DOKTORI ÉRTEKEZÉS TÉZISEI

Tóth Ágoston

Perspectives on the Lexicon

Témavezető: Dr. Hollósy Béla

DE Bölcsészettudományi Kar
2005

1. Az értekezés célkitűzése, a téma körülhatárolása

Az értekezés célja a lexikon szerepének, működésének többszempontú vizsgálata. Ennek során először elemzem a lexikai komponens pozícióját a chomskyánus generatív grammatika kiválasztott alternatív elméleteiben, vizsgálva, hogy a mondattan és az alaktan között kialakuló munkamegosztás (különösen a ragozás és a képzés kérdésében) hogyan befolyásolja a nyelvészek lexikonról kialakított elképzeléseit. A folytatásban a poliszémia és homonímia kérdéskörének tömör elemzésével szemléltetem azt, hogy a jelentéstani alapproblémák komoly buktatókat rejtenek a lexikon tervezése során. Az elmélet mellett a gyakorlatra is hangsúlyt fektet az értekezés: megvizsgálom, hogy két jelentős lexikai adatbázisban, a WordNet-ben és a FrameNet-ben a jelentés mely aspektusai köré szervezték a tárolt információkat, miközben a WordNet-tel kapcsolatban az elméleti oldalról korábban vizsgált jelenségeket, főként a ragozás és a szóképzés kezelésének módját saját vizsgálatokat is végezve elemzem. Ezek az adatbázisok kiemelkednek a természetesnyelv-feldolgozásban használatos szóalapú adatbázisok köréből szerkesztési jegyeik, valamint méretük, kidolgozottságuk miatt.

Az értekezés fontos újítása, hogy a vizsgálatba bevonja a konnekcionista, mesterséges neurális hálózat alapú kísérletek releváns tapasztalatait. Előbb egy áttekintő fejezetben látjuk, ahogy felvázolódik a nyelvi információ ábrázolásának egy lehetséges módszertana elsősorban Rumelhart és Elman kísérleteiben, majd bemutatok néhány mesterséges neurális hálózat konstrukciót, melyek a nyelvi bemenet kezelésében jelentős potenciállal rendelkeznek. Végül ismertetem önálló kutatásomat, melyhez kifejlesztettem egy sajátos neurális hálózat szerkezetet, melyet FrameNet-szerű jelentéstani keret (frame) és keret-elem (frame element) felismerésre tanítottam be generált korpuszon. A következő tervezési célokat tűztem ki magam elé:

- A bemeneten legyen lehetséges tetszőleges számú különböző input (elsősorban szavak) megkülönböztetése.
- Legyen lehetőség a nyelvi bemenetet nem atomi, hanem belső szerkezettel rendelkező objektumokként feldolgozni.
- Szükség esetén a rendszer legyen képes szavakból álló komplex kifejezésekkel dolgozni, melyek megfelelhetnek szóösszetételeknek, idiómáknak, összetevőknek, stb.
- A feldolgozás ne legyen izolált mondatokra korlátozott.

A mesterséges neurális hálózat viselkedését egy kísérletsorban vizsgáltam a betanítottól eltérő bemenet megjelenése esetén. A tézisfüzet 3. és 4. fejezetei a kísérletben alkalmazott módszereket és a mérések eredményeit ismertetik.

Az értekezés méltatja a leírás és implementáció szintjén egyaránt fontos eredményeket felmutató, modulként kezelt szimbólummanipulációs lexikon fontos szerepét a nyelvi „egyediségek" (idiosyncrasies) megfelelő leírásában. A konnekcionista nyelvi modellezés csak lassan jut túl a nyelvi alapkutatás fázisán, így a szóalapú információtárolásra vonatkozó eddigi eredményei is korlátozottak. Éppen ezért nem vállalkozom a jelen műben is csak alapjaiban megfogalmazódó konnekcionista „szórványlexikon" (az értekezésben szereplő megnevezése "dissolved lexicon – distributed lexis") szisztematikus összevetésére a hagyományosan kutatott és használt szimbólummanipulációs lexikonnal. Célom annak szemléltetése, hogy a konnekcionista szórványlexikonon alapuló nyelvi feldolgozás valós alternatíva, mely természetéből fakadóan (jelenleg csak elméleti lehetőségként) erőfeszítés nélkül kiküszöbölheti a szimbólummanipuláló nyelvi modulok munkamegosztásának újradefiniálásakor keletkező funkciózavart, csakúgy, mint a lexikai szemantika által előrevetített, a klasszikus kategorizálás és jelentésfelsorolás számára nem megoldott elméleti problémákat, gyakorlati szempontból pedig a fentiekből következő részfeladatokat (pl. szemantikai egyértelműsítés).

2. A tudományos háttér

2.1 A lexikai komponens chomskyánus megközelítése

A lexikon szerepéről és tartalmáról szóló elképzeléseinket jelentős mértékben meghatározta a chomskyánus generatív nyelvtan. Chomsky (1957) morféma-alapú lexikont és morfémákat manipuláló transzformációkat vizionál. Chomsky (1965) szintén mellőzi a morfológiát, mint önálló modult. Az inflexióval kapcsolatban kifejti, hogy a lexikonnak tartalmaznia kell bizonyos paradigma-dimenziókat, mint például a német nyelv esetében a nyelvtani nemet (gender), míg más dimenziók, így például a szám (number) a frázisszerkezetbe történő beillesztés kapcsán, vagy még később, a transzformációk során adódnak a mondat elemeihez. Chomsky rámutat, hogy a szóképzés potenciálisan „kvázi-produktív", valamint kifejti, hogy a produktívabb folyamatok eredményeként létrejövő alakokat generálni, a kevésbé produktívakat pedig tárolni érdemes. Chomsky a nominalizációt választja ki a produktívabb folyamatok problémamentes példájaként.

Chomksy (1970) visszatér a nominalizációra. Különbséget tesz gerundívumos (pl. *John's being eager to please*) és képzett (*John's eagerness to please*) nominalizáció között.

“Eléggé szabályszerű”-nek nevezve a gerundívumos alakot, ennek kezelését a transzformációs komponensre bízza; a szóképzésben megnyilvánuló nominalizáció eredményét azonban már kevésbé tartja regulárisnak, ezért ezeket az alakokat a lexikonban tárolná: ezzel Chomsky megfogalmazza a lexikalizmus programját.

Halle (1973) az erős lexikalizmus prominens alakja: alternatív modelljében a deriváció és az inflexió egyaránt a “lexikonba” kerül. Lexikon alatt itt egy több komponensből álló, visszacsatolást is tartalmazó komplexumot kell érteni, melynek részei: a szótöveket és affixumokat tartalmazó *morfémalista*, az ezekkel operáló *szóalakító komponens*, a *szűrő*, mely egyrészt megakadályozza a rosszulformált alakok továbbjutását, másrészt jólformált alakokhoz ad szabályokkal nem megfogható további információkat, végül pedig a *szótár*, mely kész szavakat tárol, melyek vagy valós időben épülnek fel morfémákból, és érkeznek ide a mondatfeldolgozás idejére, vagy hosszabb időre tárolódnak itt, így szükség esetén azonnal elérhetőek. Halle modellje tartalmaz egy mondattani komponenst, amely a szótárból veszi bemenetét, kimenete pedig egy fonológiai komponensen keresztül jut el hozzánk.

Jackendoff (1975), valamint Roeper és Siegel (1978) szóösszetételekről szóló írásainak áttekintése után ismertetem Lieber két munkáját: az 1981-es erős lexikalista rendszerét (és az ott alkalmazott háromrészes lexikonszerkezetet, mely szabályaival elvben minden morfológiai jelenséget képes kezelni morfémákból építkezve), valamint 1992-es könyvét, melyben visszatér a korábban felvázolt chomskyánus kiindulóponthoz, miszerint nincs elkülönülő morfológiai komponens a nyelvtanban, hanem a mondattan eszközrendszerét (művében a Kormányzás és Kötés eszközeit) kell képessé tenni arra, hogy a szublexikális szintet is elérje.

Az értekezés rámutat, hogy amennyiben a lexikalista megközelítést fogadjuk el, akkor vagy teljesen elválasztjuk a morfológiai folyamatokat a mondattantól, miközben a lexikon jelentős morfolexikai eszközrendszerrel bővül („erős lexikalizmus”), vagy a morfológiai feladatokat megosztjuk a lexikon és a mondattan között („gyenge lexikalizmus”). Harmadik útként, melyet Chomsky (1957) vezetett be, Lieber (1992) pedig megpróbált az értekezésben részletezett módon szisztematikusan megvalósítani, a morfológia beépül a mondattanba, így a lexikon szerepe is megváltozik. Ehhez a gondolathoz kapcsolódva a fejezet végén, külön szakaszt szentelve a témának, vázlatosan áttekintem a természetesnyelv-feldolgozásban meghonosodott morfológia-közelítéseket, releváns modellek és valós implementációk bemutatásával.

2.2 Poliszémia és szójelentés – a lexikai szemantika nézőpontja

A lexikont nem elég beillesztenünk a megfelelő nyelvmodellbe: szembe találjuk magunkat azzal a problémával is, hogy a lexikonban tárolt elemekhez *jelentést* szeretnénk rendelni, mégpedig minél többet abból, amit a nyelvi egyediségek osztályába sorolunk. Az értekezés nem vállalkozik a lexikai szemantika áttekintésére, hanem a potenciálisan felmerülő problémákat a poliszémia-monoszémia és a poliszémia-homonímia „különbségtételek" elemzésével szemlélteti. Cruse (1986, 2000) alapján a fent említett jelenség-párokat kontinuumként ismertetem, melyből következik a kérdés, hogy hogyan lehetséges a többjelentésű szavaknál a különböző olvasatok szétválasztása, mely későbbi kezelésük (ábrázolásuk, tárolásuk) alapja. A jelentéstani problémákkal foglalkozó fejezetben helyet kapott a prototípus-alapú kategorizálás és a metafora-elmélet tömör áttekintése, végül pedig az elméletet és gyakorlatot összekapcsoló szerzők véleményét ismertetem a szótárszerkesztés tradíciójáról (többek között Pustejovksy összefoglalását a jelentésfelsoroló lexikonok problémáiról és Verspoor megjegyzését arra vonatkozóan, hogy a hagyományos szótárszerkesztésben a homonímia maximalizálása és a poliszémia eliminálása figyelhető meg). Ennek kapcsán kiemelem, hogy a szótárakat emberi felhasználásra készítik, és a szótárforgató ember szempontjai nem feltétlenül egyeznek meg a számítógépes nyelvészek lexikonra vonatkozó preferenciáival. A természetesnyelv-feldolgozás szempontjait Verspoor (1997) gyűjti csokorba, és ez a szemlélet vezeti át az olvasót a következő fejezethez, melyben olyan adatbázisokat ismerünk meg, melyek jelentős potenciállal rendelkeznek a számítógépes nyelvfeldolgozásban, és amelyek a jelentés ábrázolásában fontos újításokat valósítottak meg.

2.3 WordNet, FrameNet és a számítógépes jelentésábrázolás

A szinonimacsoportokra építő WordNet adatbázis tervezési szempontjai közül kiemelem a lexikai relációkat implementáló egyedi szerkezetet és a rendkívül kis jelentéseltéréseket megkülönböztető (ugyanakkor poliszémiát nem kezelő) megközelítést. Kitérek arra, hogy a finom jelentésdistinkciók használata szakirodalmi adatok alapján megnehezíti a WordNet-alapú számítógépes nyelvészeti projektekben a szemantikai egyértelműsítést, ennek kapcsán pedig bemutatom Mihalcea és Moldovan (2001), Chen és Chang (1998) és Seagull (2000) megoldásait a szójelentések (olvasatok) WordNet-beli összevonására. Az adatbázisban felsorolt képzett szavakat számítógépes eszközökkel (erre a célra készített előfeldolgozó programmal és egy ugyancsak saját fejlesztésű konkordanciaprogrammal) vizsgálva megállapítom, hogy a WordNet adatbázisban jelentős számban szerepelnek képzett szavak, melyek becsült aránya (a vizsgált 12 képző vonatkozásában) korrelál a Baayen és Lieber

(1992) által ugyanezen képzők CELEX adatbázisban megfigyelt arányával. Nincs korreláció ugyanakkor a WordNet-beli képzett szavak becsült aránya és Baayen és Lieber (1992) által ezen képzőkhöz megadott produktivitási értékek alakulása között.

A FrameNet adatbázis bemutatása ugyan mellőzi azt az önálló szempontú vizsgálódást, ami a WordNet kapcsán megjelenik az értekezésben, ezzel együtt azt hiszem, hogy jól közvetíti azt az üzenetet, hogy egy szóalapú adatbázis túlmutathat a hagyományosan áruházott, jó esetben is legfeljebb a lexikai szemantika egy-két aspektusára kiterjedő jelentéstani szerepkörön. A FrameNet bemutatása azért is fontos, mert az értekezés későbbi részében ismertetett mesterséges neurális hálózat alapú szimuláció is FrameNet-szerű felcímkézést valósít meg. Egy harmadik adatbázis, a MindNet bemutatása közvetlen kapcsolatot teremt a szójelentéssel foglalkozó fejezet bizonyos részeivel, végül Véronis és Ide jelentéstároló rendszerének bemutatása vezeti át az olvasót a dolgozat második, konnekcionista részére.

2.4 Mesterséges neurális hálózatok a nyelvészetben

A nyelvészeti szakirodalomban fellelhető mesterséges neurális hálózat alapú kutatások közül McClelland és Rumelhart (1981) betűpercepciós kísérletét ismertetem először, mely a bemeneten megjelenő, ortografikus jegyeket tartalmazó inputvektornak szavakat feleltet meg a kimenetén. A bemeneti és a köztes rétegek kizárólag pontosan négy betűs szóhosszal képesek dolgozni, a kimeneten pedig –lokalista ábrázolást használva – minden potenciális szónak egy neuront feleltet meg a modell: a szavak megfelelő reprezentációja valójában a neurális hálózatos szimulációk egyik alapvető problémája.

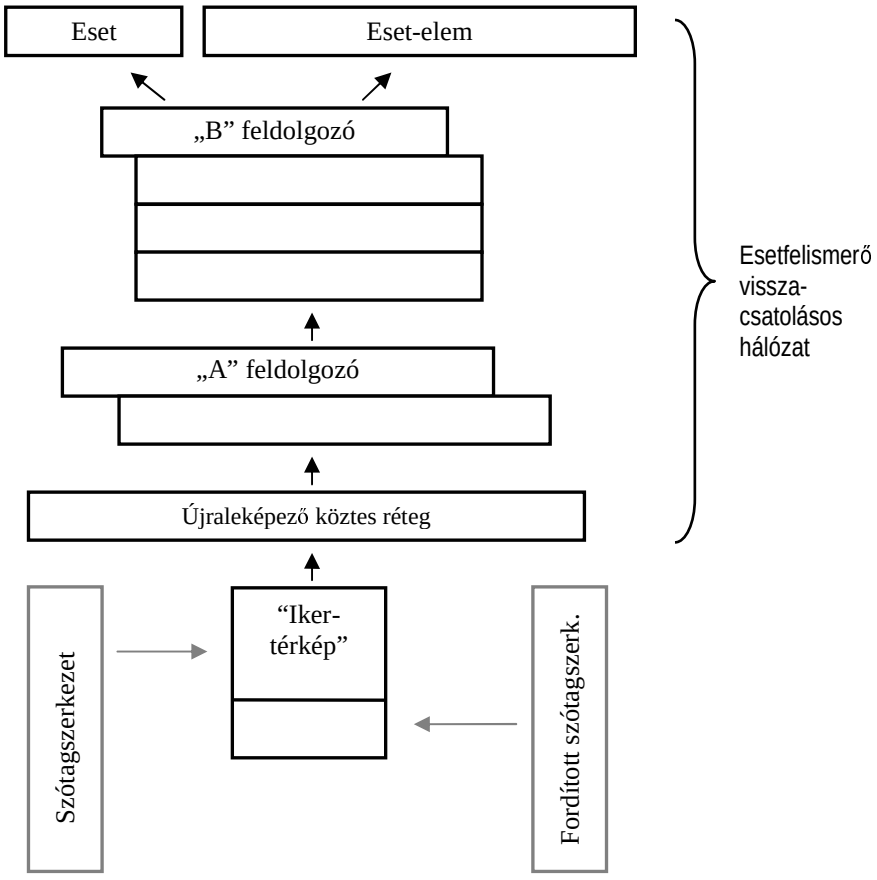
Elman (1990) kísérletsorozatának bemutatása több szempontból is fontos. Egyrészt azért, mert az Elman által bevezetett hálózatmodell („egyszerű visszacsatolásos hálózat”) frappáns megoldást nyújt *memória* kiépítésére egy mesterséges neurális hálózatban. Másrészt, nyelvészeti szempontból fontos megfigyelésként Elman megmutatja, hogy a betanító-korpuszból a hálózat által kinyert „statisztikák” („co-occurrence statistics”) olyan lexikai osztályokba szervezték a (sajnos csak néhány tucat elemből álló) input-szótárt, mint igék–főnevek és azon belül élő–élettelen főnevek. Elman (1990) alapján kiemelek néhány, a szakirodalomban azóta hagyományossá vált és általam is használt módszert (gondolva itt például az újraíró szabályokkal generált betanítókorpusz széleskörű használatára), melynek megpróbálom összegyűjteni az előnyeit és lehetséges buktatóit. A fejezet végén bemutatom Kohonen önszerveződő térképét, mely megfelelő változtatásokkal képes a nyelvi bemenet ábrázolására (az eredeti változatot Kohonen nem időbeli lefolyású folyamatok feldolgozására

tervezte). A saját mesterséges neurális hálózat alapú kísérletemről szóló fejezetben bemutatom, hogy egy szerény méretű önszerveződő térkép is, az általam javasolt konstrukcióban, akár több tízezer szó kielégítő hatékonyságú reprezentálására képes.

3. A mesterséges neurális hálózat alapú kísérletben alkalmazott módszerek ismertetése

Az értekezés 6. fejezetében bemutatom saját kísérletemet, melyben egy mesterséges neurális hálózatot hozok létre jelentéstani elemzésre: a bemeneti szavakhoz és összetevőkhöz FrameNet-szerű jelentéstani címkéket rendel a hálózat. A betanításra és tesztelésre nem természetes-nyelvi mintákat, hanem generált szövegeket használok. A tervezési szempontokat a tézisfüzet kutatási célokat összefoglaló 1. fejezete tartalmazza.

A hálózat több öntanuló térképet és módosított Elman-féle egyszerű visszacsatolásos hálózatot kombinál. A betanítást és a tesztelést Douglas Rohde programozható hálózatszimulátorában végeztem; a betanítókorpusz előállítása, valamint a teszteredmények értékelése saját fejlesztésű programokkal történt.



1. ábra

A hálózat szerkezetét az 1. ábra szemlélteti. A kísérleti modell általam ikertérképnek nevezett, két Kohonen-térképből álló, megfelelő bemenettel ellátott szerkezeti egységét kb. 20700 különböző szó felismerésére tanítottam be. Ezek a szavak 435 szótagtípust tartalmaztak. A két

„szótagszerkezet” bemeneti rétegen minden szótagtípusnak külön neuron felelt meg, melyeket a szótag szóbeli pozíciójának megfelelő aktivációs szintre állítottam be. A „fordított szótagszerkezet” rétegen a szó végéről számított pozíció határozta meg az aktivációs szintet. Ez a konstrukció egyrészt segíti annak a problémának a kiküszöbölését, hogy az ismétlődő szótagok bemeneti rétegenként csak egyszer jelenhetnek meg (tehát háromszori ismétlődés még az ikertérképes megoldással is zajhoz vezet), másrészt pedig a szóvégi szótagok előnyösebb súlyozása az ikertérkép fordított szótagsorrendhez tartozó részén lehetővé teszi hasonló végződésű szavak egymáshoz közeli elhelyezését a térkép adott részén. A betanítás végén minden egyes bemeneti szóhoz 5-5 neuron aktiválódik mindkét térképrészben, melyek együtt alkalmasak voltak arra, hogy a keretfelismerési kísérlethez generált korpuszban használt szavakat hiba nélkül, egyedileg azonosítsák. Mivel az ikertérképben az információsűrűség alacsony (a legtöbb neuron nyugalmi szinten marad), lehetséges az ikertérképek kombinálása: ez a módszer alkalmas összetett szavak, idiómák és akár konstituensek ábrázolására is; ezt a lehetőséget bizonyos mértékben ki is használtam a kísérletben.

A visszacsatolásos hálózatrész betanítására és tesztelésére 21610 mondatszekvenciát (egyenként 1-3 mondattal) generáltam 38 sablon segítségével. Összesen körülbelül 300 különböző szót használtam a kísérletben. A sablonokban elhelyezett nemterminális szimbólumok száma 41 volt, melyeket szavakkal vagy összetevőkkel (utóbbiak ikertérkép-kombinációkként jelentek meg a betanítókorpuszban) helyettesített az újraírást végző program. Összesen 268870 „esemény” (egy vagy több szót reprezentáló ikertérkép) került összesen 21 alkalommal ismétlődve a bemenetre egy néhány órás betanítási folyamat során, melynek célja az volt, hogy a hálózat megtanulja reprodukálni az összes eseményhez a megfelelő keret és keret-elem bélyegeket. Összesen 11 keretet (ebből kilencet a FrameNet-ből vettem, kettő saját volt) és 31 különböző keret-elemet használtam. Mivel mindössze 15 neuron (a „B feldolgozó” szerkezet legfelső rétege az 1. ábrán) volt felelős azért, hogy a 268870 esemény mindegyikéhez aktiválja a megfelelő keret és keret-elem neuront, így a hálózat rá volt kényszerítve, hogy felismerje a bemeneten megjelenő mondatszekvenciákban található szabályszerűségeket (például a 41 nemterminális szimbólumnak megfelelő ikertérkép-halmazokat). A hálózat közel 100 százalékos felidézést és pontosságot mutatott a keret és keret-elem beazonosításban.

Következő lépésben megvizsgáltam, hogy hogyan teljesít a hálózat, ha a 21610 mondatszekvencia 5 százalékát tanítom be, és a maradék kb. 20500 mondatsort csak

tesztelésre használom. A felidézés néhány százalékkal csökkent (így is 95% felett maradt), a pontosság pedig 99 százalékos volt.

Négy további kísérletet hajtottam végre a betanított hálózattal. Először a betanítókorpuszban 3 értelemben használt *crack* szó („betörni számítógépes rendszerbe” / „kiváló” / „(egy drogfajta)”) által keltett potenciális zavart próbáltam kimutatni. 10 olyan sablont választottam ki, mely tartalmazott olyan nemterminális szimbólumot, melyeket újraírva megjelenik az egyik *crack* homonima. Ezekkel először olyan mondatszekvenciákat generáltam, melyekben *crack* szerepelhet, majd a nemterminálisokhoz rendelt terminálishalmazok változtatásával olyan szekvenciákat is létrehoztam, amiből a *crack* alakot kizártam. Az értekezésben dokumentált adatokból, aktivációs diagramokból látszik, hogy a *crack* homonimái nem növelték a hibát, és nem okoztak felismerési bizonytalanságot.

Azt is vizsgáltam, hogy hogyan változik a teljesítmény akkor, ha a betanított hálózat olyan mondatsorokkal találkozik, amelyeket az eredeti 38 mondatszekvencia-sablontól eltérő templátumokkal generáltam. Az első ilyen kísérletben a mondatszekvenciákon belül a mondatok között kialakult kapcsolatot vizsgáltam. A kísérlethez kiválasztottam 2 olyan sablont, amely 2-2 mondatot tartalmaz. Amikor a sablonok *második mondatát* előállító részeket töröltem, akkor az immáron csak az első mondatokat tartalmazó „szekvenciák” önmagukban ugyanolyan eredményt értek el, mint a másik mondat társaságában. Amikor az *első mondatot* hagytam el a sablonból, szintén változatlan maradt a teszteredmény. Felvetődik tehát, hogy a felsorolt tervezési célokkal ellentétben a hálózat különállóként kezelte-e a szekvenciák mondatait. Amikor *felcseréltem* a sablonokban a két mondatot, a hiba jól mérhetően nőtt, és ezzel a tervezési szempontoknak megfelelő eredményt kaptam.

A következő sablonmanipulációs kísérletben egy-egy összetevőért felelős nemterminális szimbólumot töröltem, és figyeltem az így generált megrövidült szekvenciák viselkedését. A felidézés közepes mértékben (9-12 százalékkal), a pontosság ennél kevesebbel (0-7 százalék) csökkent.

Az utolsó kísérletben felhasznált sablonokban 1-1 nemterminális szimbólumot lecseréltem egy másik nemterminálisra az alább vázolt módokon, és megvizsgáltam, hogy az általuk generált mondatsorokat milyen hibával címkézi fel a betanított hálózat. A kísérletsor áttekintéséhez itt jegyzem meg, hogy a felhasznált 38 sablon 3 szituációs helyzethez kötődött: a) számítógépes bűnözés, b) drogforgalmazás és -fogyasztás, c) emberek közötti kommunikáció. A sablonokban lecserélt nemterminális szimbólumot a következőkkel helyettesítettem:

a) ismeretlen szavakat generáló nemterminális szimbólum (a vártnak megfelelően ez okozta a legnagyobb teljesítménycsökkenést),
b) ugyanebben a szituációs helyzetben (de más sablonokban) felhasznált nemterminális szimbólum,
c) más szituációhoz tartozó sablonokban felhasznált nemterminális: ez a b) esethez hasonló mértékben növelte a hibát, és végül
d) az eredeti szavak többesszámát generáló nemterminális szimbólum: ez alig okozott mérhető változást a keret- és keret-elem felcímkézésben.

4. Következtetések, eredmények

A fent bemutatott mesterséges neurális hálózatos kísérletre az értekezésben a következő megállapításokat teszem:

1) Az ikertérképes bemeneti interfész képes nagyszámú szóalak automatikus megkülönböztetésére, sőt akár több szó kombinálására egy szimulációs eseménnyé, így lehetőséget kínál a nyelvi bemenet ábrázolására, melynek megoldása a neurális hálózatos kísérletek visszatérő problémája.
2) A visszacsatolásos hálózatokból az értekezésben ismertetett módon összeállított struktúra megfelelő általánosítási képességgel és memória-potenciállal rendelkezik a kitűzött feladat végrehajtására: mondatszekvenciák (néhány mondatos szövegrészletek) jelentéstani címkékkel történő annotálására. Ezt a generált korpusz 5 százalékán betanítva (kb. 1070 példa alapján) is nagy pontossággal elérte a hálózat.
3) A *crack* homonimáit vizsgáló kísérlet alapján részben megelőlegezhető, hogy az azonos alakú szavak kezelése ilyen típusú feladat hasonló módszerekkel történő megoldásában kivitelezhető szemantikai egyértelműsítés, előzetes szófaji vagy egyéb felcímkézés nélkül.
4) A sablonmanipulációs kísérletek bizonyítják, hogy a modell képes kompenzálni a hiányos bemenetet (hiányzó mondatok és összetevők), vagy az egyéb jellegű zajt (lecserélt összetevők). A lecserélt összetevőket vizsgáló kísérlet megmutatta, hogy a manipulált bemenet felcímkézése nem csupán a kontextus érintetlenül maradt elemei alapján történt, hanem a cseréként beállított összetevők (a velük betanítás során asszociált információdarabkák) is meghatározták annak sikerét. Így az egyesszámú szavak lecserélése a nekik megfelelő többesszámú alakokra kevés problémát okozott, míg például a teljesen ismeretlen szavak felcímkézése nagyobb hibát eredményezett.

Az ismertetett kísérleti modell közvetlenül nem használható arra, hogy természetes nyelvi szövegeket címkézzen fel, így jelenleg a nyelvi alapkutatás kategóriájába sorolható. Alkalmazott kutatássá válhat, ha megjelennek azok a FrameNet címkékkel annotált korpuszok, amelyekre támaszkodva lehetségessé válik olyan modell kialakítása, amely képes legalábbis részleges jelentéstani felcímkézésre (például néhány előre meghatározott keret beazonosítására).

Az értekezésben az elméleti áttekintő fejezetek és az ismertetett kísérlet értékelése alapján kiemelem, hogy bizonyos konnekcionista megoldások, megfelelő feladatválasztással párosítva, potenciálisan szükségtelenné tehetnek több olyan eljárást, amelyek a nyelvészek számára elméleti, a számítógépes nyelvészetben pedig gyakorlati, mindennapos kihívást okoznak, beleértve a szófaji és mondattani felcímkézést és a szemantikai egyértelműsítést. A szakirodalmi előzmények azt bizonyítják, hogy a megfelelően konstruált mesterséges neurális hálózatok képesek a bemeneten megjelenő információk közül automatikusan kiválasztani azokat, amelyek a feladat megoldásához szükségesek, és megtalálni a szisztematikus ismétlődéseket, noha ezek belső ábrázolása nem transzparens abban az értelemben, hogy a kialakuló, akár a teljes hálózaton szétszóródó aktivációs mintázatok nem feltétlenül feleltethetőek meg azoknak a kategóriáknak, amelyeket a hálózat által elsajátított jelenségek leírására hagyományosan használunk. Ezért a szavakkal is operáló neurális hálózati modellekben, hacsak nem teszünk intézkedéseket a tervezés és a betanítás során lokalista információábrázolásra, a szóalakokhoz kötődő információk egy „szórványlexikonban” manifesztálódnak, neuronok közti súlyokká és aktivációs szintekké változnak.

A doktori tézisfüzetben hivatkozott irodalom

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: A corpus based study. *Linguistics, 29,* 801-843.

Chen, J. N., & Chang, J. S. (1998). Topical clustering of MRD senses based on information retrieval techniques. *Computational Linguistics, 24(1).*

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of Syntax*. Cambridge, MA: The MIT Press.

Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs, & P. Rosenbaum (szerk.), *Readings in English Transformational Grammar*. Waltham, MA: Gin and Company.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Cruse, D. A. (2000). *Meaning in language.* Oxford: Oxford University Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Halle, M. (1973). Prelogomena to a theory of word formation. *Lingusitic Inquiry, 4(1).*

Jackendoff, R. (1975). Morphological and semantic regularities in the lexicon. *Language, 51(3),* 639-671.

Lieber, R. (1981). *On the organization of the lexicon*. Bloomington, IN: Indiana University Linguistics Club.

Lieber, R. (1992). *Deconstructing Morphology*. Chicago and London: The University of Chicago Press.

McClelland, J. L., & Rumelhart, D. E. (1981). An Interactive Activation Model of Context Effects in Letter Perception. Part 1: An Account of Basic Findings. *Psychological Review, 88*, 375-407.

Mihalcea, R., & Moldovan, D. (2001). EZ.WordNet: Principles for automatic generation of a coarse grained WordNet. *Proceedings of FLAIRS 2001*, 454-459, Key West, FL.

Roeper, T., & Siegel, M. (1978). A lexical transformation for verbal compounds. *Linguistic Inquiry, 9(2).*

Seagull, A. (2000). *A Compaction of WordNet Senses for Evaluation of Word Sense Disambiguators. TR726.* Computer Science Department, University of Rochester.

Verspoor, C. M. (1997). *Contextually-Dependent Lexical Semantics*. PhD thesis. Edinburgh: The University of Edinburgh.

A szerző értekezés tárgyához kötődő publikációi és előadásai

Megjelent közlemények

(2000). Szóhálózat. *Magyar Tudomány, 10*, 1235-1237.

(2002). Derived nouns in WordNet and the Question of Productivity. *Studies in Linguistics, 6,* 433-449.

(2003). Derived Nouns and Gerunds in WordNet. In M. Fabian (szerk.), *Sučasni doslidženńa z inozemnoï filologiï. Volume 1.* 149-154. Uzhhorod: Department of Foreign languages, Uzhhorod National University.

(2004). Polysemy and Homonymy in WordNet. In M. Fabian (szerk.), *Sučasni doslidženńa z inozemnoï filologiï. Volume 2.* 28-36. Uzhhorod: Department of Foreign languages, Uzhhorod National University.

(2005). Form and Meaning in a Recurrent Neural Network. In M. Fabian (szerk.), *Sučasni doslidženńa z inozemnoï filologiï. Volume 3.* 22-31. Uzhhorod: Department of Foreign languages, Uzhhorod National University.

Recenzió

(2002). Lawler, J., & Dry, H. (szerk.) (1998). *Using Computers in Linguistics: A practical guide*. Studies in Linguistics, 6, 490-493.

Konferencián elhangzott előadások

(2003). HUSSE VI, Debrecen: *WordNet and the Lexicon*

(2005). HUSSE VII, Veszprém: *Form and Meaning in an Artificial Neural Network*