Strukturált nyelvi adatbázis létrehozása gépi tanulási módszerekkel

Kutatási terv

Gábor Kata

A kutatásom célja egy strukturált lexikai adatbázis felépítése magyar nyelvre gépi tanulási módszerek alkalmazásával. A munka eredményeképpen előálló lexikai adatbázis igei tételeket és hozzájuk tartozó szintaktikai információt (bővítménykeret), valamint az igék szemantikai csoportosítását tartalmazza. A kutatás elméleti hozadéka a számítógépes nyelvészet területén olyan módszerek kidolgozása, melyek lehetővé teszik, hogy természetes nyelvi korpuszból szintaktikai tulajdonságok figyelembe vételével szemantikai információt nyerjünk ki. Gyakorlati szempontból a lexikai adatbázisok automatikus előállításának legfőbb előnyei a rugalmasság és a bővíthetőség. A munka eredményeképpen előálló szemantikai információt tartalmazó számítógépes lexikon felhasználható például információkinyerő és kereső rendszerekben.

A kutatás jelentősége és az eredmények gyakorlati felhasználása

Az információs társadalom és az internet szerepének felértékelődésével megnőtt az érdeklődés az olyan nyelvtechnológiai alkalmazások iránt, melyek intelligens módon képesek természetes nyelvi szövegeket feldolgozni, azaz a morfológiai és szintaktikai elemzésen túl a szövegek *tartalmi* elemzésére is képesek. A könnyen hozzáférhető információ mennyisége, és ezzel párhuzamosan a feldolgozandó szövegek mennyisége is nő, így a versenyben már nem az a döntő fontosságú, hogy ki fér hozzá az információhoz, hanem hogy ki milyen gyorsan képes feldolgozni. Az információkinyerő (information extraction, IE) rendszerek, a kulcsszó-alapú módszeren túllépő intelligens keresőmotorok, a kérdés-válasz rendszerek (question answering, QA) kitüntetett területei a nyelvtechnológiai kutatásoknak, az eredmények ugyanis szinte azonnal beépülnek olyan alkalmazásokba, melyek átalakítják mindennapi internetezési szokásainkat. Az intelligens szövegfeldolgozást végző rendszerek ugyanakkor nagyon komplex, a szöveg megértésén alapuló feladatok megoldását tűzik ki célul: olyan feladatokat, melyeket mostanáig emberek végeztek. Ilyen feladatok megvalósításához nélkülözhetetlen, hogy a jelentésre vonatkozó emberi tudást, szemantikát integráljuk a rendszerekbe. Ennek első és legfontosabb lépcsőfoka a szemantikai információt is tartalmazó természetes nyelvi adatbázisok létrehozása.

A kutatás nemzetközi és magyarországi előzményei

Az elmúlt évtizedben a lexikai adatbázisok automatikus előállítása a nyelvtechnológiai kutatások kiemelt területévé vált. Ez egyfelől azzal indokolható, hogy az emberi munkaerő használata költségesebb és lassúbb, másfelől a szöveg-feldolgozási feladatokban fontos szerepet kap az automatikus bővíthetőség és más területekre való alkalmazás (*domain adaption*) szempontja. Automatikus módszer használata esetén a megvalósítás kevésbé munkaigényes, és bármikor újra alkalmazható, ezáltal az adatbázis folyamatosan frissíthető, karbantartható.

A korai automatikus lexikonépítési kísérletekben nem számítógépes célokra készült, kétnyelvű vagy értelmező szótárak elektronikus változatát használták nyersanyagként (például [Briscoe et al., 1990], [Sanfilippo and Poznanski, 1992]). Az automatikus módszerek közül ez a megközelítés áll legközelebb a kézi előállításhoz, így megvan az az előnye, hogy a szótárból kinyert információ már keresztülment egy emberi szűrőn. Ugyanakkor éppen emiatt rendelkezik a nem automatikus módszer fő hátrányaival: nem elég rugalmas, és nem teszi lehetővé az automatikus bővítést, ezáltal nem vihető át más területre. A szótár használatánál robusztusabb megközelítést jelent az igei vonzatkeret-információ automatikus kinyerése nagyméretű korpuszokból. Ezen kívül az ember által készített erőforrásokban (szótárak, nyelvi ontológiák) elkerülhetetlenül előforduló inkonzisztenciák is amellett szólnak, hogy helyettesítsük őket olyan módszerrel, amely minél több nyelvi adatra és szigorú statisztikai szűrőkre támaszkodik. Napjainkban már a legtöbb európai nyelvre, így a magyarra is rendelkezésre állnak olyan erőforrások (morfológiailag elemzett és egyértelműsített korpusz: [Váradi, 2002], szintaktikailag annotált treebank: [Csendes et al., 2005]), melyek lehetővé teszik gépi tanulási módszerek alkalmazását a vonzatkeret-információ és a szemantikai tulajdonságok kinyerése céljából.
Ilyen jellegű feladatok megoldásához először is meg kell fogalmaznunk feltételezésünket arról, hogy a szóhoz tartozó keresett lexikai információt milyen, a korpuszbeli előfordulások alapján számszerűsíthető tulajdonságok reprezentálják. A vonzatkeret-kinyerés esetében a legtöbb módszer az ige és a vonzatjelölt együttes előfordulási statisztikáiból indul ki ([Briscoe & Carroll, 1997], [Pereira et al. 1993], magyarra [Sass, 2006]). Az előfordulások természetesen szűrhetők további nyelvtani információ alapján, attól függően, hogy a korpusz milyen szintű nyelvi elemzéssel rendelkezik. [Brent, 1993] megközelítése nem igényli szintaktikailag elemzett korpusz használatát: nagyon precíz mintákat határoz meg, melyek nagyon pontos közelítését adják a szintaktikai függőségi viszonyoknak, ám ezzel a módszerrel a korpuszbeli információ nagy része elvész, és csak a legbiztosabb találatok hasznosíthatók. Ezzel szemben [Manning, 1993] minél több információt igyekszik kinyerni a korpuszból, és a zajt a szigorúbb statisztikai szűrők alkalmazásával csökkenti.
A tisztán szintaktikai vonzatkeret-információ kinyerésén túl egyre nagyobb szerepet kap a szemantikai tulajdonságok automatikus felismerése is. Az igék szemantikai szelekciós tulajdonságainak gépi tanulására [Resnik, 1993] a WordNet nyelvi ontológián alapuló módszert javasolt. [Erk, 2007] hasonló hatékonyságú eredményt ér el úgy, hogy a WordNet helyett az argumentumként előforduló főnevek disztribúciójának hasonlósági adataiból indul ki, melyet külön tanulókorpuszból sajátít el.

Kutatásom közvetlen előzményének ([Schulte im Walde & Brew, 2002], [Joanis & Stevenson, 2003], valamint a magyarra vonatkozó kísérletek közül [Gábor & Héja, 2007] és [Sass 2007] tekinthető. Ezen kutatások célja Beth Levin [Levin, 1993] igeosztályainak megfelelő igei csoportosítás automatikus kinyerése korpuszokból. Levin szintaktikai alternációk alapján definiált igeosztályokat, azzal a mögöttes feltételezéssel élve, hogy a hasonló szintaktikai mintázatokban előforduló igék jelentése is hasonló. Az általa létrehozott igeosztályok feltételezése szerint egy vagy több, ún. metapredikátumokkal leírható jelentéskomponensben osztoznak. Ez a hipotézis egybevág azzal az igénnyel, hogy a WordNet vagy egyéb erőforrások helyett korpuszbeli adatok alapján próbáljunk szemantikai csoportokat definiálni. Kutatásomhoz a Magyar Nemzeti Szövegtár (MNSz) adatait kívánom felhasználni. Olyan gépi tanulási módszert keresek, mely alkalmazható a szövegtár adataira, vagyis nem igényel szintaktikailag elemzett korpuszt, és hatékonyan működik szabad szórendű, nem konfigurációs nyelvekre.

Részfeladatok, főbb megoldandó problémák

A kutatás során megoldandó probléma egy nem felügyelt csoportosítási feladat (*unsupervised clustering*), melynél nem ismerjük előre a csoportokat, amelyekbe az elemeket (esetünkben igéket) be akarjuk sorolni. A magyarra ugyanis eddig nem született Levinéhez hasonló leírás az igékről. [Schulte im Walde, 2000], és nyomán [Gábor & Héja, 2007] szintaktikailag elemzett korpuszt használ az igék csoportosításához. Magyar szintaktikailag elemzett korpusz csak egy létezik, a Szeged Treebank, amely azonban kis méreténél (1,2 millió szó) fogva nem alkalmas arra, hogy ritka igékre vonatkozó információt nyerjünk ki belőle. Hosszú távon elkerülhetetlen, hogy a kísérleteket kiterjesszük az MNSz adataira, ehhez azonban módosítani kell a clustering során figyelembe vett jegykészletet, mivel nem annotált korpusz használatakor nem hivatkozhatunk bővítménykeret-információra. [Merlo & Stevenson, 2001] csak szófajilag egyértelműsített korpuszt használ, az ő munkájuk viszont előre megadott csoportok automatikus kinyerését célozza. A csoportosítási feladathoz előre definiálták a célcsoportokra jellemző, illetve azokat egymástól megkülönböztető szintaktikai alternációkat, valamint az alternáló vonzatkeretek korpuszbeli előfordulásainak mintázatait (vagyis azokat a számszerűsíthető disztribúciós jellemzőket, melyek a lehető legjobban közelítik az alternációkat), és algoritmusuk ezeket vizsgálva döntötte el az egyes igékről, hogy mely csoportba tartoznak. Később [Joanis & Stevenson, 2003] javaslatot tett egy olyan jegykészletre, mely segítségével *bármely*, az angol igék leírásánál használt alternáció korpuszbeli előfordulásai felismerhetők, így a módszer lehetővé teszi az összes ige besorolását az angolra jellemző igecsoportok valamelyikébe. A magyar adatok csoportosításának egy lehetséges módja ehhez hasonló jegykészlet meghatározása magyar nyelvre. Ehhez egyfelől felhasználhatók a [Gábor & Héja, 2007] és [Sass, 2007] csoportosítási kísérletei eredményeképpen kapott szemantikai csoportok, illetve a rájuk jellemző szintaktikai mintázatok, másfelől egyéb, szemantikai alapú csoportosítások, például a magyar WordNet ([Kuti et al., 2005]) és a Nyelvtechnológiai osztály által épített vonzatkeret-adatbázis összevetése.

A probléma egy másik megközelítése [Gábor & Héja, 2006] alapján az adjunktumok disztribúciójának figyelembe vétele lehet. Mind Schulte im Walde, mind Stevenson munkáira jellemző, hogy ([Levin, 1993] alapvetéseit elfogadva) csak az ige vonzatainak disztribúciójára alapoznak. [Gábor & Héja, 2006] szerint az adjunktumok előfordulásait meghatározza, hogy a mondat predikátuma milyen szemantikai csoportba tartozik. A kevésbé produktív adjuntumok használata így jellemző lehet az igei jelentésre. Ugyanakkor az összes adjunktum figyelembe vétele túl nagy jegykészletet, és ezáltal nagy zajt eredményez. A nem felügyelt gépi tanulási módszerek érzékenyebbek a zajra, mint a felügyelt módszerek, ezért nagy a jelentősége annak, hogy ki tudjuk választani a releváns jegykészletet, esetünkben az igei szemantikától nagymértékben függő adjunktumok körét, és csak ezeket vegyük figyelembe. [Sass 2007] az ige környezetében előforduló főnevek esetragjának gyakorisága alapján redukálja a jegykészletet, vagyis a leggyakoribb esetragok valamenyikét viselő főnevek lemmáját használja szintaktikai környezetként. Várakozásom szerint ennél jobb eredményt lehet elérni, ha a figyelembe vett esetragokat nem gyakoriság szerint, hanem nyelvészeti tudást felhasználva, az igéhez mint régenshez való szemantikai/szintaktikai viszony alapján szűrjük.[1]

[1] A második módszerhez, vagyis az igéhez tartozó adjunktumok és vonzatok tulajdonságainak jegykészletként való felhasználásához legalább részlegesen elemzett korpuszra van szükség. Ehhez az MNSz-t kívánom felhasználni, a részleges elemzéshez (*chunking*) pedig az osztályon rendelkezésre álló NooJ-alapú elemzőt tervezem használni.

Kutatásomban össze fogom hasonlítani az adjuntumok disztribúcióján alapuló csoportosítás, illetve az alternációkra jellemző mintázatok gyakoriságán alapuló jegykészlet használatával kapott eredményeket. Az összehasonlítás jelentőségét elsősorban abban látom, hogy segíthet meghatározni azoknak a szintaktikai tulajdonságoknak a pontos körét, melyek a magyar nyelvben az ige szemantikájától függnek. Mivel a szemantikai csoportosításhoz figyelembe veendő tulajdonságok mindenkor nyelvfüggőek, ezért az összehasonlítás nem azt hivatott eldönteni, hogy általában melyik módszer a hatékonyabb, ám hasznosnak bizonyulhat más, a magyarhoz (szórendi és morfológiai szempontból) hasonló szerkezetű nyelvek feldolgozásához. Elképzelhető eredménynek tartom, hogy az alternációk közelítésére használt jegykészlet nagy hasonlóságot fog mutatni az igecsoportokra jellemző esetragos névszói csoportok körével.

Természetesen egy korpusz sem lehet elég nagy ahhoz, hogy ne szembesüljünk az adathiány (*data sparseness)* problémájával. Mivel a korpusz sosem fogja lefedni a teljes nyelvet az adott időpillanatban, nem feltételezhetjük, hogy ami nem szerepel a korpuszban, az a nyelvnek sem része. Emellett a statisztikai módszerek nem megoldott problémája, hogy a kevés adattal példázott jelenségekről nem tudnak pontos becslést adni. A lexikonépítési módszer azonban akkor hatékony, ha a ritkább igékre is működik, valamint a módszer hordozhatósága is nagyban függ attól, hogy milyen becslést tud adni kevésszer vagy egyáltalán nem látott adatokra. A jegykészlet és a csoportosítási módszer kiválasztását is befolyásolja, hogy mennyire kell felkészülnünk a *data sparseness* problémájára. A statisztikai módszerek a nem vagy kevésszer előforduló adatokat simítással korrigálják. A megfelelő algoritmus kiválasztásának része, hogy a hozzá illő simítási eljárást meghatározzuk. [Gábor & Héja, 2007] csak technikai problémaként kezeli az adathiányt, mivel a leggyakoribb igék vizsgálatából indulnak ki, és feltételezik, hogy ezen igék esetében valamely bővítmény hiánya valódi szemantikai inkompatibilitást jelent az ige és a bővítmény között. Robusztusabb módszer kidolgozásához azonban már jósló erejű simítást kell alkalmazni.

A kutatás utolsó évében a módszer pontosítása mellett külön figyelmet szeretnék fordítani az adatbázis hasznosíthatóságának biztosítására is. Ennek érdekében az első részfeladat egy hordozható, jól karbantartható formátum kiválasztása lesz. A hasznosíthatóság garanciáját természetesen a tényleges felhasználás jelenti. Ezért kísérletet fogok tenni rá, hogy az osztály dolgozóival együttműködve az adatbázisnak legalább egy részét prototípus szinten beépítsem valamely, az osztály közreműködésével fejlesztett nyelvtechnológiai (gépi fordító, illetve információkinyerő) rendszerbe.

Időbeosztás

Feladat	Részfeladatok leírása	Hónap	Eredmény
Nem felügyelt tanulási kísérletek	MNSz részkorpusz kiválasztása, előkészítése, részleges elemzése	1-3	Annotált részkorpusz
	Jegykészlet meghatározása (releváns vonzatkeret- és adjunktum minták kiválasztása)	2-6	
	Nem felügyelt csoportosítási kísérletek (különböző klaszterezési módszerek) kipróbálása	4-8	Automatikusan előálló szemantikai osztályok
	Eredmények kiértékelése	8-10	Tanulmány
Felügyelt tanulási kísérletek	A magyarban releváns szemantikai osztályok/tulajdonságok keresése (Wnet, korábbi kísérletek adatai alapján)	11-14	Manuálisan definiált szemantikai osztályok
	Az osztályokra jellemző szintaktikai környezet meghatározása	13-15	
	Jegykészlet meghatározása (a szintaktikai környezetet közelítőleg leíró korpuszbeli minták leírása)	14-17	
	Felügyelt csoportosítási kísérletek	17-20	Automatikusan előálló szemantikai osztályok
	Eredmények kiértékelése	20-21	Tanulmány
Simítás	a módszereknek megfelelő simítási eljárás kiválasztása	2-8 14-20	
Összehasonlítás	A nem felügyelt és a felügyelt kísérletek eredményeinek összevetése	21-25	Tanulmány
	Az adatbázisba felveendő szemantikai osztályok kiválasztása	21-25	
Adatbázis-építés	Megfelelő formátum kiválasztása	25-27	
	A vonzatkeret-táblázatban kódolt információ és a szemantikai csoportok integrálása az adatbázisba	24-32	Többszintű, strukturált nyelvi adatbázis
Felhasználás	Kísérlet az adatbázis (teljes vagy részleges) beépítésére egy nyelvtechnológiai alkalmazásba	25-36	

Hivatkozások:

Brent, Michael, 1993: From grammar to lexicon: unsupervised learning of lexical syntax. Computational Linguistics 19, 2 (Jun. 1993), 243-262. o.

Briscoe, Ted, Anne Copestake and Bran Boguraev, 1990: Enjoy the paper: lexical semantics via lexicology. In Proceedings of the 13th International Conference on Computational Linguistics (COLING-90), Helsinki, 42—47. o.

Briscoe, Ted and John Carroll, 1997: Automatic Extraction of Subcategorization from Corpora. In Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97), Washington, DC, USA

Csendes, Dóra János Csirik, Tibor Gyimóthy: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. TSD 2004: 41-48. o.

Gábor, Kata & Héja Enikő.: *Clustering Hungarian Verbs on the Basis of Complementation Patterns.* In: Proceedings of the ACL'07 conference, Prága, 2007.

Gábor, Kata, Héja Enikő: *Predikátumok és szabad határozók.* In: Kálmán L. (szerk): A titkos kötet. Nyelvészeti tanulmányok Bánréti Zoltán és Komlósy András tiszteletére, Tinta kiadó, Budapest 2006

Joanis, Eric & Suzanne Stevenson. 2003. A general feature space for automatic verb classification. In Proc. of the 10th Conference of the EACL (EACL 2003), 163 - 170. o.

Kuti et al., 2005. Javaslat a Magyar igei WordNet kialakítására. In: Alexin Z., Csendes D. (szerk.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged, 2005. 79-88.o.

Levin, Beth: English Verb Classes and Alternations. A Preliminary Investigation. The University of Chicago Press, Chicago, 1993.

Manning, Christopher, 1993: Automatic acquisition of a large subcategorization dictionary from corpora. Proceedings of the 31st ACL, pp. 235-242.

Merlo, Paola and Suzanne Stevenson, 2001: Automatic Verb Classification Based on Statistical Distributions of Argument Structure. In Computational Linguistics. 27: 3, 373-408. o.

Pereira, Fernando, N. Tishby és L. Lee 1993: Distributional Clustering of English Words. In Proceedings of the 31st Annual Meeting of the ACL, 183-190. o.

Resnik, Philip, 1993: Selection and Information: A Class-Based Approach to Lexical Relationships. Doctoral Dissertation, Department of Computer and Information Science, University of Pennsylvania

Sanfilippo, Andrea & Victor Poznanski, 1992: The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Resources. In Proceedings of the Applied Natural Language Processing Conference, Trento, Italy, 80—87. o.

Sass Bálint, 2006: Extracting Idiomatic Hungarian Verb Frames. In Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (szerk.): Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006 Turku, Finnország, pp. 303-309.

Sass Bálint, 2007: First Attempt to Automatically Generate Hungarian Semantic Verb Classes. In: Proceedings of the 4th Corpus Linguistics Conference, Birmingham, 2007.

Schulte im Walde, Sabine & Chris Brew, 2002: Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA.

Váradi, Tamás, 2002: The Hungarian National Corpus. In Proceedings of the Second International Conference on. Language Resources and Evaluation, Las Palmas