úgy érnek el javulást az eredményekben, hogy több módszert összekombinálva, szavazással hozzák meg a döntéseket, (Tjong Kim Sang, 2000) öt különféle módszert kombinált össze.

Hivatkozás	$F_{\beta=1}$	Teszt adatbázis	Módszer
Abney, 1991	-	-	Chunking
Ramshaw el al., 1995	92.0	Penn Treebank	Transzformációs tanulás
Argamon, 1998	91.6	Penn Treebank	NP- és VP-szerkezetek
Tjong K. S. et al., 1999	92.37	Penn Treebank	Többszintű nyelvtan
Tjong K. S., 2000	93.26	Penn Treebank	Több módszer kombinációja

1. táblázat: Néhány fontosabb angol nyelvre elért eredmény, 1993 óta az összehasonlíthatóság érdekében a Penn Treebank adatain történik a tesztelés.

2.4 A magyar nyelv szintaktikai elemzésének szakirodalma

A magyar nyelvre ez idáig lényegesen kevesebb szintaxiselemző készült. Az előzőekben vázolt nehézségek miatt szinte lehetetlen olyan nyelvész szakértők által kézzel készített szabályrendszert megalkotni, ami megfelelő hatékonyságú, és minden lehetséges esetre kiterjed. A másik probléma, hogy idáig nem volt elegendő mennyiségű annotált magyar szöveget tartalmazó korpusz, ami a gépi módszerek alkalmazását lehetővé tette volna.

A MorphoLogic Kft. által kifejlesztett HumorESK mondatelemző (Kis, 2003) 1995 óta folyamatosan fejlődik. Ez idő alatt különféle nyelvészeti területeken alkalmazták. Fő jellemzője, hogy a szimbólumokhoz jegyszerkezeteket (*feature structure*) kapcsol, és elemzési erdőt épít az egyes jegyek örököltetésével. Az elemzőben használt nyelvtan nyelvész szakértők közreműködésével állt elő. A Nyelvtudományi Intézet beszámol egy készülő szintaktikai elemzőről (Váradi, 2003), ami főnévi szerkezeteket ismer fel reguláris kifejezésekkel leírt, többfokozatú (kaszkád) szakértői szabályrendszer alkalmazásával. A Szegedi Egyetemen a nyelvtan előállítása gépi tanulási módszerekkel történt. A főnévi szerkezetek (Hócza, 2004), valamint a teljes szintaxis (Hócza, 2005) felismerésére készült módszerek modelljének forrását a Szeged Treebank annotált szövegei adták. Az eredmények összefoglalása a 2. táblázatban található:

Hivatkozás	$F_{\beta=1}$	Teszt adatbázis	Módszer
HumorESK (Kis, 2003)	-	-	Szakértői szabályok
Váradi, 2003	58.78	100 annotált mondat	Szakértői szabályok
Hócza, 2004	83.11	Üzleti hírek	NP-tanulás
Hócza, 2005	78.59	Szeged Treebank	Rövid faminták tanulása

2. táblázat: A magyar nyelvre elért eddigi eredmények

3 Szintaktikai elemzők összehasonlítása

Két módszer hiteles összehasonlítása megkívánja, hogy ugyanazokon az adatokon alkalmazzuk őket. A magyar nyelvre eddig még nem volt kialakítva olyan nyilvános adatbázis, mint az angol esetén a Penn Treebank. Cikkünk fő célja beszámolni arról, hogy hogyan próbáljuk megteremteni a hiteles összehasonlíthatóság lehetőségét a magyar szintaktikai elemzőkre. Ennek érdekében a következő módszert dolgoztuk ki:

- Kifejlesztettünk magyar nyelvre egy teljes szintaktikai elemzőt (Hócza, 2005), és alkalmaztuk a Szeged Treebank adatain.
- A Szeged Treebank adataiból kialakítottunk egy mintaadatbázist, amely a jövőben lehetőséget teremt minden érdekelt számára, hogy a treebank adatait felhasználva kipróbálja a szintaktikai elemzőjét, és összevesse annak hatékonyságát a magyar nyelvre alkalmazott más módszerekével.
- Néhány további módszert is kipróbáltunk a mintaadatbázison (3. táblázat).
- A rendelkezésre álló módszereket alkalmaztuk a Penn Treebank adatain is a hatékonyság lemérésére, valamint, hogy összehasonlítsuk az angol és a magyar nyelv összetettségét a szintaxis elemzés feladatának szempontjából.

3.1 A Szeged Treebank szövegeiből kialakított adatbázis

Példák véletlenszerű kiválasztása még nem garantálja azt, hogy egy módszer tesztelésénél ne kapjunk torz eredményeket. Például ha véletlenül egymáshoz nagyon hasonló példák kerülnek a tesztadatokba, a kapott végeredmény pontossága több százalékkal is eltérhet egy másik felosztással kapott értéktől. A nemzetközi szakirodalomban az eredmények közlésénél ennek a problémának az elkerülésére alkalmazzák a *tenfold cross-validation* módszert, melynek lépései a következők:

1. A példákat véletlenszerűen szétosztjuk 10 egyforma méretű csoportba.
2. Előállítjuk a tréningállományokat úgy, hogy mindegyik 9 különböző csoportból adódjon össze (mindig 1 kimarad). A különböző lehetőségek száma 10.
3. A tréningállományokhoz hozzárendeljük azok tesztpárját; ez az a csoport lesz, ami kimaradt a tréningből.
4. Lefuttatjuk a tanulást a 10 tréningcsoportra, majd elvégezzük a tesztelést a tréningállomány megfelelő teszt párján.
5. A teszteredményeket összesítjük és átlagoljuk; ez az átlag lesz a példákon végzett tanulás végső eredménye.

A Szeged Treebank adataiból ezt a 10 részre történő felosztást végeztük el ajánlásként a jövőben kipróbálandó módszerek egységes felkészítéséhez és összehasonlításához.

4 Eredmények

Ebben a részben arról számolunk be, hogy a számunkra rendelkezésre álló módszerekkel milyen eredményeket értünk el magyar és angol szövegeken.

4.1 Néhány módszer alkalmazása a magyar nyelvű adatbázison

A Szeged Treebank mintaadatbázisán 4 módszert alkalmaztunk a teljes szintaxis felismerésére. Az első, alapnak (baseline) tekinthető módszer a treebank adataiból kigyűjtött környezetfüggetlen valószínűségi nyelvtant (PCFG) használ, amely chart parsing segítségével építi fel a teljes szintaxisfát. A második módszer esetén a chart parser nyelvtanát kb. 22 ezer szabály képezte, amelyet nyelvész szakértők állítottak elő. A harmadik egy memória alapú módszer volt, amely a treebank adataiból kigyűjtött teljes mondatok szintaxisfáját illesztette a nyelvtani kódok figyelembevételével. A negyedik módszerben a chart parser gépi tanulással előállított famintákat alkalmazott.

Módszer	$F_{\beta=1}$	Teszt adatbázis	Megjegyzés
Baseline	56.01	Szeged Treebank	PCFG, chart parsing
Szakértői szabályok	55.58	Szeged Treebank	~22k szabály, chart parsing
Mondat memória	7.04	Szeged Treebank	Hasonló mondat keresése
Faminták	75.47	Szeged Treebank	Részfák illesztése, chart p.

3. táblázat: A magyar nyelv teljes szintaxiselemzésében a különféle módszerekkel elért eredmények a Szeged Treebank mintaadatbázisán modellt építve és tesztelve.

4.2 A famintákon alapuló módszer alkalmazása angol nyelvre

A magyar nyelvre alkalmazott módszerek közül a faminták tanulásán alapuló módszer érte el a legjobb eredményt, ezért ezt a módszert választottuk arra, hogy az angol nyelvre is kipróbáljuk. Így összehasonlítási alapunk lehet a szakirodalomban jó eredményeket elért további módszerekkel. A körülbelül 20 millió szót tartalmazó Penn Treebank szintaktikailag annotált szövegei szekciókra tagolódnak. Ebből a nagy adatbázisból elkülönítettek egy kisebb, körülbelül 1 millió szót tartalmazó részt, a 2-24-es szekciókat, melyen az összehasonlítani kívánt módszerek egységesen építhetnek modellt, és a pontosság mérése is egységes adatokon történik. Az általunk alkalmazott módszer esetén is ezeknek a feltételeknek megfelelően jártunk el.

A famintákon alapuló módszer alkalmazása során kipróbáltuk azt a speciális esetet is, amikor a faminták egymélységű fák. Ez az eset lényegében a környezetfüggetlen valószínűségi nyelvtant (PCFG-t) alkalmazó módszer, amit a magyar nyelv elemzése során is alapnak tekintettünk. Ezen a modellen az elemző 81.31%-os pontosságot ért el. A többmélységű famintákat alkalmazó elemzővel 85.73% volt az eredmény, ami 4,42%-os javulást jelent az alapmódszerhez képest. Az így elért pontosság megközelíti az angol nyelvre publikált eredményeket.

5 Összefoglalás és fejlesztési lehetőségek

A dolgozatban bemutatásra került egy a Szeged Treebank adattárára épülő adatbázis kialakítása. Ez a magyar nyelvű szintaktikai elemzőkhöz készül, hogy az adatbázis segítségével előállított elemzőket egységesen össze lehessen hasonlítani. Ezen az adatbázison kipróbáltunk néhány rendelkezésre álló módszert, melyek eredményei összehasonlítási alapot adhatnak a jövőben ezen az adatbázison alkalmazandó további algoritmusokhoz. A famintákon alapuló módszert – amely a legjobb eredményt érte el a magyar szövegeken – alkalmaztuk a Penn Treebank angol szövegeire is, és így képet kaptunk a két nyelv összetettségéről szintaxiselemzési szempontból.

A közeljövőben szeretnénk alkalmazni olyan szintaktikai elemzőket magyar nyelvre, amelyek jó eredményeket értek el angol szövegeken.

Irodalom

Abney S. (1991) Parsing by chunks, in *Principle-Based Parsing*. Kluwer Academic Publishers.

Argamon, S., Dagan, I., and Krymolowski, Y. (1998) A memory-based approach to learning shallow natural language patterns, in *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, Montreal, pp. 67-73.

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A., (2005) The Szeged Treebank, in *Proceedings of the TSD*, Karlovy Vary, pp. 123–131.

Hócza, A (2004) Noun Phrase Recognition with Tree Patterns, in the *Acta Cybernetica*, vol. 16, pp. 611–623.

Hócza, A., Felföldi, L., Kocsor, A., (2005) Learning Syntactic Patterns Using Boosting and Other Classifier Combination Schemas, in *Proceedings of the TSD*, Karlovy Vary, pp. 69–76.

Kis, B., Naszódy, M., Prószéky, G. (2003) Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer, *MSZNY 2003 konferenciakiadványa*, Szeged, 145-151 oldal.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993) Building a large annotated corpus of English: the Penn Treebank, Association for Computational Linguistics.

Ramshaw, L. A., and Marcus, M. P. (1995) Text Chunking Using Transformational-Based Learning, in *Proceedings of the Third ACL Workshop on Very Large Corpora,* Association for Computational Linguistics.

Simov K. (2001) CLaRK – an XML-based System for Corpora Development, in *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster, pp. 553-560.

Tjong Kim Sang, E. F., and Veenstra, J. (1999) Representing text chunks, in *Proceedings of EACL '99*, Association for Computational Linguistics.

Tjong Kim Sang, E. F. (2000) Noun Phrase Recognition by System Combination, in *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, Seattle, pp. 50-55.

Váradi T. (2003) Shallow Parsing of Hungarian Business News, in *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster, pp. 845-851.