Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra

Siklósi Borbála[1], Novák Attila[1,2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
[2] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport,
1083 Budapest, Práter utca 50/a
e-mail:{siklosi.borbala,novak.attila}@itk.ppke.hu

Kivonat A neurálishálózat-alapú szemantikai beágyazási modelleket létrehozó algoritmusok a diszrtibúciós szemantika egy viszonylag új, de egyre népszerűbb alkalmazási területe. A szavakhoz vagy kifejezésekhez rendelt folytonos reprezentációk azok jelentését jól reprezentálják angol nyelvű tanítóanyagok esetén. Cikkünkben arra vonatkozó vizsgálatokat mutatunk be, hogy magyar nyelvre mennyire használhatóak ezek a modellek, illetve egy konkrét kategorizációs feladatban is kiértékeljük ezek hatékonyságát.

1. Bevezetés

A szavak reprezentációjának meghatározása a nyelvtechnológiai alkalmazások számára alapvető feladat. A kérdés az, hogy milyen reprezentáció az, ami a szavak jelentését, vagy azok morfoszintaktikai, szintaktikai viselkedését is meg tudja határozni. Angol nyelvre egyre népszerűbb a kézzel gyártott szimbolikus és nyers szövegből tanulható ritka diszkrét reprezentációk helyett a folytonos vektorreprezentációk alkalmazása, melyek hatékonyságát a neurális hálózatokra alapuló implementációk használatával több tanulmány is alátámasztotta [5,8,2]. Ezekben a kísérletekben és alkalmazásokban azonban a leírt módszereket általában egy a magyarhoz képest jóval kevesebb szóalakváltozattal operáló, kötött szórendű és egyszerű szószerkezeteket használó nyelvre alkalmazzák.

Cikkünk célja a folytonos reprezentációt implementáló modellek használhatóságának és hatékonyságának vizsgálata magyar nyelvre.

Vizsgálatunk motivációja azonban kettős. Egyik célunk a módszer szemantikai érzékenységének felderítése, azaz, hogy mennyire alkalmas arra, hogy magyar nyelvű korpuszon tanítva a szavakat a szemantikai térben konzisztensen helyezze el. Másrészt pedig egy konkrét alkalmazás támogatása is a célok között szerepelt: egy morfológiai elemző adatbázisának kiegészítése olyan szemantikai jegyekkel, amelyek hatással vannak a szavak morfológiai, helyesírási, illetve szintaktikai viselkedésére. Ilyenek például a színek, anyagnevek, népnevek, nyelvek, foglalkozások, stb. Ezek kézzel való összegyűjtése és az adatbázishoz való hozzáadása igen idő- és munkaigényes feladat, ezért ennek a feladatnak az automatizálása szintén céljaink között szerepelt, kísérleteink egy része ezeknek a szemantikai csoportoknak a létrehozására ad módszert.

2. Folytonos disztribúciós szemantikai modellek

A disztribúciós szemantika lényege, hogy a szavak jelentése szorosan összefügg azzal, hogy milyen kontextusban használjuk őket. A hagyományos disztribúciós szemantikai modellek létrehozásakor az egyes szavak előre meghatározott méretű környezetét az azokban előforduló szavak nagy korpuszból számított előfordulási statisztikái alapján határozzuk meg.

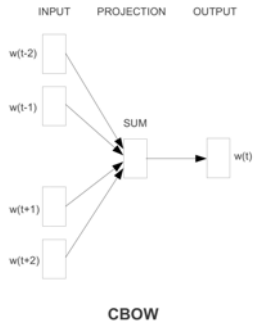
Ezzel szemben a nyelvtechnológiai kutatások egyik kurrens módszere a folytonos vektoros reprezentációk alkalmazása (*word embedding*), melyek nyers szöveges korpuszból szemantikai információk kinyerésére alkalmazhatók. Ebben a rendszerben a lexikai elemek egy valós vektortér egyes pontjai, melyek konzisztensen helyezkednek el az adott térben, azaz, az egymáshoz szemantikailag és/vagy morfológiailag hasonló szavak egymáshoz közel, a jelentésben eltérő elemek egymástól távol esnek. Mindemellett, a vektoralgebrai műveletek is alkalmazhatók ebben a térben, tehát két elem szemantikai hasonlósága a két vektor távolságaként meghatározható, illetve a lexikai elemek pozícióját reprezentáló vektorok összege, azok jelentésbeli összegét határozzák meg [8,6]. A módszer hátránya csupán az, hogy önmagában nem képes a poliszémia, illetve homonímia kezelésére, tehát egy többjelentésű lexikai elemhez is csupán egyetlen jelentésvektort rendel, azonban a szakirodalomban erre a problémára is találunk sikerrel alkalmazott módszereket [1,3,10].

Ennek a modellnek a tanítása során is az egyes szavak fix méretű környezetét vesszük figyelembe, az ezekből álló vektor azonban egy neurális hálózat bemenete. A környezetet reprezentáló vektorok összegét használja a hálózat arra, hogy megjósolja a célszót. A tanítás során a hiba visszaterjesztésével és ennek megfelelően a környezetet reprezentáló vektorok frissítésével jön létre a célszót helyesen megjósoló súlyvektor, ami a neurális hálózat megfelelő rétegéből közvetlenül kinyerhető. Mivel a hasonló szavak hasonló környezetben fordulnak elő, ezért a szövegkörnyezetre optimalizált vektorok a hasonló jelentésű szavak esetén hasonlóak lesznek. Az erre a feladatra felépített neurális hálózat a CBOW (*continous bag-of-words*) modellt implementálja, ami az 1. ábrán látható. Egy másik lehetőség az ún. skip-gram modell alkalmazása, amikor a hálózat bemenete a célszó, az optimalizálás célja pedig e szó környezetének megjósolása.

3. Kísérletek

A kísérleteinkben használt modelleket a `word2vec`[3] eszközzel hoztuk létre, ami mind a CBOW, mind a skip-gram modellek implementációját tartalmazza és a lexikai elemeket reprezentáló vektorok közvetlenül kinyerhetőek belőle. Mivel a két modell közül a CBOW modell betanítása hatékonyabb nagy tanítókorpuszok esetén, ezért mindegyik tanítás során ezt alkalmaztuk. Tanítóanyagként pedig egy majdnem 4 milliárd szavas magyar nyelvű webkorpuszt használtunk. Minden modell esetén 300 dimenziós vektorokat definiáltunk a lexikai elemek

[3] https://code.google.com/p/word2vec/



1. ábra. A CBOW (*continous bag-of-words*) modell

reprezentálására és 5 token sugarú mintavételezési ablakot a szövegkörnyezet kinyerésére.

3.1. Nyers szövegen tanított modell

Először egy a korpusz nyers változatán tanított modellt hoztunk létre (`SURF`), ami a szavak felszíni alakját reprezentáló vektorokat határozott meg, így az azonos tőhöz tartozó különböző ragozott alakok külön pozícióba kerültek a szemantikai térben. Ez a modell tehát a különböző morfológiai analógiák felderítésére használható. Például a *jó – rossz* és a *jobb – rosszabb* szópárok hasonlósága sokkal erősebb, mintha az azonos tő szerint hasonlítjuk őket össze (*jó – jobb*, illetve *rossz – rosszabb*). Ez a modell tehát jól reprezentálja a szemantikai és szintaktikai hasonlóságot. Néhány további példa az ebben a modellben az egy-egy szóhoz legközelebb álló szavakra a 1. táblázatban látható. A példákban a szavak melletti számok a korpuszbeli előfordulások számát adják meg.

3.2. Előfeldolgozott szövegen tanított modell

A másik modellben a korpusz szófaji egyértelműsített változatát használtuk oly módon, hogy a szavak lemmáját tartottuk meg, melyek után, külön tokenként szerepeltek a morfológiai elemző által generált címkék `ANA`. Mivel ezek a címkék az aktuális szó környezetében megmaradtak, ezért az általuk reprezentált szintaktikai információ továbbra is szerepet kapott az egyes szavakat reprezentáló vektorok létrehozásában, azonban a modell csak lemmákat tartalmazott, így robusztusabb modell jött létre az adatritkaság csökkenése miatt. A 2. táblázat néhány példát tartalmaz az ezzel a modellel kapott hasonlósági listákra. Látható, hogy a modell rangsorolása jól működik a szavak gyakoriságától függetlenül, hiszen a nagyon gyakori szavak nem előzik meg a szemantikailag jobban hasonló kifejezéseket.

1. táblázat. Példák a nyers szövegből kinyert modellek alapján kapott hasonló kifejezésekre. A zárójeles számok a korpuszbeli előfordulások számát mutatják.

kenyerek	pirosas	egerekkel	fiaik	megeszi
kiflik(349)	lilás(2476)	patkányokkal(524)	lányaik(593)	eszi(12615)
zsemlék(283)	rózsaszínes(1638)	férgekkel(513)	leányaik(251)	megenné(563)
lepények(202)	barnás(6463)	majmokkal(606)	férjeik(759)	elfogyasztja(1129)
pogácsák(539)	sárgás(7365)	hangyákkal(343)	gyermekeik(12028)	megeszik(6433)
pékáruk(771)	zöldes(5215)	nyulakkal(366)	feleségeik(638)	Megeszi(189)
péksütemények(997)	fehéres(2517)	legyekkel(252)	gyerekeik(5806)	megette(7868)
sonkák(613)	vöröses(5496)	rágcsálókkal(259)	asszonyaik(458)	megrágja(477)
tészták(2466)	feketés(1157)	hüllőkkel(241)	gyermekei(31241)	megeheti(287)
kalácsok(277)	narancssárgás(429)	pókokkal(436)	fiak(1523)	bekapja(977)
kekszek(1046)	sárgászöld(723)	bogarakkal(425)	unokái(3528)	lenyeli(1862)

2. táblázat. Példák a tövesített és elemzett szövegből kinyert modellek alapján kapott hasonló kifejezésekre. A zárójeles számok a korpuszbeli előfordulások számát adják meg.

kenyér	eszik	csavargó	csónak	franciakulcs
hús(136814)	iszik(244247)	koldus(15793)	tutaj(3950)	feszítővas(846)
kalács(10658)	főz(120634)	zsivány(3497)	ladik(3895)	csípőfogó(345)
rizs(31678)	csinál(1194585)	haramia(2024)	motorcsónak(4079)	csavarkulcs(473)
zsemle(6690)	megeszik(68347)	vadember(2497)	hajó(238807)	kisbalta(491)
pogácsa(11066)	fogyaszt(160724)	csirkefogó(2019)	kenu(6649)	konyhakés(1501)
sajt(46660)	etet(43539)	szatír(1649)	kocsi(283438)	pajszer(567)
kifli(9715)	zabál(13699)	útonálló(1942)	gumicsónak(1033)	partvis(648)
krumpli(37271)	megiszik(31002)	bandita(6334)	mentőcsónak(2511)	villáskulcs(764)
búzakenyér(306)	eszeget(3928)	suhanc(4144)	dereglye(962)	erővágó(360)
tej(113911)	alszik(359268)	vándor(14070)	sikló(4394)	péklapát(475)

3.3. Helyesírási hibák és nem sztenderd szóalakok

A modell vizsgálata során fény derült arra is, hogy a jelentésben hasonló szavak között megjelentek a különböző elírt változatok is. Ezek adták az ötletet arra, hogy olyan szóalakokhoz tartozó listákat is lekérdezzünk, melyek eleve hibásak. Ebben az esetben olyan szóalakokat kaptunk eredményül, melyek ugyanolyan vagy hasonló jellegű helyesírási hibát tartalmaznak, vagy amiket a lemmatizáló ugyanúgy rontott el, ugyanakkor ezekben a listákban is érvényesül a szemantikai rangsor. A 3. táblázat első két oszlopa ilyen példákat tartalmaz. A rendszernek ez a képessége jól hasznosítható hibák felderítésére és javítására, illetve egy adott nyelvtechnológiai feladat hibatűrővé tételére azáltal, hogy a számára ismeretlen szavakat is egy ismert szóhoz való hasonlósága révén kezelhetővé tesszük.

Mivel a tanítókorpusz a webről gyűjtött szövegekből áll, ezért sok nem sztenderd vagy szleng szóalak is előfordul benne. A modell ezekre is jól működik, ami szintén jól hasznosítható a csupán sztenderd szóalakokat ismerő szövegfeldolgozó

eszközök támogatása során. A 3. táblázat utolsó két oszlopa ilyen kifejezésekre kapott eredményeket tartalmaz.

3. táblázat. Példák a rendszer által a hibásan lemmatizált (első oszlop) és a hibásan írt (második oszlop) szavakhoz visszaadott hasonló kifejezésekre, illetve nem sztenderd szóalakokra (utolsó két oszlop).

pufidzsek	angolúl	mittomén	hehehe
rövidnac(43)	magyarúl(486)	mittudomén(2969)	hihihi(1203)
napszemcs(37)	németűl(132)	mifene(2455)	hahaha(3822)
szemcs(37)	francziául(25)	mittoménmi(412)	höhö(1827)
szmöty(45)	angolol(27)	mittudoménmi(441)	brr(1212)
zacs(170)	írül(95)	nemtommi(469)	muhaha(1498)
suzuk(131)	mindenrõl(422)	neadjisten(1741)	heh(1603)
sap(374)	minderrõl(129)	blablabla(2590)	Muhaha(879)
törcs(11)	ilyenröl(58)	stbstb(1739)	muhahaha(428)
kispolszk(41)	Amirõl(143)	bla-bla-bla(711)	hajaj(1579)
sokmindenk(58)	olyasmirôl(38)	jahh(466)	höhöhö(361)

3.4. Analógiavizsgálatok

A beágyazási modellek kiértékelésének egyik módszere az angol nyelvű modellek esetén az analógiatesztek elvégzése [7]. Ezeknél a teszteknél egy szópárosból és egy tesztszóból indulnak ki. A rendszer feladata annak a szónak a megtalálása, ami tesztszóhoz az eredeti szópáros közötti relációnak megfelelően viszonyul. Például a *férfi – nő* páros és a *király* tesztszó esetén a várt eredmény a *királynő*. Elvégeztünk ugyan néhány ilyen tesztet, azonban mivel a többértelmű szavakhoz egy reprezentációs vektor tartozik, ezért a szópárok közötti relációkat kevésbé sikerült jól modellezni. Az előbbi példában a *nő* szó igei és főnévi jelentései keverednek, ezért a *férfi* és a *nő* szavak közötti távolság nem pontosan felel meg a *király* és a *királynő* közötti távolságnak (aminek oka a *király* szó többértelműsége is). Így csupán elvétve találtunk olyan analógiapéldákat, melyek helyes eredményt adtak. Ilyen volt például a *hó – tél* páros és a *nap* tesztszó esetén eredményül kapott *nyár*. Részletes kiértékelést azonban ebben a feladatban nem végeztünk, hiszen előbb a jelentés-egyértelműsítés problémakörének megoldását tartjuk kritikus fontosságúnak.

3.5. Szemantikai csoportok kinyerése

A fenti modelleket szemantikai csoportok kinyerésére használtuk fel. Mivel a cél ebben a részfeladatban a kifejezések szemantikai besorolása volt, ezért ehhez csak az `ANA` modellt (tehát a lemmákat tartalmazót) használtuk. Minden szemantikai csoporthoz meghatároztunk egy kezdő szót, ami az adott csoportba tartozik.

Ehhez a szóhoz meghatároztuk a 200 leghasonlóbb szót a létrehozott modellből, majd ennek a listának a 200. eleméhez szintén lekérdeztük a 200 leghasonlóbb szót és ezt a lépést ismételtük legfeljebb 10 alkalommal. Az így létrejött max. 2000 elemű listában ellenőriztük, hogy melyik indikátorszó nem járult hozzá a korábbiakhoz képest új elemekkel, ezeket a szavakat töröltük a lekérdezések közül, majd újra lefuttattuk az algoritmust. Így minden szemantikai csoporthoz, a csoportba tartozó egyetlen kiindulási szó meghatározása után több száz vagy akár ezer, az azonos csoportba tartozó kifejezést nyertünk ki automatikusan. Úgy találtuk, hogy bizonyos (szűkebb) szemantikai mezőkben a 200 szavankénti lekérdezés túl sok zajt eredményezett, például amikor kifejezetten ruhaanyagok gyűjtése volt a cél. Ekkor az egyszerre lekérdezett kvantum 50 eleműre csökkentésével kaptunk viszonylag jól használható eredményt.

4. Eredmények

Az eredmények vizsgálatát több módszerrel végeztük. A szemantikai kategorizációs feladatban kézzel számoltuk meg az eredményül kapott listában a helyes és nem helyes szavak arányát. Ahhoz azonban, hogy a kézzel történő ellenőrzést hatékonyabban tudjuk végezni, egy klaszterezést is alkalmaztunk az eredménylistára, illetve az eredménylistában szereplő szavak sokdimenziós reprezentációját leképeztük egy kétdimenziós térbe, ahol a klaszterezés eredményével együtt jelenítettük meg a szavakat, jól áttekinthető vizuális megjelenítéssel támogatva az ellenőrzést.

4.1. Klaszterezés

A lexikai elemek klaszterezéséhez hierarchikus klaszterezést alkalmaztunk, melynek bemenete a csoportosítandó szavakat tartalmazó listán szereplő lexikai elemekhez tartozó szemantikai vektor, a klaszterezés során pedig a vektorok távolságát Ward [11] módszere alapján határoztuk meg. Ennek köszönhetően a kapott dendrogram alsó szintjein tömör, egymáshoz közel álló kifejezésekből álló csoportok jöttek létre. Célunk azonban nem egy bináris faként ábrázolt teljes hierarchia meghatározása volt, hanem a fogalmak elkülönülő csoportjainak meghatározása, azaz a kapott dendrogram egyes kompakt részfái. A klaszterezés és a részfák kivágására szolgáló módszer részleteit [9]-ben közöltük. A 4. táblázatban néhány eredményül kapott klaszterre láthatunk példát egy-egy szemantikai kategórián belül. Jól látható, hogy az egy klaszterbe sorolt kifejezések egymáshoz szorosabban kapcsolódnak az adott kategórián belül is. Természetesen, az algoritmus lehetőséget biztosít a klaszterezés kifinomultságának állítására, így akár nagyobb, vagy még kisebb csoportosítás is könnyen kinyerhető. A példák között a foglalkozások között kiemelendő a különböző katonai rangok rövidített alakjainak csoportja, illetve a nyelvek esetén a magyar nyelvjárásokat összegyűjtő csoport. Külön klaszterekbe gyűltek össze az adott feladat szempontjából ugyan szemantikailag releváns, de önmagában nem tökéletes megoldások is, például a

nyelveknél azok a földrajzi nevek, amelyek egy-egy nyelvváltozat jelzői, de önmagukban nem nyelvnevek, a nyelvpárok, illetve a kifejezetten tévesen a listán feltűnő elemek, például színpárok. Ez meglehetős mértékben megkönnyíti a generált listák kézi ellenőrzését is, mert a nyilvánvalóan hibás csoportok gyorsan kiszűrhetők.

4. táblázat. Klaszterekbe rendezett kifejezések a négy vizsgált szemantikai csoport esetén

Foglalkozások
író költő író drámaszerző prózaíró novellista színműíró regényíró drámaíró
ökológus entomológus zoológus biológus evolúcióbiológus etológus
hidegburkoló tapétázó mázoló szobafestő festő-mázoló szobafestő-mázoló bútorasztalos
tehénpásztor kecskepásztor birkapásztor fejőnő marhahajcsár tehenész marhapásztor
őrm ftörm zls alezr vőrgy szkv ezds hdgy őrgy szds fhdgy
Nyelvek
kuwaiti szaudi szaúdi kuvaiti jordán szaúd-arábiai jordániai
lengyel cseh bolgár litván román szlovák szlovén horvát szerb
osztrák-német német-osztrák elzászi dél-tiroli flamand
bánsági háromszéki gömöri széki gyimesi felföldi sárközi
Anyagnevek
feketeszén kőszén barnaszén lignit feketekőszén barnakőszén
fluorit rutil apatit aragonit kvarc kalcit földpát magnetit limonit
konyhasó kálium-klorid nátriumklorid nátrium-klorid
Textilek
selyemszatén béléselyem düsesz shantung
posztó szűrposztó abaposztó őzbőr teveszőr kendervászon házivászon háziszőttes
csipke bársony selyem kelme brokát selyemszövet tafota damaszt batiszt

4.2. Vizualizáció

Mivel a fogalmakat reprezentáló vektorok egy szemantikai térben helyezik el az egyes lexikai elemeket, ezért gyakran alkalmazott módszer ennek a szerveződésnek a vizualizációja. Ehhez a sokdimenziós vektorokat egy kétdimenziós térbe képeztük le a t-sne algoritmus alkalmazásával [4]. A módszer lényege, hogy a szavak sokdimenziós térben való páronkénti távolságának megfelelő eloszlást közelítve helyezi el azokat a kétdimenziós térben, megtartva tehát az elemek közötti távolságok eredeti arányát. Így könnyen áttekinthetővé válik a szavak szerveződése, a jelentésbeli különbségek jól követhetőek és felmérhetőek.

A vizualizáció során a klaszterezés eredményeit is megjelenítettük, a különböző klaszterbe került szavakat különböző színnel jelenítve meg. Az így létrejött ábrán jól követhetővé váltak a klaszterek közötti távolságok is.

2. ábra. A foglalkozásnevek t-sne vizualizációjának egy kis részlete. A kifejezések egymástól való távolsága az ábrán arányos a jelentésbeli hasonlóságukkal.

4.3. Kvantitatív kiértékelés

A klaszterezésnek köszönhetően tehát könnyebbé vált az egyes szemantikai kategóriákba sorolt szavak ellenőrzése, hiszen az egy klaszterbe sorolt szavak csoportosan ellenőrizhetőek voltak, mivel a részfák kivágása során úgy állítottuk be a paramétereket, hogy inkább kisebb, de tömör és koherens csoportok jöjjenek létre. Az így kapott eredményeket a 5. táblázat foglalja össze.

5. táblázat. A szemantikai kategorizáció eredményei

	nyelvek	foglalkozások	anyagnevek	textilek
Helyes	755	2387	1139	120
Hibás	98	134	162	114
Átmeneti	391	37	54	0
Összesen	1244	2558	1355	234

Négy szemantikai kategóriára végeztük el a méréseket: nyelvek, foglalkozások, anyagnevek és ezen belül a textiltípusokra. A kapott lista szavait (illetve a klaszterezés eredményeként kapott klasztereket, amikor ezek homogének voltak) három kategóriába soroltuk: helyes, hibás, illetve létrehoztunk egy átmeneti kategóriát, amelybe a feladat szempontjából nem egészen hibás, de nem is tökéletes megoldások kerültek (például a foglalkozásnevek között az egyéb embert, emberszerű lényt vagy foglalkozástól különböző emberi szerepet jelölő szavak, pl. *srác, öregasszony, hölgy, albérlő, élettárs, kobold*). Látható, hogy a textilek kategóriáját kivéve mindegyik esetben igen nagy volt a helyes szavak aránya (a textilek esetében pedig nagyon egyszerű volt a találatok kézi ellenőrzése, mert gyakor-

latilag az összes téves találat külön klaszterekbe gyűlt össze, amelyek kizárólag ruhaanyagokból készült cikkeket: ruhadarabokat, lábbeliket, lakástextiltermékeket tartalmaztak). A 10 indikátor szó alapján 755 nyelv, 2387 foglalkozás és 1139 anyagnév gyűlt össze, amik igen jó eredménynek számítanak ahhoz képest, ha ezeket a listákat kézzel kéne összeállítani. Sok esetben az átmeneti jelölést kapott szavak is helyesek lehetnek egy-egy feladatban, most azonban a legszigorúbb értékelést alkalmaztuk, ezért nem jelöltük őket elfogadottnak.

5. Részletes hibaelemzés

A négy kategória közül az egyikre (nyelvek) részletes hibaelemzést is készítettünk. Az egyes szavak helyességének, illetve a nem nyelvként szereplő nevek hibatípusának megítélésekor az eredeti célt tartottuk szem előtt, azaz a morfológiai adatbázis szemantikai jegyekkel való bővítését. Így, ebben az esetben több szóalakot is elfogadhatónak tekintettünk.

A 6. táblázat a különböző nyelvkategóriák disztribúcióját tartalmazza, melyek a következők:

Az első csoport nyelveket, nyelvtípusokat tartalmaz.

- Sztenderd nyelvek: egy nyelv hivatalos neve, a helyesírási szabálynak megfelelő alakban.
- Kitalált nyelv: egy irodalmi alkotás szerzője által kitalált nyelv neve.
- Egy nyelvcsoport vagy nyelvcsalád neve: pl. *uráli*
- Népcsoport neve, de nem nyelv: pl. *zsidó*. Ezeket a kifejezéseket a köznyelvben gyakran használják úgy, mintha nyelvek lennének (pl. *zsidó nyelv, zsidóul*).
- Egy írásrendszer neve: pl. *dévanágari*, *cirill*. A nyelvtani konstrukciók, amikben ezek szerepelnek hasonlóan viselkednek a nyelvekkel használt konstrukciókhoz.
- Nyelvtípus: pl. *kreol*, *patois*, *pidzsin* (az ilyen típusú nyelvek összetett nevének utolsó része)

A második csoportba nyelvek attribútumait sorolhatjuk:

- Földrajzi helyet jelölő tulajdonság: egy nyelv, dialektus vagy nyelvcsoport jelzője, ami önmagában nem használható a nyelv neveként, pl. *iraki* (arab), *mezopotámiai* (nyelvek)'
- Más (nem földrajzi) attribútumok: *rabbinkus* (héber)'

A harmadik csoportba helyesírási változatokat, szinonimákat és elírt változatokat soroltunk:

- Szinonimák: egy nyelv alternatív (pl. régies) megnevezése, pl. *tót – szlovák* , *hellén – görög*.
- Helyesírási változatok (nyelv, nyelvcsoport vagy dialektus esetén): archaikus alakok, fonetikai variánsok vagy latin helyesírás szerinti alakok, pl. *franczia*, *bulgár*, *szittya*, *scytha*

– Súlyosabb elírások: egy nyelv, dialektus vagy nyelvcsoport nevében hiányzó, fölösleges, vagy felcserélt betűk

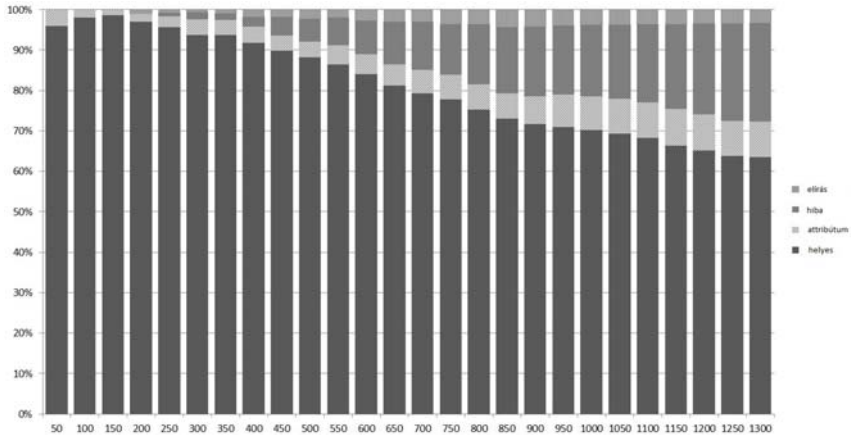
Az ebbe a három csoportba tartozó szóalakok a morfológiai elemző adatbázisának bővítése szempontjából nyelvnek tekinthetők. Ezek a közel 1300 szónak a 74,96%-át teszik ki. A többi 25,04% nem nyelvmegnevezés. Ide soroltuk például azokat a nyelvpárokat (pl. *magyar-angol*), ahol a nyelvpár nem egy nyelvcsoportot jelöl, viszont az olyan párokat, mint pl. a *bajor-osztrák*, ahol a két nyelv együtt alkot egy dialektust, azokat nyelvként fogadtuk el.

6. táblázat. A nyelvekre készített részletes hibaelemzés eredménye. A százalékértékek az 1244 elemű listából számított arányok.

típus	**példa**	**pontosság**
sztenderd nyelv	*joruba*	39,83%
kitalált nyelv	*újbeszél*	1,11%
dialektus neve	*Cockney*	5,33%
nyelvcsoport vagy nyelvcsalád neve	*uráli*	4,37%
népcsoport, de nem nyelv	*zsidó*	1,03%
írásrendszer	*cirill*	0,72%
nyelvtípus	*kreol*	0,32%
írásváltozat	*scytha*	10,25%
szinonima	*hellén*	2,07%
elírás	*ngol*	3,42%
földrajzi jelző	*iraki*	8,51%
más jelző	*rabbinkus*	0,40%
		74,96%
nem nyelv, nyelvpár	*magyar-angol*	25,04%

A 3. ábra a módszer pontosságának alakulását mutatja az automatikusan kinyert nyelvnévlista hosszának függvényében. Látható, hogy a lista elején sokkal kevesebb hiba található, míg ha az eredeti indikátorszavaktól egyre távolabb kerülünk a szemantikai térben, úgy kerül be egyre több zaj a kinyert listába. Az ábra jól illusztrálja a word2vec algoritmusban implementált hasonlóságszámítás hatékonyságát is, ami alapján ez a rangsorolás létrejön.

A módszer által adott lista fedésének becslése jóval nehezebb feladat, mint a pontosság meghatározása, mivel magyar nyelven nem találtunk a nyelveket, nyelvcsaládokat és nyelvcsoportokat tartalmazó teljes listát. (Ha létezne ilyen, akkor ezt használhattuk volna az eredeti feladatban is.) Ugyanez igaz a többi szemantikai kategóriára (foglalkozások, anyagnevek, stb.), ráadásul a bemutatott módszer tetszőleges szemantikai csoport kinyerésére alkalmazható.



3. ábra. A módszer pontossága az automatikusan kinyert lista hosszának függvényében. A *helyes* szavak azok, amiket nyelvnek fogadtunk el, az *attribútumok*, amiket nyelvek jelzőinek, az *elírások* olyan nyelvnevek, nyelvcsoportok, nyelvcsaládok, stb., amikben kisebb elírás szerepel, a *hiba* kategóriába pedig azok a szavak tartoznak, amik a fentiek közül egyik kategóriába sem tartoznak.

6. Konklúzió

Cikkünkben bemutattuk, hogy az egyre népszerűbb, neurális hálózatok betanításán alapuló szemantikai beágyazási modellek magyar nyelvre is jó eredménnyel működnek kellő méretű és elemzett tanítóanyag alkalmazása esetén. Néhány általános kísérlet elvégzése mellett a létrejött szóreprezentációk egy konkrét feladatra való felhasználhatóságát is megvizsgáltuk. Ennek során célunk többek között egy meglévő morfológiai elemző lexikonában a morfológiai, szintaktikai, szemantikai szempontból releváns kategóriainformáció gazdagítása, illetve ellenőrzése. Mivel a modell alkalmasnak bizonyult arra, hogy szavakhoz azokhoz valamilyen szempontból hasonló szavakat rendeljen, ezért az egy kategóriába (foglalkozások, nyelvek, anyagnevek) tartozó szavak automatikusan kinyerhetőek. Továbbá, a modellek folytonosságából adódóan a hasonlóság mértéke tetszőlegesen állítható, így a kategorizálás különböző absztrakciós szinteken valósítható meg. Az eredményekben megmutattuk, hogy számos olyan szót tudtunk a megfelelő kategóriacímkével ellátni, melyre kézi gyűjtés esetén csak nagyon sok további munka árán lett volna lehetőség. Ugyancsak alkalmasnak bizonyult a módszer a különböző annotációs és egyéb korpuszhibák kimutatására és osztályozására is.

Hivatkozások

1. Banea, C., Chen, D., Mihalcea, R., Cardie, C., Wiebe, J.: Simcompass: Using deep learning word embeddings to assess cross-level similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 560–565.

Association for Computational Linguistics and Dublin City University, Dublin, Ireland (August 2014), http://www.aclweb.org/anthology/S14-2098
2. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 238–247. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/P14-1023
3. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Sensembed: Learning sense embeddings for word and relational similarity. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 95–105. Association for Computational Linguistics, Beijing, China (July 2015), http://www.aclweb.org/anthology/P15-1010
4. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne (2008)
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013), http://arxiv.org/abs/1301.3781
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. pp. 3111–3119 (2013), http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality
8. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 746–751. Association for Computational Linguistics, Atlanta, Georgia (June 2013), http://www.aclweb.org/anthology/N13-1090
9. Siklósi, B., Novák, A.: Közeli rokonunk, az autó. In: Tanács, A., Varga, V., Vincze, V. (eds.) XII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 27–36. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2016)
10. Trask, A., Michalak, P., Liu, J.: sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. CoRR abs/1511.06388 (2015), http://arxiv.org/abs/1511.06388
11. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963), http://www.jstor.org/stable/2282967