A Dologfelismerő

Novák Attila[1,2], Siklósi Borbála[2]

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

Kivonat A szóbeágyazási modellek megjelenése az utóbbi években forradalmi változásokat hozott a nyelvtechnológia számos területén. A tömör valós vektorokkal való jelentésreprezentáció ugyanakkor közvetlenül nem interpretálható az emberek számára, bár a különböző vizualizációs technikák segítenek a modellek értelmezésében. Jelen cikkben egy olyan technikát mutatunk be, amely a szavakhoz diszkrét szemantikai jegyeket rendelve segíti a folytonos modellben ábrázolt jelentések értelmezését, ugyanakkor hozzáférhetővé teszi az azokban reprezentált tudást a diszkrét jegyekkel dolgozó gépi tanuló vagy keresőalgoritmusok számára is. Kísérleteink során hasonló magyar nyelvű erőforrások hiányában angol nyelvű lexikai erőforrásokban szereplő kategóriacímkéket rendeltünk magyar szavakhoz. Az alkalmazott transzformációk ellenére a modell jól címkézi a gyakori szavak mellett a semmilyen kézzel készített erőforrásban nem szereplő ritka, szleng szavakat, a neveket és a rövidítéseket is.

1. Bevezetés

A szavak disztribúciós viselkedésének reprezentálására az utóbbi években egyre népszerűbbé vált szóbeágyazási modellek igen hatékony módszernek bizonyultak [13]. Azonban a szavak sokdimenziós térben való absztrakt reprezentációja az emberek számára önmagában nehezen értelmezhető. Ennek orvoslására egy olyan módszert mutatunk be, ami egy nagy korpuszból létrehozott szóbeágyazási modellhez olyan szemantikai kategóriacímkéket tesz hozzá, amik az eredeti modell értelmezését segítik. A hozzáadott címkéket létező lexikai erőforrásokból, azok automatikus transzformációjával illesztjük az eredeti beágyazási térbe. Ennek köszönhetően az eredetileg nagyon sok szót tartalmazó szemantikai tér felbontható jóval kisebb számú, címkézett altérre, ami a modellt átláthatóbbá teszi.

A bemutatott algoritmus az eredeti korpuszban lévő összes szóhoz képes kategóriacímkét rendelni, függetlenül attól, hogy az adott szóalak a címkék létrehozásához használt lexikai erőforrásban szerepelt-e. Továbbá, a módszer nyelvfüggetlen, a felhasznált erőforrások nyelve nem szükségszerűen azonos az eredeti szóbeágyazási modell nyelvével. A szavak kategorizálásakor a szóbeágyazási modellek természetéből adódóan nem egy előre definiált tudományos rendszertani besorolás érvényesül, hanem a szavak reprezentációjának alapja azok disztribúciós viselkedése, tehát a tényleges nyelvhasználat.

A módszert magyar nyelvre mutatjuk be, de más nyelvre is könnyen adaptálható, amennyiben egy szóbeágyazási modell (vagy egy elégséges méretű korpusz) rendelkezésre áll.

2. Kapcsolódó munkák

A szóbeágyazási modellek az utóbbi évek egyik legnépszerűbb eszköze a szavak jelentésének hatékony reprezentációjára [12,19]. Akár szó-, akár jelentésbeágyazásról van szó, az ezekben használt folytonos vektor reprezentációk nem alkalmasak az emberi értelmezésre. Történtek kísérletek ezeknek a beágyazási modelleknek olyan létező szemantikai erőforrásoknak az „összeházasítására", mint a BabelNet [18] vagy a WordNet [6,1]. Rothe és Schütze a szóbeágyazási vektorok kombinálásával próbált WordNet synseteket belevetíteni az eredeti beágyazási térbe [20]. Más megközelítések pedig kézzel annotált adathalmazhoz próbálták adaptálni az eredeti modellt [9]. Több kutatás során pedig tudásbázisok felhasználásával próbálták javítani a beágyazási modellek minőségét [25,2,4]

3. Lexikai erőforrások

Az eredeti modellhez hozzárendelt kategóriacímkéket létező, angol nyelvű lexikai erőforrásokból nyertük ki.

Angol nyelvre az egyik legnépszerűbb ilyen erőforrás a **WordNet** [7,14], ahol a fogalmak egy szigorú hierarchikus rendszerbe vannak besorolva. A probléma azonban ezzel az erőforrással az, hogy egyrészt az alacsonyabb szinteken igen nagy a felbontása, a magasabb szintű kategóriák viszont túl általánosak [3]. Másrészt, a középső szinteken olyan mesterkélt kategóriákat tartalmaz, ami egy tudományos taxonómiában indokolt, azonban a mindennapi nyelvhasználatnak nem feltétlenül része (pl. *páros ujjú patás*). Ráadásul a WordNetben a synseteknek nincs neve sem, csak azonosítója és definíciója. Ezért a WordNetet végül nem használtuk az itt leírt kísérleteinkben.

Egy másik, gyakran használt, bár kissé elavult lexikai erőforrás a **Roget's Thesaurus** [5]. A digitálisan elérhető változata 990 szemantikai kategóriát tartalmaz. Minden kategória alatt 5 szófaj szerinti bontásban (főnév, ige, melléknév, határozószó, kifejezés/indulatszó) az adott kategória/szófaj alá sorolható szavak listája található. Az eredeti tezaurusz 91608 szót, illetve kifejezést tartalmaz, azonban a kísérleteink során használt az angol Wikipédiából épített szóbeágyazási modellünkben ezekből csak 51108 szó szerepel.[3] Mivel a modellben csak szavak szerepelnek, ezért a két halmaz metszetéből hiányoznak a többszavas kifejezések, az elavult szavak, illetve a téves szófajcímkével ellátott szavak.

Szintén online elérhető a **Longman Dictionary of Contemporary English** (LDOCE) [23] digitális változata, amiből könnyen előállítható egy az előző erőforráshoz hasonló gyűjtemény is, hiszen a benne lévő címszavak egy része 213

[3] Hogy a Wikipédiából épített modellt hogyan építettük, és pontosan mire és hogyan használtuk, az a 4. részben fog kiderülni.

szemantikai kategóriába van besorolva, szintén szófaj szerinti bontással együtt. Mivel azonban ez a szótár jóval modernebb szókincset tartalmaz, ezért az angol Wikipédia-modellel való metszés után az eredeti 28257 kategorizált szóból 21546 megmaradt.

A harmadik erőforrás a szintén az LDOCE-n alapuló **4lang** szótár volt. Ebben az erőforrásban az eredeti szótár definícióinak formális átirata szerepel [8], ahogy az az alábbi példán látszik:

```
bread: food, FROM/2742 flour, bake MAKE
show: =AGT CAUSE[=DAT LOOK =PAT], communicate
```

Ezt az ábrázolásmódot tovább alakítottuk úgy, hogy az előzőeknek megfelelő formát kapjunk. Ehhez a formális definíciókat feldaraboltuk (szóközök és zárójelek mentén) és minden egyes így kapott darabot címkének tekintettünk, összegyűjtve hozzá minden olyan szót, aminek a leírásában az adott címke szerepelt. Így 1489 kategóriacímke jött létre ebből a szótárból, amelyek összesen 12507 szóval voltak összerendelve. Ezekből a Wikipédia-szókinccsel való metszés után 11039 szó maradt. Annak ellenére, hogy ebben a szótárban főleg gyakori szavak szerepelnek, mégis voltak olyan elemek, amik nem szerepeltek az elemzett Wikipédiából készített modellben. Ezek többnyire toldalékok, illetve toldalékolt szóalakok.

Az 1. táblázatban néhány példa látható a kategóriacímkékre és a hozzájuk rendelt szavakra.

Erőforrás	**Kategória**	**Példák az eredeti erőforrásból**
ROGET	Mean_N	medium#NN generality#NN neutrality#NN middle_state#NN median#NN golden_mean#NN middle#NN
ROGET	Rotundity_ADJ	spherical#JJ cylindric#JJ round_as_an_apple#JJ bell_shaped#JJ spheroidal#JJ conical#JJ globated#JJ
LDOCE	Cooking	allspice#NN bake#VB barbecue#VB baste#VB blanch#VB boil#VB bottle#VB bouillon_cube#NN
LDOCE	Mythology	centaur#NN chimera#NN Cyclops#NN deity#NN demigod#NN faun#NN god#NN griffin#NN gryphon#NN
4LANG	food	sandwich#NN, fat#NN, bread#NN, pepper#NN, meal#NN, fork#NN, egg#NN, bowl#NN, salt#NN
4LANG	=DAT	say#VB, show#VB, allow#VB, swear#VB, grateful#ADV, let#VB, teach#VB, give#VB, help#VB

1. táblázat. Példák az eredeti lexikonokban található kategóriákból és hozzájuk tartozó szavakból azok azonos formára való átalakítása után

A 2. táblázat első négy oszlopa foglalja össze a felhasznált erőforrások jellemzőit. A 4.2 részben leírt módon az angol Wikipédiából épített modell alapján klaszterezést is végeztünk az egyes kategóriákat jellemző szavakon. Ennek eredménye látható az utolsó három oszlopban.

4. Módszer

Célunk egy olyan eszköz létrehozása volt, ami egy tetszőleges szóhoz hozzárendeli a megfelelő szemantikai kategóriacímkéket, akkor is, ha az adott célszó nincs benne egyik lexikai erőforrásban sem, illetve, ha ilyen lexikai erőforrás az adott nyelven nem is létezik. Ezért két problémát kellett kezelni: a kategóriacímkék hozzárendelését és a nyelvi különbség áthidalását.

Erőforrás	Eredeti			Metszet és klaszterezés után		
	kategória	szó	szó/kat.	kategória	szó	szó/kat.
LDOCE	213	28257	132,66	3069	21546	7,02
ROGET	3077	91608	29,77	7066	51108	7,23
4LANG	1489	12507	8,39	2249	11039	4,91

2. táblázat. A felhasznált erőforrások jellemzői (különböző kategóriák száma, szavak száma, átlagos szószám kategóriánként; az angol modellel való metszés, illetve a klaszterezés előtt és után).

4.1. Szóbeágyazási modellek létrehozása

A nyelvtechnológiai kutatások egyik kurrens módszere a folytonos vektoros (*word embedding*) reprezentációk alkalmazása, 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. A módszer hátránya csupán az, hogy önmagában nem képes a poliszémia, illetve homonímia kezelésére, tehát egy többjelentésű lexikai elemhez is csupán egyetlen jelentésvektort rendel. Ennek részleges kezelésére egy egyszerű megoldást alkalmaztunk azokban az esetekben, ahol az azonos alakok különböző szófajúak. Ehhez a modell építése előtt szófaji egyértelműsítést és lemmatizálást alkalmaztunk a korpuszra[4] a PurePos szófaji egyértelműsítő [17] és a Humor morfológiai elemző [15,16] használatával, majd a fő szófajcímkéket hozzáfűztük a szótövekhez, így az azonos alakú, de különböző szófajú szavaknak külön reprezentációja jött létre. Korábban azt is megmutattuk, hogy az összetett morfológiájú nyelvek esetén jobb minőségű szóbeágyazási modell hozható létre, ha a további morfológiai címkékben kódolt információk különálló tokenként maradnak meg a modell építéséhez használt szövegben, így a hozzájuk tartozó szótő kontextusában jelennek meg [22,21].

Mivel a jelen cikkben bemutatott címkézőrendszer megvalósítása során a magyar modellt egy angol szóbeágyazási modellnek is meg akartuk feleltetni, ezért ezt is hasonló módon hoztuk létre. A 2,25 milliárd szavas angol Wikipédia[5] szövegeit a Stanford tagger [24] használatával elemeztük, a szófajcímkéket a szótövekhez csatoltuk, a további morfológiai címkéket pedig külön tokenként leválasztottuk.

Mind az angol, mind a magyar modell tanításához a word2vec[6] eszközben implementált CBOW modellt használtuk, 5 token sugarú szövegkörnyezetet véve figyelembe és 300 dimenziós beágyazási modelleket hozva létre.

4.2. Szemantikus kategóriacímkék beágyazása

Ha van egy beágyazási modellünk, akkor az abban szereplő szóvektorok klaszterezésével könnyen létrehozható egy az eredeti szótárnál kevesebb elemből álló reprezentáció, amiben az egyes klaszterekbe tartozó szavakat valamilyen szempont szerint hasonló szemantikai jegyekkel rendelkező szavaknak tekinthetjük.

[4] A korábbi modelljeink [22,21] építéséhez használt webkorpuszt alkalmaztuk itt is.
[5] letöltve: https://dumps.wikimedia.org/ 2016. május
[6] https://code.google.com/p/word2vec/

Ebben az esetben azonban ezeknek a közös szemantikai jegyeknek a meghatározása csupán kézzel lehetséges, de még akkor is nehézséget jelenthet ezeknek a csoportoknak a felcímkézése ember által értelmezhető formában. Továbbá, hacsak nem valamilyen probabilisztikus klaszterezést alkalmazunk, minden szó csak egy klaszterbe kerül.

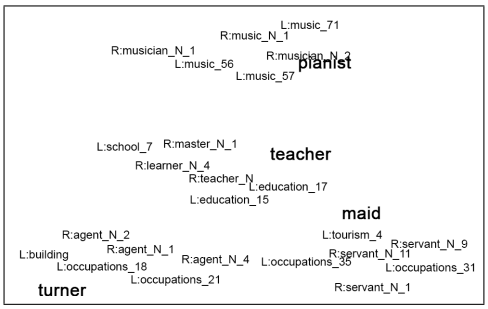
Az elnevezés problémája kezelhető lenne úgy, hogy a klaszter egy reprezentatív elemét (pl. a centroidhoz legközelebbit) kiválasztjuk, azonban ebben az esetben nem a csoportba tartozó szavak közös szemantikai kategóriáját határoznánk meg, hanem csupán a kategória egy példányát jelölnénk meg. A probléma megoldására tehát azt a módszert alkalmaztuk, hogy a fent felsorolt lexikai erőforrások kategóriacímkéit helyeztük el az eredeti szóbeágyazási modell által létrehozott szemantikai térben.

Az eredeti erőforrásokban szereplő kategóriacímkék azonban néha túl általánosnak bizonyultak, ezért a hozzájuk rendelt szólista is igen heterogén volt. Ezért először az eredeti kategorizációt tovább bontottuk úgy, hogy egy hierarchikus klaszterezési algoritmus [22] segítségével csoportosítottuk az 5-nél több szót tartalmazó kategóriacímkéhez tartozó szólistákat. Ennek eredményeképpen minden ilyen címkéhez alcsoportok jöttek létre, melyeket numerikus indexszel különböztettünk meg egymástól. A 2. táblázat utolsó három oszlopában láthatók a klaszterezés eredményeként kapott lexikonok jellemzői.

A klaszterezés során a Roget's Thesaurus elavult szóhasználatából adódó problémát is sikerült némileg áthidalni. Mivel az egyes szavakhoz tartozó reprezentációt a klaszterezéshez a modern Wikipédiából épített modell alapján nyertük ki, ezért az esetleg korábban más jelentéssel bíró szóalakokhoz is azok modern jelentését tudtuk reprezentálni. Például a *Combatant* kategóriába tartozó szavak közül a *charger, battery, file, monitor* külön klaszterbe került, hiszen ezek ma már inkább számítástechnikai/elektronikai jelentést hordoznak. Így bár maga a kategóriacímke nem feltétlenül jellemzi jól a hozzá tartozó szemantikai jegyet, de a klaszterezés során hozzáadott numerikus index alapján azonosítható és jól elválasztható ez a kategória a *Combatant* címkéhez tartozó szavakból létrejött többi, katonai kifejezéseket tartalmazó kategóriától.

Ezután minden így létrejött új címkéhez hozzárendeltük a benne felsorolt szavak beágyazási vektorának átlagát a szófajcímkékkel ellátott és tövesített angol Wikipédia-modellből. Így megkaptuk a kategóriacímkék pozícióját az angol szóbeágyazási térben. A címkékhez tartozó vektorokat külön tároltuk, hogy a lekérdezés során könnyen le lehessen szűkíteni az eredményt kategóriacímkékre, illetve szavakra. Egy angol, szófajcímkével ellátott szóhoz tehát úgy kaphatjuk meg a megfelelő kategóriacímkéket, hogy az őt reprezentáló vektorhoz koszinusztávolság alapján legközelebbi vektorokat kérdezzük le a kiválasztott erőforrásban szereplő kategóriacímkék vektorai közül .

Az 1. ábrán négy angol szó (*pianist 'zongorista', teacher 'tanár', turner 'esztergályos', maid 'takarítónő'*) és a LDOCE és Roget modellekből a hozzájuk tartozó 3 legközelebbi kategóriacímke elhelyezkedése látható két dimenzióba leképezve.



1. ábra. A *pianist, teacher, turner, maid* szavakhoz a LDOCE és a Roget modellekből lekérdezett 3-3 legközelebbi címke elhelyezkedése a szemantikai térben

4.3. Nyelvek közötti leképezés

Korábbi kutatások bemutatták, hogy a különböző nyelvekre létrehozott szóbeágyazási modellek által definiált szemantikai terek leképezhetők egymásba a leképezés során egy kiindulási szótár alapján megtanult páronkénti lineáris transzformáció alkalmazásával [11]. Ha a kiindulási szótárban n darab (w_x, w_z) szópár van, ahol w_x fordítása w_z, a vektorreprezentációik pedig $(x_i, z_i)_{i=1}^n$, akkor a W transzformációs mátrix az alábbi optimalizációs probléma megoldásaként meghatározható:

$$\min_W \sum_{i=1}^{n} ||Wx_i - z_i||^2 \tag{1}$$

Így minden forrásnyelvi x vektorra alkalmazható a $z = Wx$ transzformáció. A célnyelvi modellben a z vektorhoz közel található szavak x közelítő fordításai.

A 4lang-szótár 3477 angol-magyar szópárból álló részhalmazát használtuk fel a transzformációs mátrix tanításához szükséges kiinduló szótárként. Az eredeti szótárból azokat a párokat tartottuk meg, amiknek mindkét tagja legalább 10000-szer előfordul a megfelelő nyelvű korpuszban. További 100 szópáron kézzel kiértékelve a transzformáció 38%-os pontosságot adott az első legközelebbi szóra nézve, és 69%/81%-ot az első 5/10 legközelebbi szóra nézve. Ez azt jelenti, hogy a transzformáció során a megfelelő környékre történik a leképezés az esetek nagy részében. Mivel célunk nem a pontos fordítások azonosítása volt, hanem a magyar és az angol nyelvű szemantikai tér egymásra illesztése, ezért ez az eredmény igazolta a transzformáció alkalmazhatóságát. Ez a módszer tehát lehetővé teszi, hogy az angol erőforrásokból létrehozott szemantikai kategóriacímkékhez rendelt vektorokat az angol térből leképezzük a magyar nyelvű szóbeágyazási modell terébe.

5. Kísérletek és eredmények

A módszerünk elsődleges célja a szóbeágyazási modellek értelmezhetőségének támogatása, illetve a beágyazási tér különböző részeinek szemantikai jegyekkel való

automatikus annotálása. Ezért először az eredmények kiértékeléséhez egy webes felületbe integrált ábrázolásmódot használtunk, ami a t-sne algoritmus [10] alkalmazásával az eredetileg 300 dimenziós beágyazási teret kétdimenziós ábrán jeleníti meg. A felület lehetőséget ad arra, hogy magyar szavakhoz bármely modellből (Roget's, LDOCE, 4lang) lekérdezhessünk tetszőleges számú kategóriacímként és az így kapott annotált szemantikai teret megjelenítsük. A vizualizáció mellett azonban kvantitatív kiértékelést is végeztünk, különböző típusú szavakra vizsgálva a kategorizáció minőségét.

5.1. Általános szavak

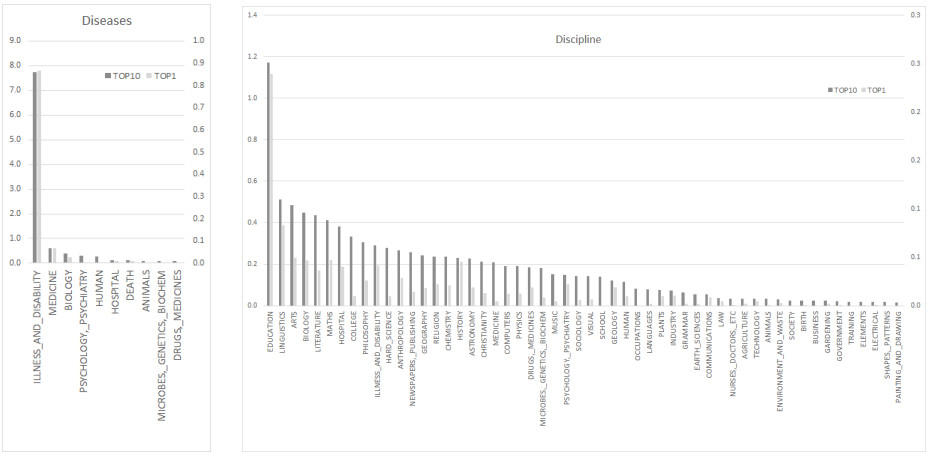
A sztenderd szóalakok kiértékeléséhez különböző szemantikai csoportokba sorolható szavakat gyűjtöttünk össze a [22]-ben bemutatott félautomatikus módszerrel, majd kézzel ellenőriztünk 35 ilyen csoportot (3. táblázat), amiben összesen 50507 szó szerepelt. Így ezeket gold standardnek tekintettük.

Csoport	**Szavak száma**	**Csoport**	**Szavak száma**
Foglalkozások	1332	Épületek	1123
Tudományágak	1051	Idő	380
Mértékegységek	1229	Esemény	3538
Elektronika	935	Színek	907
Betegségek	1359	Ruhák jellemzői	432
Állatok	1189	Emberek jelzői	980
Konyhai eszközök	769	Ételek jelzői	1141
Ételek	1662	Mozgást jelentő igék	1113
Járművek	1084	Szabadidős tevékenységek	1953
Ruhák	915	Pusztulást jelentő igék	909
Vizek	881	Magyar vezetéknevek	4197
Területek	1313	Latin vezetéknevek	2752
Természeti események	643	Angol vezetéknevek	1738
Domborzat	678	Szláv nevek	1479
Városok	4568	Becenevek	537
Helyek	3295	Emberi kapcsolatok	959
Embercsoportok	1206	Köszönések	322
Sportolók	445	**Sum**	**50507**

3. táblázat. A kiértékeléshez használt kézzel ellenőrzött szócsoportok és azok mérete

Ezután minden szóhoz mindhárom modellből lekérdeztük a 10 legközelebbi kategóriacímkét, majd összeszámoltuk, hogy az egyes csoportokon belül melyik címke hányszor fordult elő (függetlenül attól, hogy hányadik helyen szerepelt a 10-es listában). Egy másik esetben pedig csak az első helyen szereplő címkéket számoltuk össze, aminek célja a módszer pontosságának kiértékelése volt, azonban a tágabb értelmű csoportok esetén értelmetlennek bizonyult csupán egyetlen címke hozzárendelése, így az ebben az esetben mért alacsonyabb pontosság értékek nem feltétlenül jelentenek rosszabb teljesítményt.

A 2. ábra a *betegségek* és a *tudományágak* csoportjaira kapott eredményt mutatja a LDOCE címkék hozzárendelésére vonatkozóan. A grafikon egyes oszlopai a csoportban szereplő azon szavak arányát jelölik, amikhez az adott címkét rendelte a rendszer (az első 10 címke közül bármelyik pozícióban, illetve első címkeként). A címkékre vonatkozó összesítést a hozzájuk a klaszterezés során rendelt



(a) Betegségek (b) Tudományágak

2. ábra. A *betegségek* és a *tudományágak* csoportjába tartozó szavakhoz rendelt első 10 (TOP10) és az első (TOP1) címke eloszlása

index elhagyásával számoltuk, így ugyanaz a főcímke egy szó első 10 címkéjének listájában többször is előfordulhat, ezért jelennek meg 1-nél nagyobb aránynak megfelelő értékek az ábrán. Annak ellenére, hogy az összesítésben elhagytuk ezeket az indexeket, az indexek által reprezentált különbségek jelentősek lehetnek. Például a *biology* címkén belül elválnak a betegségekre, az emberi szervekre vagy a sejtbiológiára vonatkozó címkék, azonban ezek a csupán számokkal jelölt különbségek az emberi értelmezést (ilyen formában) nem segítik (egy gépi tanulási algoritmusban való felhasználás során azonban mindenképpen érdemes ezeket is figyelembe venni).

Látható, hogy a betegségeket tartalmazó csoport esetén a leggyakoribb címke az 'ILLNESS AND DISABILITY', ami az elsőként hozzárendelt címkék 78%-a, ami mellett csak néhány további címke jelenik meg számottevő arányban (*medicine; biology; psychology, psychiatry; human; hospital; death; animals; stb.*). A tudományágak csoportjában azonban sokkal kevésbé meredek a címkék eloszlásának íve. Bár a leggyakoribb címke ('EDUCATION') itt is kiemelkedik a többi közül és általánosságban jellemzi a csoportot, ezt a különböző tudományágak nevei követik közel egyenletesen csökkenő eloszlás szerint (*linguistics, arts, biology, literature, maths, etc.*). Látható tehát, hogy csoportonként eltérő lehet a címkék eloszlásának jellege, ezért a kiértékelés során nem a tényleges minőséget jellemezte volna az összes csoportra vonatkozó összesített eredmény. A 4. táblázat további néhány csoport szavaihoz az első 10 közül bármelyik helyen leggyakrabban előforduló kategóriacímkéket tartalmazza az egyes modellekből (L: LDOCE, R: Roget, F: 4lang). A példaként felsorolt csoportoknál látható az is, hogy a negyedik oszlopban szereplő összes különböző kategóriacímkék száma a csoport heterogenitásától függően eltérő, ugyanakkor minden esetben jóval kisebb, mint a csoportban szereplő szavak száma, tehát elmondható, hogy az alkalma-

Csoport	szavak	TOP 10 kategóriacímke	D	COV
Foglalkozások	1332	L: occupations, education, college, newspapers, publishing, painting and drawing, nurses, doctors, etc, music, construction, building, literature	80	71.25%
		R: Agent_N, Scholar_N, Remedy_N, Artist_N, Experiment_N, Book_N, Servant_N, Clothing_N, Painting_N, Accounts_N	97	61.26%
		F: HAS, profession, person, skill, scientist, educate, job, practice, science, IN/2758	112	67.64%
Mértékegységek	1229	L: measurement, currencies, computers, electricity, broadcasting, drink, maths, jewellery, numbers, elements	41	91.70%
		R: Length_N, Money_N, Gravity_N, Receptacle_N, Greatness_N, Littleness_N, Smallness_N, Period_N, Calefaction_N, Heat_N	64	89.50%
		F: unit, length, HAS, measure, =REL, temperature, cent, small, pound, mass	105	77.14%
Betegségek	1359	L: illness_and_disability, medicine, biology, psychology, psychiatry, human, hospital, death, animals, microbes, genetics, biochem, drugs, medicines	15	99.19%
		R: Disease_N, Death_N, Deterioration_N, Agitation_N, Hindrance_N, Disease_ADJ, Convexity_N, Remedy_N, Violence_N, Evil_N	35	91.83%
		F: bad, health, body, ill, disease, organ, situation, injury, damage, harm	37	89.92%
Természeti események	643	L: nature, meteorology, geography, illness_and_disability, geology, physics, earth_sciences, chronology, astronomy, power	41	80.87%
		R: River_N, Wind_N, Disease_N, Violence_N, Deterioration_N, Revolution_N, Evil_N, Agitation_N, Rotation_N, Resentment_N	88	57.39%
		F: cloud, wind, weather, ice, IN/2758, atmosphere, sudden, damage, AT/2744, HAS	92	55.99%
Emberek jelzői	980	L: animals, hair_and_beauty, clothes, colours, occupations, illness_and_disability, nature, clothes_and_fashion, biology, psychology, psychiatry	66	67.55%
		R: Size_ADJ, Clothing_N, Love_ADJ, Beauty_ADJ, Adolescence_ADJ, Animal_N, Sexuality_ADJ, Servant_N, Vulgarity_ADJ, Pleasurableness_ADJ	159	49.90%
		F: HAS, lack, kind, CAUSE, mad, IN/2758, bad, much, intelligent, body	100	66.94%
Mozgást jelentő igék	1113	L: transport, air, computers, theatre, swimming, government, water, insects, illness_and_disability, motor_vehicles	48	73.76%
		R: Journey_VB, Velocity_VB, Arrival_VB, Depression_VB, Navigation_VB, Departure_VB, Ascent_VB, Supposition_VB, Offer_VB, Haste_VB	81	71.70%
		F: =AGT, after, lack, AT/2744, go, =PAT, surface, rush, long, ON	48	62.71%

4. táblázat. Néhány kategóriához rendelt leggyakoribb címkék a három modellből. D=különböző címkék száma, COV=az első 10 címke aránya az összes hozzárendelt címkéhez képest

zott módszerrel hatékonyan sikerült az eredeti szóbeágyazás által meghatározott szemantikai térben szereplő sűrű numerikus vektorokat emberi értelmezésre is alkalmas szimbolikus jellemzők egy korlátozott méretű halmazára leképezni.

Bár a LDOCE címkék elnevezései a legérthetőbbek, a Roget és 4lang modellek alapján is hasznos szemantikai jegyeket határoztunk meg. Míg például a Roget modellben a melléknevek kategorizációja sokkal kifinomultabb, a 4lang szótárból kinyert címkék másfajta értelmezést rendelnek a szavakhoz. Mivel ebben az esetben a címkék a szótárban szereplő definíciók részei, néhánynak önmagában nincs valódi jelentéstartalma (pl. HAS), viszont a szavakhoz rendelt 10 legközelebbi címkét együttesen vizsgálva a szótárban eredetileg nem szereplő szavakhoz is egy definíció-szerű leírást adnak meg.

A címkézés további jellemzője, hogy mivel a szavak reprezentációja a tényleges nyelvhasználat alapján jött létre, ezért a rendszer kategóriákat is ehhez a fajta reprezentációhoz rendel, nem pedig egy előre definiált tudományos rendszerezés szerint. Tehát például a *macska* szónak több közös címkéje van a *kutya* szóval, mint az *oroszlán* vagy *tigris* szavakkal. Egy biológiai rendszertan természetesen a macskaféléket tekinti közelebbi rokonoknak, azonban a mindennapi életben a háziállat–vadállat megkülönböztetés sokkal jellemzőbb.

5.2. Tulajdonnevek

A szóbeágyazási modellek a létrehozásukhoz használt korpuszban implicit megtalálható világismeretet is hatékonyan tükrözik. Ezért a kategóriacímke-hozzárendelés különböző típusú tulajdonnevek, vagy akár rövidítések esetén is működik. Az

5. táblázatban látható, hogy személynevekhez is releváns címkéket rendel, még akkor is, ha az adott név nem feltétlenül gyakori, de egyértelműen azonosítható. Hasonlóan jól működik a hozzárendelés a különböző szervezetek, intézmények rövidített nevei esetén, ahol még az állami és egyházi oktatási intézmények közötti különbség is megjelenik az *ELTE*, illetve a *PPKE* címkehalmazaiban.

Látható tehát, hogy a módszerünk az olyan szavakhoz is releváns címkéket rendel, amik sem a felhasznált lexikai erőforrásokban, sem az angol szóbeágyazási modellben nem szerepeltek. Az is látszik ezekből az eredményekből, hogy a többszörös transzformáció során sem veszett vagy torzult el a lényegi szemantikai információ jelentős része.

Szó	**TOP 10 kategóriacímke**
Bartók	L: MUSIC.20, MUSIC.71, PERFORMING.12, MUSIC.51, MUSIC.52, MUSIC.54, MUSIC.40, MUSIC.19, LITERATURE.14, MUSIC.41, MUSIC.21
	R: Music.N.5, Music.N.1, Music.N.6, Precursor.N.1, Poetry.N.3, Musician.N.2, Musician.N.5, Lamentation.N.3, Music.N.7, Music.N.9, Poetry.N.2
	F: HAS.27, music.2, art, poem, poet, poetry, WRITE, sound/993.2, text.2, musician, '.7
Obama	L: OFFICIALS.12, GOVERNMENT.17, GOVERNMENT.15, OFFICIALS.13, GOVERNMENT.18, OFFICIALS.10, GOVERNMENT.19, LAW.29, GROUPINGS.10, VOTING.7, GROUPINGS.4
	R: Government.N.14, Politics.N.2, Authority.N.4, Director.N.2, Council.N.2, Politics.N.5, Conduct.N.3, Direction.N.1, Participation.N.1, Government.N.12, Compact.N.2
	F: country.13, government, politician, HAS.22, @United.States, state/76.2, LEAD/2617, place/1026.3, president, republic, country.8
Einstein	L: HARD.SCIENCE.2, PHYSICS.1, PHILOSOPHY.1, MATHS.19, ASTRONOMY.6, LINGUISTICS.14, CHEMISTRY.22, OCCULT.1, ELECTRICITY.6, OCCUPATIONS.3, EDUCATION.14
	R: Heterodoxy.N.5, Scholar.N.2, Experiment.N.2, Smallness.ADJ.2, Intellect.N.7, Conversion.N.3, Production.N.1, Irreligion.N.1, Knowledge.N.1, Life.N.2, Irreligion.N.4
	F: @Karl.Marx, science, man/744.2, atom, scientist, poet, ABOUT.3, NOTPART.OF, prove, exact, politician
ELTE	L: COLLEGE.11, COLLEGE.13, EDUCATION.13, COLLEGE.12, EDUCATION.9, EDUCATION.10, COLLEGE.14, EDUCATION.12, SCHOOL.7, SCHOOL.2, SCHOOL.9
	R: Knowledge.N.2, School.ADJ, Language.N.1, School.N.5, Skill.N.4, Learner.N.4, Teaching.ADJ, Learner.N.3, Evidence.N.4, World.N.3, Receptacle.N.4
	F: educate, institution, study, student, degree, science, AT/2744.27, numbers, atom, GIVE.2, IN/2758.22
PPKE	L: COLLEGE.12, COLLEGE.13, EDUCATION.9, COLLEGE.8, COLLEGE.11, EDUCATION.13, EDUCATION.15, OCCUPATIONS.7, SCHOOL.9, EDUCATION.12, CHRISTIANITY.2
	R: School.ADJ, Knowledge.N.2, School.N.4, Teaching.ADJ, Churchdom.N.6, Churchdom.ADJ.1, Publication.ADJ.2, Churchdom.N.1, Skill.N.4, Evidence.N.4, Learner.N.3
	F: educate, institution, science, group.5, study, student, degree, society/2285.2, sleeve, @Catholic.Church, LEAD/2617
IBM	L: COMPUTERS.33, COMPANIES.3, PLANTS.21, COMPUTERS.34, COMPANIES.2, BUSINESS.BASICS.5, FACTORIES.3, INDUSTRY.3, COMPUTERS.62, COMMUNICATIONS.3, COMPUTERS.27
	R: Servant.N.4, Numeration.N.3, Convexity.N.14, Jurisdiction.N.2, Support.N.7, Merchant.N.3, Action.N.1, Participation.N.2, Receiving.N.2, Receptacle.N.29, Falsehood.N.3
	F: business, factory, computer, IN/2758.22, company/2549, unit.4, INSTRUMENT.5, machines, AT/2744.27, produce, method

5. táblázat. Néhány példa a hozzárendelt címkékre tulajdonnevek és rövidítések esetén a három modellből (L:LDOCE, R:Roget, F:4lang)

5.3. Szubsztenderd nyelvhasználat

Már az eredeti magyar szóbeágyazási modellben is érzékelhető volt, hogy a hasonló annotációs hibát vagy elírást tartalmazó szóalakok egymáshoz közel helyezkedtek el a modellben [22]. Bár az ilyen hibatípusok azonosítása is hasznos funkciója lehet ezeknek a modelleknek, ezek részben elfedik az azonos hibatípusba tartozó szavak közötti szemantikai különbségeket. A kategóriacímkék hozzárendelésekor azonban az ilyen hibás szóalakokhoz is helyes címkéket rendelt a modellünk.

Ugyanez igaz a szleng és más nem sztenderd szóalakokra, amik igen gyakoriak a webről gyűjtött korpusz felhasználói hozzászólásokat, fórumokat tartalmazó részében. Ráadásul ezek gyakran igen erős érzelmi töltetet is tartalmaznak. Ez jól tükröződik az olyan szavakhoz rendelt kategóriacímkékben, mint a *nyugger*, *proli*, *bolsi* vagy *cigó*, amikhez a leggyakoribb címkék például Deceiver',

‘Obstinacy’, ‘Ignorance’, Thief’, CRIME’, ‘POLITICS’ ‘RACE RELATIONS’ ‘PSYCHOLOGY, PSYCHIATRY’, stupid’, criminal’ a mindegyikre illeszkedő ‘person’ mellett.

6. Konklúzió

Bemutattunk egy olyan módszert, melynek segítségével a szóbeágyazási modellekben implicit jelen lévő jelentéscsoportokat emberek által is értelmezhető szimbolikus jegyekké transzformáltuk. A módszer olyan nyelvek esetén is alkalmazható, mint a magyar, amelyekre nem áll rendelkezésre olyan lexikai erőforrás, amelyben szereplő kategóriarendszer közvetlenül felhasználható lenne az osztályozás során. Bemutattuk, hogy egy angol szóbeágyazási modellen átvetítve sem torzul lényegesen az információ, az angol nyelvű erőforrások alapján meghatározott címkék hozzárendelése még tulajdonnevek, rövidítések, illetve a semmilyen külső erőforrásban nem szereplő nem sztenderd szóalakok esetén is jól működik.

Hivatkozások

1. Agirre, E., Martínez, D., de Lacalle, O.L., Soroa, A.: Evaluating and optimizing the parameters of an unsupervised graph-based wsd algorithm. In: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. pp. 89–96. TextGraphs-1 (2006)
2. Bordes, A., Weston, J., Collobert, R., Bengio, Y.: Learning structured embeddings of knowledge bases. In: AAAI (2011)
3. Brown, S.W.: Choosing sense distinctions for wsd: Psycholinguistic evidence. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. pp. 249–252 (2008)
4. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: A unified multilingual semantic representation of concepts. In: Proceedings of ACL-IJCNLP 2015 – Volume 1. pp. 741–751. Association for Computational Linguistics, Beijing, China (July 2015)
5. Chapman, R.: Roget’s International Thesaurus. Harper Colophon Books
6. Chen, X., Liu, Z., Sun, M.: A unified model for word sense representation and disambiguation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1025–1035. Association for Computational Linguistics, Doha, Qatar (October 2014)
7. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press (1998)
8. Kornai, A., Ács, J., Makrai, M., Nemeskey, D.M., Pajkossy, K., Recski, G.: Competence in lexical semantics. In: Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. pp. 165–175. Association for Computational Linguistics, Denver, Colorado (June 2015)
9. Labutov, I., Lipson, H.: Re-embedding words. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics – Volume 2. pp. 489–493. Association for Computational Linguistics, Sofia, Bulgaria (August 2013)
10. van der Maaten, L., Hinton, G.E.: Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research 9, 2579–2605 (2008)
11. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR abs/1309.4168 (2013)

12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013. pp. 3111–3119 (2013)
13. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL 2013. pp. 746–751 (2013)
14. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM 38, 39–41 (1995)
15. Novák, A.: A new form of Humor – mapping constraint-based computational morphologies to a finite-state representation. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)
16. Novák, A., Siklósi, B., Oravecz, C.: A new integrated open-source morphological analyzer for Hungarian. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
17. Orosz, G., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013). pp. 539–545. Hissar, Bulgaria (2013)
18. Panchenko, A.: Best of both worlds: Making word sense embeddings interpretable. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
19. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
20. Rothe, S., Schütze, H.: Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In: Proceedings of ACL-IJCNLP 2015 – Volume 1. pp. 1793–1803. Association for Computational Linguistics, Beijing, China (July 2015)
21. Siklósi, B., Novák, A.: Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. XII. Magyar Számítógépes Nyelvészeti Konferencia pp. 3–14 (2016)
22. Siklósi, B.: Using embedding models for lexical categorization in morphologically rich languages. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016. Springer International Publishing, Cham., Konya, Turkey (April 2016)
23. Summers, D.: Longman Dictionary of Contemporary English. Longman Dictionary of Contemporary English Series
24. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of NAACL 2003 - Volume 1. pp. 173–180. NAACL '03 (2003)
25. Yu, M., Dredze, M.: Improving lexical embeddings with semantic knowledge. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics – Volume 2. pp. 545–550. Association for Computational Linguistics, Baltimore, Maryland (June 2014)