Lemmaasszociáció és morfológiai jegyek mesterséges neurális hálózatokban

Tóth Ágoston[1], Csernyi Gábor[1]

[1] Debreceni Egyetem, Angol Nyelvészeti Tanszék
{toth.agoston, gabor.csernyi}@arts.unideb.hu

1 Bevezetés

Kutatásunk célja egy lemmatizálást és korlátozott morfológiai elemzést mintaasszociáció segítségével megvalósító mesterséges neurális hálózat implementálása, továbbá a neurális modellezés erősségeinek és nehézségeinek dokumentálása.

2 A kísérleteink

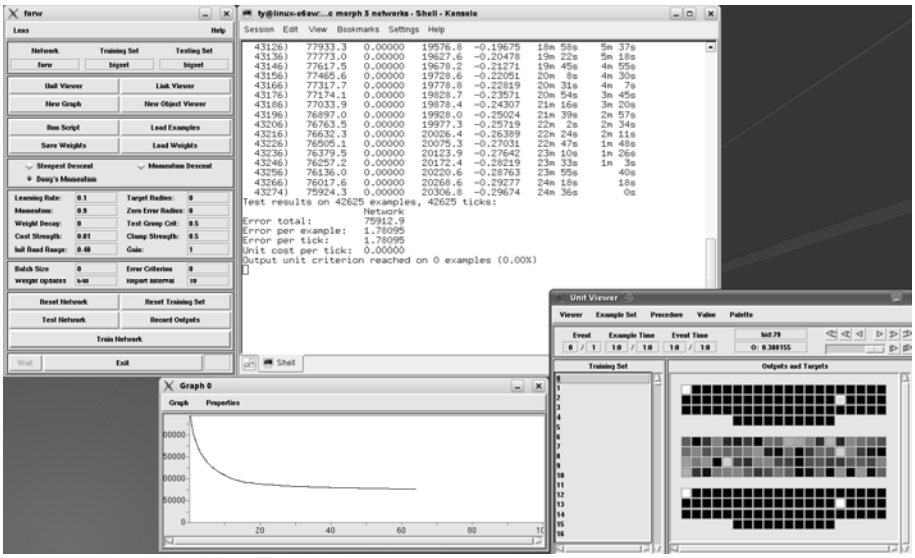
A kísérletekhez használt tanító adatokat a *Magyar Webkorpusz* [1] 100000 leggyakoribb szóalakját tartalmazó listáról nyertük, melyet feldolgozás előtt szűrtünk. Az így előállt, körülbelül 82 ezer szavas szólista 63531 elemére adott a *Hunmorph* [4] legalább egy elemzést. A szóalakokhoz az elemzés során kapott lemmát, valamint kiválasztott (egyelőre korlátozott számú) morfológiai jegyet tanítottunk be.

A kísérleteket neurális hálózatokkal végeztük. A bemeneti rétegen (70 neuron) szóalakokat helyeztünk el egy első alkalommal felhasznált szóreprezentációs technikát használva. Az aktivációk innen egy rejtett rétegbe (80 neuron) haladtak tovább tanítható, súlyozott kapcsolatokat használva, 1:N projekcióval. A rejtett rétegből hasonlóan kialakított kapcsolatok vezettek a kimeneti réteghez, ahol egyrészt 70 neuron végezte a szóalakkal asszociált lemma reprezentációját ugyanazzal a módszerrel, amivel a bemenetet kezeltük (elméletileg végtelen számú szó ábrázolását lehetővé téve), másrészt bizonyos mennyiségű, alapvető morfológiai információkat ábrázoló neuronokat is betanítottunk, az adott kísérlet függvényében. A tanítás a „visszafelé terjesztés" módszerével történt (minden bemenetre képeztük az aktuális súlyokat használva a kimeneteket, kiszámítottuk a teljes hibát, majd a hibát visszafelé terjesztve módosítottuk a súlyokat).

Minden minta (szóalak-lemma pár) legalább 650 alkalommal került betanításra. A bemeneteken és a kimeneteken [0;1] intervallumba eső valós értékek jelentek meg. A kimeneten mind a lemmát, mind a morfológiai jegyeket osztályoztuk a következő módon: a 70 valós értékből álló lemma-kimenetet a legközelebbi ismert lemma célvektornak feleltettük meg, a morfológiai jegyeket pedig 0,4 kimeneti érték alatt 0-nak (jegy hiánya), 0,4-től pedig 1-nek (jegy megléte) osztályoztuk.

A betanítást és a tesztelést a LENS neurális hálózat szimulátorban végeztük [2]. Az 1. ábrán példaként egy hálózat betanításának szimulációs eredményét mutatjuk be, amelyen alul, balra megfigyelhető a hibadiagram, a jobb alsó sarokban pedig a betaní-

tási és tesztelési minták egyenkénti vizsgálatára alkalmas „unit viewer” ablakban az első mintára (az *a* határozott névelőre) kapott aktivációs szintek (alul a bemeneti csoport, fölötte a 80 neuronos „rejtett” réteg, felettük a kimenetek).



1. ábra: LENS képernyőfotó.

Fontos kiemelni, hogy az itt bemutatott kísérleteinkben a többértelműség (az alternatív alaktani elemzések) kezelése komoly problémát okozott már a tervezés fázisától kezdve. Adott keretek közt alternatívák betanítása nem lehetséges, hiszen egy alternatíva jelenléte (azonos inputra különböző kimeneti célok) a betanítást elrontja. Természetesen a valóságban a környezet különbözősége jelenti azt az információt, ami alapján az egyértelműsítés elvégezhető. A morfológiai elemzés szokásos, véges állapotú automatákat használó változata olyan kimenetet ad, amiben az alternatívák mind megjelennek, és egy későbbi mondattani elemzés során ez vagy egyértelműsíthető, vagy további elemzések bevezetéséhez vezet (és ekkor a problémát tovább delegáljuk a szemantikai szintre). A többértelműség kezelésében azonban nem feltétlenül jelent megoldást az összes elemzés visszaadása egy későbbi egyértelműsítés reményében (ahogyan azt a lexikai szemantika vonatkozásában a SenseEval/SemEval versenyekben láthattuk). Éppen ezért a későbbiekben sem az alternatívák enumerációja, hanem a figyelembe vehető paraméterek bővítése (pl. a mondatban szereplő további szavak, morfémák figyelembevétele) és ezek alapján egyértelmű kimenet előállítása a hosszú távú célunk. Jelen rendszerünket úgy terveztük, hogy szófajonként egy elemzést tudunk kezelni; ha egy szó Hunmorph-os elemzése ennek nem felelt meg, akkor kizártuk a kísérletből. Ezen a szűrőn 42625 szóalak ment át, ami a Hunmorph által összesen elemzett 63531 alak 67%-a (ez egyben a felidézési érték, amely mellett rendszerünk Hunmorph-hoz viszonyított pontossága értendő).

A bemeneten megjelenő szóalakok és a kimeneten elvárt lemmák reprezentálására olyan vektorokat képzünk, amelyben az ABC minden betűjének két vektorelem felel

meg. Az egyik azt mutatja meg, hogy az adott betű a szó hányadik karakterpozícióján fordul elő *először*, a másik pedig azt, hogy az adott betű a szó (szó végétől számítva) hányadik karakterpozíción fordul elő *utoljára*. Ha egy szóban egy betű kettőnél többször szerepel, ami nem ritka jelenség, akkor az adott betű első és utolsó előfordulásának helye lesz rögzítve, a többiről nem tárolunk információt. A módszert Tóth [3] javasolta, ahol több reprezentációs eljárás is szerepel, és a módszerek előzetes tesztelését angol írott, angol fonetikusan átírt és magyar szavakon végezte el. Az ottani kísérletekből látszik, hogy a betűk *utolsó* előfordulásának jegyzése önmagában is nagyon hatásos eszköz egy szó felismerésében, de egy további adat (itt: az első előfordulások felhasználása) fokozza az eljárás pontosságát. Ezek a módszerek nem kölcsönösen egyértelmű leképezéseket valósítanak meg, de ha ez az adott felhasználáshoz szükséges, akkor is rendkívül alacsony a hiba. Mostani kísérletünkben 23 olyan szópár volt, melyek olyan szavakból álltak, amelyeknek reprezentációja azonos volt. Ez a jelenség a vizsgált 42625 szónak kevesebb mint 1 ezrelékét érintette, ezért nem tekintettük jelentős hibaforrásnak, és ezeket a szavakat is megtartottuk.

Első kísérletünkben a szófaji felismerést mértük, miközben a kimeneten a lemmát leíró egységek teljesítményét nem figyeltük. A *főnév* jegyet 82%, az *igét* 90%, a *melléknevet* 84%, a *határozószót* 96%, az *egyéb* kategóriát (*névelő*, *kötőszó*, *számnév*, stb.) 97% pontossággal jelezte a rendszer a 42625 szavas szólistán mérve.

Második kísérletünkben öt hálózatot tanítottunk be, ezek sorrendben a főneveket, igéket, mellékneveket, határozókat és végül az egyéb morfológiai kategóriákat kezelték, és *alak–lemma*, valamint *alak–morfológiai jegy* asszociációt végeztek úgy, hogy bemenetükön a szóalakok, a kimenetükön pedig a lemmák és morfológiai jegyek voltak ábrázolva. A *főnévi* hálózat esetében a figyelt jegyek (gyakoriságuk alapján kiválasztva) a többes szám, a birtokos eset és a tárgyeset, az *igei* hálózatban a többes szám, a múlt idő, az 1. és 2. személy, valamint a tárgyas ragozás voltak; a *melléknéveknél* a többes számot vizsgáltuk, a *határozószóknál* nem volt megfigyelt jegy. Az *egyéb* kategóriában (5. hálózat) a Hunmorph további főkategóriáit (*névelő*, *kötőszó*, *számnév* stb., összesen 9 db) azonosítottuk 1-1 neuronnal. Amennyiben a bemeneten megjelent szóalaknak nem volt az adott hálózatnak megfelelő kategóriájú elemzése, a kimeneten a „lemmahiány" lemma megjelenését vártuk, a lemma neuronok egyedi mintázatát figyelve (tehát szintén lemmaasszociációs feladatként); a morfológiai kimenetek ekkor inaktívak voltak. A hálózatokon mért pontosságot az 1-5. táblázatokban foglaltuk össze.

1. táblázat: A főnévi hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány" (= inkompatibilis kat.)	15528	12667	82%
helyes lemma (kivéve: „lemmahiány") (baseline: 1:8297 ≈ 0,01%)	27097	18818	69%
lemmaasszoc. összesen	42625	31486	*74%*
morfológia (27097 főnévre)			*87%-97%*

2. táblázat: Az igei hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány” (= inkompatibilis kat.)	32393	31716	98%
helyes lemma (kivéve: „lemmahiány”) (baseline: 1:3102 ≈ 0,03%)	10232	5204	51%
lemmaasszoc. összesen	42625	36920	*87%*
morfológia (10232 igére)			*94%-97%*

3. táblázat: A melléknévi hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány” (= inkompatibilis kat.)	32533	31830	98%
helyes lemma (kivéve: „lemmahiány”) (baseline: 1:6325 ≈ 0,02%)	10092	3675	36%
lemmaasszoc. összesen	42625	35505	*83%*
morfológia (1 jegy, 10092 melléknév)			*91%*

4. táblázat: A határozói hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány” (= inkompatibilis kat.)	40448	40380	99%
helyes lemma (kivéve: „lemmahiány”) (baseline: 1:2079 ≈ 0,05%)	2177	233	11%
lemmaasszoc. összesen	42625	40613	*95%*

5. táblázat: Az „egyéb” hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány” (= inkompatibilis kat.)	41554	41554	100%
helyes lemma (kivéve: „lemmahiány”) (baseline: 1:678 ≈ 0,1%)	1071	8	1%
lemmaasszoc. összesen	42625	41562	*98%*
morfológia (1071 szóalakra)			*80%-99%*

A hálózatok a nem kompatibilis kategóriát, „lemmahiány” lemmát visszaadva, 82-100% pontossággal jelezték. Helyes kategóriájú alak esetén a legközelebbi lemmát 1-69% közötti pontossággal adták vissza. A gyakoribb kategóriák esetén a (létező szavakra utaló) lemmaasszociáció pontossága magasabb volt, lásd a főnévi hálózat adatait. Az adatokból az is látható, hogy a baseline értéket (ami az adott hálózat lemma kimenetén várt összes *különböző* lemmareprezentáció mennyiségével fordítottan arányos) mindegyik hálózat esetében sikerült jelentősen meghaladni. A *határozószó* és *egyéb* kategóriák nagyon kevés alakkal voltak képviselve, az elért alacsony pontosság ennek is köszönhető, ilyenkor azonban a morfológiai inkompatibilist jelző „lemmahiány” állapot visszaadása igen pontos volt. A figyelt morfológiai jegyeket (pl. többes szám, birtokos eset, tárgyeset stb.) meglehetősen jó eredménnyel jelezték a hálózatok, adott jegytől függően tartalmi szavaknál 87-97%, funkciószavaknál 80-

99% pontossággal. További kísérletekben a jegyek köre bővíthető, a skálázhatóság egyelőre nem ismert.

Utolsó kísérletünkben a mintákat véletlenszerűen, ¾ részben tanító és ¼ részben tesztelő adathalmazra osztottuk. A főnévi hálózatot a tanító mintákkal betanítottuk, majd a tesztmintákkal (melyeket a hálózat nem ismert) kiértékeltük. A főnévi elemzések esetén a lemma kimenet 71%, az inkompatibilis kategória („lemmahiány") jelzése pedig 80% pontossággal zajlott, összességében a lemmaasszociáció 74%-ban volt sikeres. A három megfigyelt főnévi morfológiai jegyet 86-96% pontossággal becsülte a rendszer, jegytől függően. Ezeket az adatokat az 1. táblázat főnévi oszlopával összevetve láthatjuk, hogy a hálózat általánosító képessége mind a lemmaasszociáció, mind a morfológiai jegyek tekintetében igen jó (a tesztadatokon mért teljesítmény semmiben sem marad el a tanítón mért pontosságtól), tehát kijelenthetjük, hogy nem a konkrét alakokat, hanem a szabályszerűségeket tanulta meg a hálózat.

Köszönetnyilvánítás

A publikáció elkészítését részben az OTKA (K 72983), részben a TÁMOP 4.2.1./B-09/1/KONV-2010-0007 számú projekt támogatta az Új Magyarország Fejlesztési Terven keresztül az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával, továbbá támogatta a TÁMOP-4.2.2/B-10/1-2010-0024 projekt az Európai Unió és az Európai Szociális Alap társfinanszírozásával.

Bibliográfia

1. Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In: Kilgarriff, A., Baroni M. (eds.): Proceedings of the 2nd International Workshop on Web as Corpus (2006)
2. Rohde, D. L. T.: LENS: The light, efficient network simulator. Technical Report CMU-CS-99-164. Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA (1999)
3. Tóth, Á.: Perspectives on the Lexicon. Akadémiai Kiadó, Budapest (2008)
4. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, G., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software (2005)