

Mondatrész-felfedezés önszervező tanulással

Szepesvári Csaba¹, Kálmán László², Lukács Ágnes^{2,3}, Rebrus Péter²

¹ MTA SZTAKI, Budapest,
szcsaba@sztaki.hu

² ELTE/MTA Elméleti nyelvészet szakcsoport, MTA Nyelvtudományi Intézet
{kalman, rebrus, lukacsag@nytud.hu}

³ BME Kognitív tudományi tanszék
alukacs@cogsci.bme.hu

Kivonat: A cikkben egy újszerű gépi tanulási módszer elméleti háttérét és az első kísérlet végrehajtását ismertetjük. A módszer lényege, hogy a nyelvmodellben nem valamilyen generatív nyelvtant, hanem pusztán mondattani mintákat feltételezünk, és a rendszert nem elemzési vagy eldöntési feladattal teszteljük, hanem mondatok hasonlóságának felismerésével.

1 Bevezetés

Egy rövid évszázadra feledésbe merült Saussure-nek az a megállapítása, amely szerint az analógia nem a nyelvi változásoknak a szabályszerűtől eltérő formái mögött meghúzódó mechanizmus, hanem a szinkrón nyelvi rendszer összetartó ereje. Azért ismerjük fel egymás megnyilatkozásainak szerkezetét, mert azok más, már ismert kifejezések mintájára, analógiájára épülnek fel, és azért vagyunk képesek megérteni őket, mert a jelentésükre is ugyanez igaz. Ezeket a szerkezeti és jelentéstani mintázatokat a Saussure utáni évszázadban szabályok segítségével próbálták leírni. Pontosabban: az analógiás mintákat a szabályokkal próbálták helyettesíteni a tényleges megnyilatkozások, kifejezések szerkezetének és jelentésének magyarázatához.

1.1 Önszervező tanulás és nyelvtechnológia

A szintaktikai mintázatok felismerése és megfelelő kezelése előfeltétele a legtöbb természetesnyelv-feldolgozással kapcsolatos feladatnak. A szintaktikai mintázatok felfedezésének legkevésbé „elfogult” (biased), vagyis a legkevésbé előre megadott nyelvészeti kategóriát feltételező és legkevésbé elméletterhelt megközelítése az önszervező (unsupervised) tanulás [3], melynek során pusztán a korpuszban előforduló mondatok alapján a mintázatok kinyerésére és általánosítására. A legtöbb ilyen módszer a nyelvészetből ismert disztribúciós elemzés [8] valamilyen változatát alkalmazza a szintaktikai összetevők és ezeket összekapcsoló szabályok azonosítására.

A nyelvet önszervező tanulással megközelítő elképzelés olyan nyelvmodellen alapul, mely szerint a nyelvtudás vagy a nyelvtani tudás megfelelő megközelítése nem egy absztrakt gép, amely kizárólag grammatikus mondatokat képes előállítani, hanem egy olyan képesség, amely a nyelvi hasonlóságok és különbségek felismerésén alapul. Ennek a megközelítésnek a következménye, hogy a működő nyelvmodell szükségképpen alkalmas arra, hogy újszerű mintázatokról megállapítsa, hogy újszerűek, illetve, hogy az újszerű mintázatok a korábbi mintázatok tudása alapján beépítse, azaz a nyelvmodell képes önszervező módon továbbfejleszteni „saját magát”.

Kiinduló feltételezésünk szerint azért ismerjük fel egymás megnyilatkozásainak szerkezetét, mert azok más, már ismert kifejezések mintájára, analógiájára épülnek fel, és azért vagyunk képesek megérteni őket, mert a jelentésükre is ugyanez igaz. Az analógiák mögött rejtőzködő szerkezeti és jelentéstani mintázatok a nyelvészeti elméletek többsége a tényleges megnyilatkozások, kifejezések szerkezetének és jelentésének magyarázatára hivatott szabályok segítségével próbálja leírni. A mi kísérletünknek azonban nem célja sem a grammatikus megnyilatkozások nyelvtani jellemzése, „elemzése”, sem egy mögöttes „nyelvtan” rekonstruálása; a modellel szembeni elvárás „mindössze” annyi, hogy a gyakori mintázatok kinyerése után a alkalmas legyen analógiás problémák (aránypárok) megoldására.

1.2 Önszervező tanulás és nyelvelsajátítás

Az önszervező gépi nyelvtanulás eredményei a nyelvelsajátítás-kutatók számára is érdekesek lehetnek. A szakértők egyetértenek abban, hogy az anyanyelv elsajátítása az önszervező tanuláshoz hasonlóan explicit tanítás hiányában történik; a vita a tanuló mechanizmusok jellegéről és a kiinduló reprezentációk gazdagságáról és absztraktságáról folyik. Az önszervező modellek megjelenése a számítógépes nyelvmodellezésben párhuzamba állítható a nyelvelsajátítás-kutatás újabb eredményeivel és modelljeivel. Ezek a generatív nyelvészeti hagyományból kinőtt, veleszületetten absztrakt szabályokat és reprezentációkat és bonyolult algoritmusokat feltételező elméletek helyett egyszerű és a nyelv kognitív tartományán kívül is általánosan működő statisztikai tanulóméchanizmusokkal magyarázzák a komplex nyelvtan elsajátítását. Ahogy a nyelvészetben egyre nagyobb teret kapnak az új, lexikalista irányzatok — nemcsak a generatív táborral szemben, hanem azon belül is — a pszicholingvisztikában is az asszociatív tanulásra építő kutatási irányba tolódik el a kutatások vonala a nativistától az empirista szemlélet felé.

A nyelvelsajátítás a mai empirista felfogásban leginkább a nyelvi adatok és érzékelési bemenetek alapján történő szerkezet-elvonatkoztatás: a veleszületett tudás leginkább perceptuális primitívekből, szerkezetabsztraháló eljárásokból és a nyelv elemzésére és produkciójára szolgáló műveletekből áll. Bár egy tanulóalgoritmus sikeressége egy adott területen semmi esetre sem jelenti azt, hogy a gyerek is ugyanazzal a mechanizmussal sajátítja el az érintett nyelvi jelenséget, annyit azért elárul, hogy az adott bonyolultságú jelenség megtanulásához nem szükségszerű bonyolultabb procedúrát feltételeznünk. Az algoritmus sikertelenségéből pedig arra következtethetünk, hogy a gyerek vagy gazdagabb inputhoz fér hozzá, vagy erősebbek a kezdeti megszorításai [9].

A felügyelő nélküli tanulás egyik alapmechanizmusa a strukturalista nyelvészeti hagyományokban gyökerező disztribúciós elemzés, amely általában a nyelvi elemzés különböző szintjein címkézett elemeken (fonémák, morféma, szavak és nyelvtani kategóriák, frázisok) történhet — ilyenkor persze nem a címkéket tanulja az algoritmus, hanem valami mást: például nyelvtani kategóriák szerint címkézett korpuszon a jelentést (szemantikát). Vannak azonban olyan algoritmusok is, amelyek címkézetlen korpuszokból, pusztán a nyelven belüli disztribúciós információból próbálnak nyelvtant tanulni. A disztribúciós elemzést általában még különböző általános statisztikai elvek egészítik ki: ilyen lehet például a súlyok valószínűségének bayesi elvek alapján történő maximalizálása, a leírás hosszának minimalizálása (válaszd azt az elemzőt és elemzést, amelyek hosszainak az összege minimális), vagy a maximum entrópia elve (az elemzési kategóriák eloszlása legyen a megszorításokkal összeegyeztethető lehető legegyszerűsebb). A tanulás felügyelő nélkül történik, vagyis nincsen explicit segítség vagy visszajelzés; az elsajátítás csak közvetlenül a kijelentésekből (vagy írott szövegből) történik. A címkézetlen korpuszokon való ilyen tanulás mindenképpen hasznos lehet a nyelvészeti kutatás számára, mivel a tanulóalgoritmusnak elég explicitnek kell lennie ahhoz, hogy számítógépes program formájában implementálható és tesztelhető legyen, és így fényt deríthet arra, mi az, aminek nem előfeltétele a veleszületett tudás.

1.3 Az ADIOS rendszer

A pusztán disztribúciós elemzésre építő algoritmusok valójában számos területen meglepően sikeresnek bizonyultak. Kizárólag az együttes előfordulások statisztikájára építve eredményesen tanulnak szóosztályokat: feljegyzik azt a kontextust, amiben egy szó előfordul, majd hasonlóságot számolnak és klaszteranalízist végeznek, amelyek segítségével egybe csoportosítják azokat a szavakat, amelyek hasonló kontextusokban fordulnak elő. A módszer nem eleve megadott szintaktikai kategóriák szerint csoportosítja a szavakat, a létrejövő klaszterek azonban a megfelelő szinten címkézve jól megfelelnek a nyelvészeti szintaktikai kategóriáknak.

A nyelvészeti felügyelő nélküli modelljei közül az ADIOS (Automatic Distillation of Structures [6] [7] [18], más hasonló megközelítésről Alignment Based Learning: [20]).

A modell nagyon jó eredményeket ért el a nyelv különböző aspektusainak elsajátításában annotálatlan korpuszokon. A modell nyelvtani szemlélete a kognitív nyelvészet és a konstrukciós nyelvtan hozzáállását tükrözi, vagyis a nyelvtan maga a nyelvi egységek listája, fokozatos általánosságot, komplexitást és absztrakciót mutató mintázatok gyűjteménye [11]. A hasznos nyelvi egységek meghatározásának eszköze ennek megfelelően a disztribúciós elemzés: azoknak a mondatoknak az azonosítása, amelyek osztoznak bizonyos szósorokban, de egy helyen paradigmatis variabilitást mutatnak. Ez a rendszer két alapvető építőköve: a mintázat (vagy szintagma) és a változatosságot mutató helyen előforduló komplementáris disztribúcióban álló szimbólumok ekvivalenciaosztálya.

Az ADIOS tehát címkézetlen elemeken, alulról fölfelé építkezik, reprezentációs ereje három elven nyugszik: 1) azok a mintázatok fontosak, amelyek kevés szóval gyakran előfordulnak és jól általánosíthatók; 2) mivel a változatosság csak a mintázat

által meghatározott kontextusban lehet, ez a kontextusérzékeny általánosítás biztonságosabb, mint egyetemes szófajok vagy szabályok alapján; 3) a komplex mintázatok rekurzívan, hierarchikus mintázatban épülnek fel.

Az ADIOS empirikusan és automatikusan magától fedezi fel a nyelvi építőelemeket. Az egyik tanítókörpusza a CHILDES [5] [15] gyerekekhez szóló, 300000 mondatot tartalmazó beszédátirata volt. 14 nap alatt 3400 intuitíven is fontos mintázatot és 3200 szemantikailag megfelelő ekvivalenciaosztályt talált. Az új mondatokat létező mintázatok megosztott reprezentációjaként alakítja ki, és ezt használta fel azokban a tesztekben, amelyeken új inputokat kellett kezelnie. 10000 mondatos tréning után 1000 új CHILDES mondatnak illetve ugyanezek véletlen szórendű változatának elfogadhatóságát kellett megítélnie, ebben a feladatban már kevés tréning után is jól teljesített, és ugyanez igaz volt a távoli függőségi viszonyok (a modell szempontjából beágyazott ekvivalenciaosztályok) kezelésében. Szokatlan módon a teljesítményét az emberivel is összehasonlították. Egy fejlődési nyelvmegértési tesztben, ahol a feladat mondatok helyességének a megítélése volt, egy 8-8,5 éves gyerek szintjén teljesített. Kényszerválasztásos angol nyelvvizsgateszten 60%-ot teljesített (a véletlen 33% lett volna), és egy hasonló elfogadhatósági tesztben is nagyon hasonló teljesítményre az emberére.

Összefoglalva, az ADIOS-algoritmus háttérében álló elvek:

1. a mintázat fontosságának probabilisztikus inferenciája
2. kontextusfüggő általánosítás
3. a komplex mintázatok rekurzív felépítése

2 A felhasznált módszer

2.1 Elméleti alapok

Az általunk követett eljárás egy fontos vonásban különbözik az ADIOS filozófiájától, és ennek messzire ható következményei vannak. Mi ugyanis a nyelvészet fő irányzataival szemben azt gondoljuk, hogy a nyelvtannak és a nyelvtudásnak nem az a legfontosabb célja, hogy a „jó” mondatokat a „rossz” mondatoktól elkülönítse, hanem hogy a kommunikáció céljait szolgálja, vagyis nyelvi formákhoz nyelvi jelentéseket társítson. Bár a jelentéssel egyelőre nem foglalkozunk, ennek a megközelítésnek az a közvetlen következménye, hogy a nyelvtan nem a „jó” mondatok halmazának rekurzív megadására törekszik. Ez mind a feltárni kívánt mondat szerkezetekre, mind a rendszer lehetséges tesztelésére vonatkozó megfontolásokra kihat.

A szokásostól eltérően nem feltételezzük a mondat szerkezetek szigorúan hierarchikus felépítését, tehát még közvetve sem egy újraíró szabályrendszert próbálunk rekonstruálni a korpuszból. Az újraíró szabályrendszerek a nyelv formális nyelvi modelljeiből származó, lényegében ma is egyetlen hatékony eszközei annak, hogy a nyelvet mint mondathalmazt jellemezzük. Mivel nem ilyen jellemzés a célunk, nem szükséges ezt az eszközt használnunk. Az eljárás során felismert minták átfedhetik egymást, az alá-fölrendeltségi viszonyokban nem szükséges döntenünk, sőt, azt is toleráljuk, ha a mondatokban nem minden elemet tudunk lefedni. Ez a robusztusság

szempontjából is fontos (az ismeretlen szavak a legtöbb esetben nem okoznak felismerési problémát).

Ami a tesztelés lehetőségét illeti, a mondat szerkezet felismerését a fenti okokból nem tudjuk azzal tesztelni, hogy „jól tudunk-e elemezni” korábban még nem látott mondatokat. Ehelyett olyan tesztek alkalmazunk, amelyek azt mutatják meg, hogy észreveszünk-e a tanultak alapján fontos hasonlóságokat korábban még nem látott mondatok között. [19].

Az eljárás megvalósításához elsőként azoknak a hasonlóságoknak a körét rögzítjük, amelyek elvileg relevánsak lehetnek. (Ez nyelvi szintenként eltérő lehet: például a mondatban az *ab* és a *ba* sorozatok hasonlóan minősülhetnek, míg az alaktanban nem — a magyar morfológia ilyen elemzését ld. [10]). Ezek képezik az adott nyelvi szintről alkotott „elfogultságainkat”.

A második lépésben egy gyermeknyelvi korpusz alapján (CHILDES, [5] [15]) a valóságosan relevánsnak minősülő hasonlóságokat kerestük több módszer összehasonlításával. A relevancia mutatójaként elsődlegesen a gyakoriságot és másodlagosan információelméleti kritériumokat használtunk: a gyakori hasonlóságok egyúttal gyakori különbségeket is jelentenek, hiszen ha mondatok gyakran hasonlítanak egy bizonyos tulajdonságban, akkor ennek a komplementerében gyakran különböznek. A feldolgozásnak ez a fázisa sok más kísérletre is jellemző [14] [18] [20], de a célkitűzés minden esetben eltér a mienktől, mert valamilyen mögöttes kategóriarendszer vagy nyelvtan rekonstruálására irányul. A megközelítésünkben egyedi, hogy a mintázat-jelöltek hatékony felfedezésére hatékony adatbányászati algoritmusokat (pl. A PRIORI [1] [2]) használtunk — nagyságrendekkel csökkentve a futásidőt a korábbi egyszerűbb eljárásokhoz képest [18] [20].

Végül a gyakori különbségekből és mintázatokból olyan adatbázist építünk, amelynek használatával megoldhatók a minket érdeklő, $A:B = C:x$ alakú aránypárok, ahol az aránypárok tagjai mondatok és mondatrészek. A nyelvtudásnak ilyen feladatokkal való tesztelése nem ismeretlen sem a mindennapi életben (pl. GRE-tesztek), sem a számítógépes nyelvészeten [12] [19]. Például a *Mary is sleeping : Mary = Joe left : x* aránypárnak $x = \text{Joe a megoldása}$, amit úgy is meg lehet fogalmazni, hogy a rendszer „felismeri az alanyt”, anélkül, hogy akár mondat szerkezetet, akár mondatrész-címkéket rendelne a kifejezésekhez. Az aránypárok megoldásához feltesszük: az $A : B$ különbségnek maximálisan relevánsnak kell lennie, és meg kell egyeznie a $C : x$ különbséggel, és ugyanígy fontosnak és azonosnak kell lennie az $A : C$ és a $B : x$ különbségeknek is. Előadásunkban a módszerek bemutatása mellett összehasonlítjuk az általunk vizsgált különböző önszervező tanulási algoritmusok sikerességét ilyen analógiás feladatok megoldásában.

2.2 Az algoritmus

Az algoritmus többféle lépés iterált alkalmazásából áll, ami mindaddig folyik, amíg „érdekes jelenségeket” fedezünk fel a korpuszban. „Érdekes jelenségeknek” minősülnek a következők:

1. Gyakran ismétlődő sorozatok („kollokációk”). Ha egy sorozat kis különbségekkel ismétlődik gyakran, és a különbségek nem túl sokfélék, akkor is „kollokációról” beszélünk.
2. Érdekes kontextusok. Azok a kontextusok számítanak érdekesnek, amelyekben elég sokféle egység fordul elő ahhoz, hogy feltételezzük, egy osztályt alkotnak (ld. 3.), de nem olyan sokféle, hogy az együttes előfordulás irrelevanciájára gyanakodjunk. Itt „egységen” szavakat, szóosztályokat és ezekből alkotott gyakori sorozatokat egyaránt értünk.
3. Azonosan viselkedő osztályok. Azokról a szavakról, szóosztályokról, sorozatokról, amelyek nagyjából ugyanazokban az „érdekes” kontextusokban fordulnak elő, feltételezzük, hogy ugyanabba az osztályba tartoznak, legalábbis az illető kontextus szempontjából.

Minden iterációs lépésben átírjuk a kiinduló korpuszt a felfedezett jelenségek felhasználásával. Ez egyfajta címkézés, csak annyiban sajátos, hogy a következő lépésekben a címkék és a korábbi információ egyaránt a felfedező eljárás tárgyát képezi. Az algoritmus visszalépési lehetőséget is tartalmaz: ha azt tapasztaljuk, hogy „túláltalánosítottunk”, azaz egy osztálynak elég nagy részosztályai nem viselkednek egyöntetű módon, akkor részekre bontjuk az osztályt, és visszalépünk a keletkezésekor érvényes korpuszhoz. Osztályok egyesítését azonban nem engedjük meg (ha már különböző viselkedést tapasztaltunk, azt nem tekinthetjük meg nem történtnek).

2.3 Eredmények

A felfedező eljárás eredményét kétféleképpen teszteltük: aránypárok megoldásával és mondatkiegészítési feladatokkal. Az aránypárok megoldása bonyolultabb eljárás, de jobban megfelel az elméleti alapfeltevéseinknek, amelyekről fent szoltunk. A mondatkiegészítés egyszerűbb próba, viszont feltételezi, hogy van valamilyen módszerünk annak ellenőrzésére, hogy két mondat közül melyikre tudunk „elfogadhatóban” felfedezett mintákat illeszteni.

A tesztelésben az az „illesztőprogram” játssza a középponti szerepet, amely a korpuszban felfedezett jelenségeket megpróbálja még nem látott mondatban felfedezni. Az illesztéseken egy kvázi-algebrai struktúrát definiáltunk (metszet, különbség), valamint egy „lefedési mértéket”, amely megadja, hogy mennyire „részletesen” fedi le az illető illesztés a mondatot.

Az $A : B = C : x$ alakú aránypárok megoldásánál (ahol x értékét megadott kifejezések közül kell kiválasztani) a következőképpen döntünk: az $A - B$ és a $B - A$ különbségekhez hasonlítjuk a $C - X$, illetve $X - C$ különbségeket (ahol X a vizsgált jelölt). Az összehasonlítás azt jelenti, hogy a metszetek nagyságát hasonlítjuk össze. Az lesz a nyertes jelölt, amelynél ezek a különbségek a legjobban hasonlítanak (a két metszet nagyságának összege a legnagyobb). A kiegészítési feladatnál a lehetséges jelöltekkel kiegészített mondatokon futtatjuk az illesztőprogramot, és az eredmények „lefedési mértékét” hasonlítjuk össze. A „lefedési mérték” annál nagyobb, minél több szóra minél hosszabb sorozatokat tudunk illeszteni. (A szóosztályba sorolás 1 hosszú sorozatnak minősül.)

A cikk leadásának órájában még csak megkezdttük a nagyobb korpuszsal való kísérletezést. A „játék” korpuszunkból, amelyben csak névelőt, 50 főnevet, 50 melléknevet és 25 igét használtunk, és ezek minden nyelvtanilag helyes kombinációját be-

vettük (a nyelvtan természetesen „Det (A) N V” volt), tanulás után a rendszer 100% pontossággal képes volt megoldani az aránypáros és a kiegészítő feladatokat.

Bibliográfia

- [1] Agrawal, R., T. Imielinski, and A. N. Swami (1993) Database Mining: A Performance Perspective. *IEEE Trans. Knowl. Data Eng.* 5(6): 914-925
- [2] Agrawal, R. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. *VLDB 1994*: 487-499
- [3] Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. *Int. J. Digit. Libr.* 1 (1997) 108–121
- [4] Bruce, K.B., Cardelli, L., Pierce, B.C.: Comparing Object Encodings. In: Abadi, M., Ito, T. (eds.): *Theoretical Aspects of Computer Software. Lecture Notes in Computer Science*, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York (1997) 415–438
- [5] <http://childes.psy.cmu.edu/>
- [6] Edelman, S., Solan, Z., Horn, D. and E. Ruppín (2003) Rich Syntax from a Raw Corpus: Unsupervised Does It; a position paper to be presented at Syntax, Semantics and Statistics; a NIPS-2003 workshop, Whistler, BC.
- [7] Edelman, S., Z. Solan, D. Horn and E. Ruppín (2004) Bridging computational, formal and psycholinguistic approaches to language. to appear in *Proc. of the 26th Conference of the Cognitive Science Society*, Chicago, IL, Aug. 2004.
- [8] Harris, Z. S. (1954). Distributional structure. *Word*, 10:140–162.
- [9] Johnson, Mark and Riezler, Stefan (2002) Statistical models of syntax learning and use. *Cognitive Science*, 26, 239-253.
- [10] Kálmán, L., P. Rebrus and M. Törkenczy (2005): Hungarian linking vowels: An analogy-based approach. Paper presented at the 7th International Conference on the Structure of Hungarian Veszprém, Hungary, May 29-31, 2005
- [11] Langacker, R. W. (1987) *Foundations of cognitive grammar: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
- [12] Lepage, Y. (1998) Solving Analogies on Words: an Algorithm. *Proceedings of COLING-ACL'98*, Montréal, August vol. I, pp. 728-735.
- [13] van Leeuwen, J. (ed.): *Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science*, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)
- [14] Lieven, E., Tomasello, M., Behrens, H. Speares, J. (2003) Early syntactic creativity: a usage-based approach. *Journal of Child Language*, 30, 333-370
- [15] MacWhinney, B. and C. Snow (1985) The child language exchange system. *Journal of Computational Linguistics*, 12:271–296.
- [16] Michalewicz, Z.: *Genetic Algorithms + Data Structures = Evolution Programs*. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
- [17] Pedersen, B., Edelman, S., Solan, Z., Horn, D., and E. Ruppín, (2004) Some Tests of an Unsupervised Model of Language Acquisition, in *Proc. COLING-2004 Workshop on Psycho-computational Models of Human Language Acquisition*, Geneva, Switzerland.
- [18] Solan, Z., Horn, D., Ruppín, E. and Edelman, S. (2003) Unsupervised Context Sensitive Language Acquisition from a Large Corpus; in *Proc. NIPS-2003*..
- [19] Turney, P.D., and Littman, M.L. (2005), Corpus-based learning of analogies and semantic relations, *Machine Learning*, 60 (1-3), 251-278.
- [20] van Zaanen, M. M. (2003) Theoretical and Practical Experiences with Alignment-Based Learning. *Proceedings of Australasian Language Technology Workshop*, Melbourne, Australia, December.