PRÓSZÉKY GÁBOR
Nyelvtudományi Kutatóközpont, Budapest
proszeky.gabor@nytud.hu

Terminológia és neurális hálók

1. Az ember-gép viszonyról

A szóbeágyazási módszerek is használhatók terminológiai célokra. A tanulmány célja az arra vonatkozó első kísérletek és fejlesztési eredmények bemutatása, hogy hogyan lehetséges ezeket a módszereket a magyarra felhasználni. A szóbeágyazás először hasonló fogalmak készletével lát el bennünket, majd klaszterezési módszereket alkalmazunk ezeken, végül a klaszterek reprezentatív elemei más nyelvek (esetünkben az angol) segítségével generálásra kerülnek a már létező forrásaik felhasználásával. Röviden leírjuk, hogyan lehet ezt megvalósítani.

Az ember megszületik, az első pillanattól kezdve hallja a nyelvet, látja a világot, és kezdi érteni az abban előforduló jelenségeket. A számítógép nem ilyen: ha nyelvi információ megértését várjuk el tőle, jelentős hátránnyal indul, ugyanis teljesen kész szövegeket kap, nincs semmi evolúció, ráadásul a gép általában nem ismeri a világot, azaz éppen azt, ami körülveszi. Röviden: a gép sose volt ember – és nem is lesz…

Disztribúciósnak nevezett modellek régóta léteznek a nyelvészetben, de eddig inkább csak egyfajta metaforaként, hiszen az egyes szavak összes szóba jövő környezetét nemcsak felsorolni, de megközelíteni is reménytelennek tűnt. Amikor Wittgenstein (1989) azt javasolta, hogy ha a szavakat meg akarjuk érteni, akkor azoknak ne a jelentését, hanem a használatát keressük, akkor valószínűleg ő se gondolta, hogy ezt a felszólítást a maga konkrét mivoltában a 21. században kézzelfogható közelségbe hozzák a számítógépes nyelvészeti megoldások. Azt eddig is tudtuk, hogy minden nyelvi jelenség csak a környezettel együtt értelmezhető, anélkül nem megy semmi. Ha tehát a meglevő szövegeinkből a gép valahogy ki tudja találni a jelentést és a világismereti relációk egy részét, akkor közelebb kerülünk a gépi szövegmegértéshez.

2. A szóbeágyazási modellekről

A szóbeágyazási modellekben a lexikai elemek vektorok, azaz egy valós vektortér egyes pontjai, ahol az egymáshoz szemantikailag és/vagy

Prószéky Gábor 2023. Terminológia és neurális hálók. In: Prószéky Gábor (főszerk.) – Fóris Ágota – Bölcskei Andrea – B. Papp Eszter – Lipp Veronika (szerk.) *A magyar terminológiastratégia kialakítása. Zöld könyv.* Nyelvtudományi Kutatóközpont, Budapest. 185–197.
DOI: https://doi.org/10.18135/term.2023.10

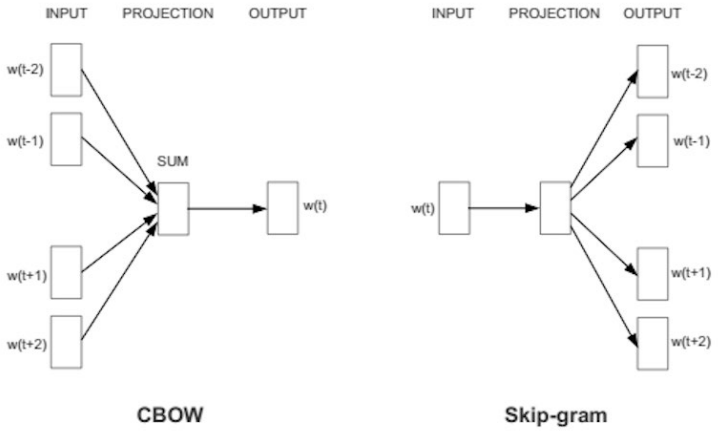
morfológiailag hasonló szavak egymáshoz közel, a jelentésben eltérő elemek egymástól távol esnek. Elvileg ez érthető, de hogy jön létre egy ilyen modell? A szóbeágyazás adatait Mikolov et al. (2013a) módszerével nagy méretű elemzetlen szövegkorpuszokból automatikusan nyerjük. Az eredményül kapott térben két elem jelentésbeli hasonlósága meghatározható a két vektor távolságaként. Az első eredmények igen meggyőzők, bár egy ilyen alapmodellben vannak még problémás esetek, hiszen ezek a modellek például még nem tudják kezelni a többértelműségeket, és csak önálló szavakra működnek, szókapcsolatokra még nem. Persze a szavaknak egy szóba vagy kettőbe írása alapvetően helyesírási kérdés, de témánk szempontjából azért is fontosak a több szóból álló kifejezések, mert a terminológia világa javarészt ilyenekből áll. A szavakból épített vektortérben aztán alapvető vektoralgebrai műveletek is alkalmazhatók, például összeadás és kivonás, amivel a nyelvészetben a 20. század eleje, egészen pontosan Ferdinand de Saussure óta jól ismert, de matematikai módszerekkel gyakorlatilag nemigen modellált analógia fogalma is megfoghatónak tűnik (Saussure 1967).

A legalapvetőbb módszer, amellyel az MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoportja néhány kiinduló kísérletet végzett, a word2vec (Mikolov et al. 2013b). Ennek két alapmodellje a CBOW, ahol a környezetből következtetünk az egyes szavakra és a Skip-gram modell, amivel az egyes szavakból következtetünk a környezetre (1. ábra). A vektoros reprezentációk előállításához alapvetően sok adat, jó algoritmus, hatékony szoftveres megvalósítás, gyors kompilálási idő és intuíció kell a rengeteg paraméter beállításához. Ami a kísérlet megkezdésekor (Siklósi–Novák 2016) rendelkezésre állt, az egy 3 milliárd szavas magyar szövegkorpusz, a word2vec algoritmus (esetünkben a CBOW modellel, 5-5 szavas szókörnyezetre építve, és az előzetes kísérletek tapasztalata alapján 300 dimenziós vektortérrel), valamint elfogadható kompilálási idő, továbbá az az intuíció, amivel az MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport két kutatója, Siklósi Borbála és Novák Attila rendelkezett.

3. Az első kísérletek magyarra

A most következő kísérleteket tehát kiértékelésükkel együtt a kutatócsoport tagjai végezték (Novák–Novák 2017). Először a word2vec algoritmus teljesen elemzetlen szövegkorpuszon futó eredményeit illusztráljuk azzal, hogy egy megadott szóalakra visszakapott környező alakok közül megjelenítjük az első nyolcat. A *kenyerek* szó vektorának lekérdezése olyan vektorokat ad vissza, melyek tartalma a pékek által készített napi ételek világába visz el (2. ábra). Érdemes észrevenni, hogy csak többes számú alakok jelennek meg a lekérdezéshez használt többes számú szó legközelebbi „rokonaiként". Az egyes szóalakokat követő numerikus értékek közül az első a keresőszóhoz való hasonlóságot mutatja, a másik a korpuszbeli

gyakoriságot. Jól látható, hogy más statisztikai módszerekhez képest itt egyértelműen nem a gyakoriság, hanem a hasonlóság a fő szempont. Hasonlóságon ezekben a modellekben a keresőszó és a rokon szó vektorának koszinuszmértékét, azaz a két vektor által bezárt szög koszinuszát értjük.



1. ábra. A word2vec algoritmus két alapmodellje
(Forrás: https://deeplearning4j.org/word2vec)

0	kenyerek	1	2270
1	zsemlék	0.8105	283
2	péksütemények	0.8048	997
3	kekszek	0.7972	1046
4	pékáruk	0.7957	771
5	tészták	0.7881	2466
6	lepények	0.7849	202
7	kiflik	0.7843	349
8	kalácsok	0.7841	277

2. ábra. Köznevek

A módszer előnye, hogy nemcsak a szótárakban felsorolt alakokat, hanem akár a tulajdonneveket is meg tudja jeleníteni a megfelelő morfológiai formában. A 3. ábrán a *Trabantok* keresőszóra többes számú gépkocsinevek jelennek meg (ráadásul, mint látszik, elsősorban a rendszerváltás előtti időszak tipikus gépkocsikéi).

0	Trabantok	1	277
1	Wartburgok	0.8822	142
2	Skodák	0.8569	237
3	Zsigulik	0.8537	111
4	Ladák	0.8511	410
5	Moszkvicsok	0.8506	91
6	Volgák	0.8189	90
7	Daciák	0.7917	115
8	Suzukik	0.7893	272

3. ábra. Tulajdonnevek

A word2vec módszer természetesen akkor is működik, ha a korpusznak nem az eredeti szóalakjai, hanem azok szótövei kerülnek vektorosításra. Így a különböző szavak morfoszintaktikai jegyei nem lesznek már jelen (hiszen ezekre utaltak az aktuális toldalékok), viszont a jelentés még pontosabban körvonalazható, hiszen egy-egy szónak sokkal több előfordulása elemezhető, ha minden alakot visszavezetünk a szótári alapformájára. Ha valakinek a *franciakulcs* tematikusan rokon alakjaira volna szüksége, valószínűleg maga is a 4. ábrán látható szerszámokat sorolná fel.

0	franciakulcs	1	255
1	feszítővas	0.8590	846
2	csavarkulcs	0.8445	473
3	csípőfogó	0.8242	345
4	pajszer	0.8219	567
5	hidegvágó	0.8054	156
6	csavarhúzó	0.7984	4369
7	csőfogó	0.7890	111
8	villáskulcs	0.7890	764

4. ábra. Tövesített korpusz: fogalmi kapcsolatok pontosabb kimutatása

A szóbeágyazási modellek triviális tulajdonsága, hogy minden olyan szóalak, mely általában hasonló szövegkörnyezetben jelenik meg, a felsorolt, rokonnak mondható szavak között lesz. Mivel a módszer a polaritásra nem érzékeny, ezért óvatosnak kell lennünk, hiszen könnyen lehet, hogy egyes szavaknak nemcsak a szinonim, hanem az antonim alakjai is megjelennek a szűkebb környezetben. Ezt mutatja egyfajta illusztrációként az 5. ábra, ahol a *lerombol* szónak nemcsak a rokon értelmű *szétrombol, szétzúz, elpusztít,* hanem a hasonló szövegkörnyezetekben megjelenő, ám ellentétes értelmű *újjáépít* is megjelenik a holdudvarában.

0	lerombol	1	18374
1	szétrombol	0.9202	3158
2	szétzúz	0.8700	3662
3	elpusztít	0.8554	38350
4	újjáépít	0.8423	8664
5	szétver	0.8360	15517
6	újraépít	0.8344	3063
7	feléget	0.8329	4243
8	rombol	0.8175	28932

5. ábra. Hasonló pozíciók: antonimák

Hasonló a helyzet a hibásan írt szavakkal is, hiszen minden olyan szó, mely elütést vagy bármely más helyesírási hibát tartalmaz, szövegkörnyezetét tekintve pontosan ott fordul elő, ahol helyesen írt megfelelői. Nyilván a Google keresőprogramjának az esetleges pontatlanságok kiküszöbölésére vonatkozó *Did you mean …?* kérdése is hasonló felismerésen alapul. Azonban a formai pontatlanságok mellett továbbra is ott vannak a tartalmi hasonlóságok is. Ezért nem meglepő, de talán mosolygásra okot adó hasonlóság a 6. ábrán mutatott és a korpuszban az utolsó karakter levágásával keletkezett helytelen *rövidnac* alak esete: ennek legközelebbi „rokonai" a szövegek tanúsága szerint az ember szélesebb értelemben vett ruhatárába tartozó *pizs, napszemcs* vagy *sap* szavak, és a szintén *i*-hiányos, de szemantikusan távolabbi szavak csak ezek után következnek.

0	rövidnac	1	43
1	pizs	0.7731	180
2	napszemcs	0.7584	37
3	sap	0.7460	374
4	zacs	0.7259	170
5	szemcs	0.7209	37
6	pih	0.7198	149
7	suzuk	0.6943	131
8	nemtomm	0.6795	47

6. ábra. Hasonló pozíciók: elütések

Végül egy megfigyelés, ami némiképp már az eddigi példák alapján is feltűnhetett, hogy a formai és a jelentéstani hasonlóságon túl a stilisztika is megjelenik a csoportosításokban. A 7. ábrán látható három magyar női keresztnév – *Katalin, Eufrozina* és *Kincső* – rokonai egyaránt női keresztnevek, ám *Katalin* környezetében a mai magyar társadalom középkorú hölgyeinek tipikus nevei jelennek meg, míg *Eufrozina* esetében elsősorban a kimondottan régies, szokatlan, ritka, javarészt idegen nevek állnak, *Kincső* mellett pedig a listában a mai névadás népszerű nevei szerepelnek.

0	Katalin	1	88546
1	Zsuzsanna	0.8893	30461
2	Ilona	0.8783	33342
3	Ágnes	0.8750	69813
4	Gabriella	0.8735	27494
5	Judit	0.8730	74435
6	Szilvia	0.8483	18932
7	Ildikó	0.8465	55454
8	Klára	0.8442	22176

0	Eufrozina	1	254
1	Jolánta	0.7732	307
2	Konstancia	0.7679	275
3	Gertrúd	0.7469	1530
4	Eugénia	0.7418	342
5	Adelhaid	0.7410	185
6	Amália	0.7187	1748
7	Sarolt	0.7168	795
8	Gertrud	0.7093	802

0	Kincső	1	1242
1	Csenge	0.8689	4680
2	Evelin	0.8662	3497
3	Bianka	0.8620	4242
4	Fanni	0.8465	10955
5	Kitti	0.8452	6544
6	Cintia	0.8387	1194
7	Villő	0.8358	542
8	Lilla	0.8296	15016

7. ábra. Stilisztikai különbségek: több mint egyszerűen hasonló jelentés

A következőkben a 8–10. ábrák segítségével bemutatunk néhány olyan lekérdezést is a számítástechnika, a biológia, az elektronika, a sörkészítés és a fizika területéről, melyek esetén a rendszer a terminológiai világra oly jellemző egyértelmű szakszavak valamilyen értelemben vett, de szinonimának csak kivételes esetben tekinthető rokonait adja meg.

0	merevlemez	1	22422
1	merevlemez	1.0000	22422
2	HDD	0.9006	7475
3	meghajtó	0.8834	21503
4	háttértár	0.8721	3516
5	memóriakártya	0.8492	6245
6	diszk	0.8484	3693
7	pendrive	0.8300	7121
8	háttértároló	0.8240	1174

0	sejtmembrán	1	1475
1	sejtmembrán	1.0000	1475
2	sejthártya	0.9419	726
3	sejtfal	0.8994	1581
4	riboszóma	0.8420	517
5	mitokondrium	0.8392	1631
6	bélfal	0.8316	1404
7	citoplazma	0.8233	555
8	célsejt	0.8138	284

0	félvezető	1	3342
1	félvezető	1.0000	3342
2	vékonyréteg	0.8208	268
3	polimer	0.8194	4180
4	szilárdtest	0.8098	401
5	tranzisztor	0.7986	4445
6	szilícium	0.7978	3417
7	germánium	0.7594	379
8	dióda	0.7522	3216

8. ábra. A *merevlemez* (számítástechnika), a *sejtmembrán* (biológia) és a *félvezető* (elektronika)

0	sörélesztő	1	642
1	sörélesztő	1.0000	642
2	búzacsíra	0.8492	1306
3	lecitin	0.8396	1157
4	lenmag	0.8297	2093
5	szójalecitin	0.8241	216
6	citromsav	0.8220	1728
7	szójafehérje	0.8180	490
8	szójatej	0.8178	872

0	erjesztés	1	2500
1	erjesztés	1.0000	2500
2	lepárlás	0.8319	1912
3	fermentáció	0.8238	873
4	fermentálás	0.8055	350
5	erjedés	0.7897	4521
6	préselés	0.7897	2332
7	érlelés	0.7878	4652
8	desztilláció	0.7807	631

0	ale	1	310
1	ale	1.0000	310
2	lager	0.7510	279
3	felsőerjesztésű	0.7310	74
4	búzasör	0.7289	609
5	cider	0.7268	663
6	gyömbérsör	0.6886	153
7	lambic	0.6867	52
8	stout	0.6830	75

9. ábra. A sörkészítés területéről három lekérdezés: a *sörélesztő,* az *erjesztés* és az *ale*

0	hologram	1	2852
1	hologram	1.0000	2852
2	holografikus	0.7097	1906
3	lézerfény	0.6994	1134
4	lézersugár	0.6980	1728
5	mátrix	0.6741	16844
6	lézernyaláb	0.6717	405
7	hőkép	0.6631	288
8	röntgensugár	0.6603	1165

0	lézersugár	1	1728
1	lézersugár	1.0000	1728
2	lézernyaláb	0.8997	405
3	lézerfény	0.8966	1134
4	sugárnyaláb	0.8781	622
5	röntgensugár	0.8198	1165
6	lézer	0.8056	14436
7	elektronsugár	0.8006	306
8	nyaláb	0.8001	2333

0	kvantumfizika	1	1964
1	kvantumfizika	1.0000	1964
2	kvantummechanika	0.9156	2939
3	kvantumelmélet	0.8683	1036
4	részecskefizika	0.8591	796
5	húrelmélet	0.8555	591
6	káoszelmélet	0.8555	706
7	kozmológia	0.8343	2063
8	magfizika	0.8220	376

10. ábra. Három keresőszó és rokonai a fizika területéről: a *hologram,* a *lézersugár* és a *kvantumfizika*

4. Csoportosítás klaszterekbe

Azt jól láthatjuk, hogy a rokon szavak nem feltétlenül szinonimák: sokszor hiponimák vagy hiperonimák is lehetnek, illetve olyan szavak, melyeket a hagyományos tezauruszokban sokszor 'related term' viszonynak mondanak. Ezeknek a hasonlósági listáknak az első nyolc elemét mutattuk meg az előző példákban, de a nyolcas szám természetesen semmilyen bűvös határt nem jelent, hiszen van olyan lista, amely sokkal lejjebb is tartalmaz rokon szavakat, míg másoknál már az első nyolc szó közé is keverednek kevésbé jól magyarázható szóalakok. Klaszterezéssel a kapott szóvektorok különféle elemszámú csoportjaiból olyan reprezentációt hozhatunk létre, amelyekben a szavak valamilyen szempont szerint hasonló szemantikai jegyekkel rendelkeznek. Erre mutat példákat a 11. ábra.

Foglalkozások
iró 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 bélésselyem düsesz shantung
posztó szűrposztó abaposztó őzbőr teveszőr kendervászon házivászon háziszőttes
csipke bársony selyem kelme brokát selyemszövet tafota damaszt batiszt

11. ábra. Néhány automatikusan létrejött klaszter

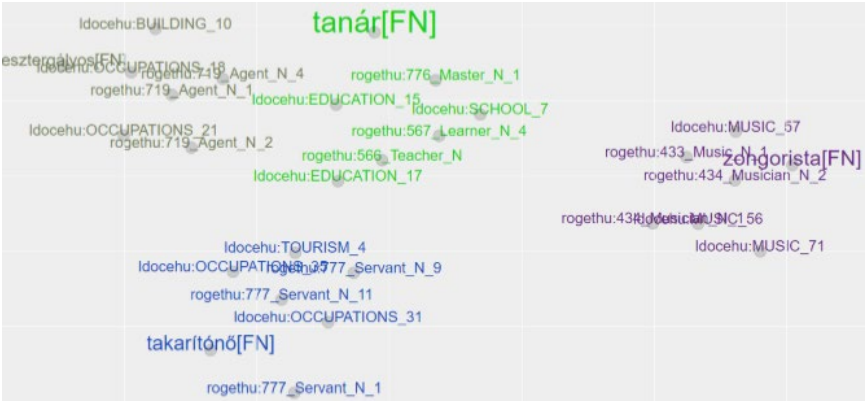
5. A reprezentánselem-generálás felé

A kapott klaszterek tagjai valamilyen értelemben „egyenértékűek", azaz nincs az egész halmazt meghatározó kifejezés hozzájuk rendelve, nincs tehát ún. reprezentánselemük, kategóriacímkéjük. Azonban ilyen elemekre a későbbiekben egy terminológiai adatbázis (fél)automatikus építéséhez szükségünk lesz. A szóbeágyazási modellek természetéből adódóan ilyenkor nem egy előre definiált tudományos rendszertani besorolást szeretnénk érvényesíteni, hiszen a szavak reprezentációjának alapja is azok disztribúciós viselkedése, tehát a tényleges nyelvhasználat.

A célból, hogy legyen egy olyan más nyelvű szövegkorpuszunk is, melyhez több komoly és könnyen hozzáférhető erőforrás létezik, az angol nyelv adja magát. Az angol *Wikipedia* szövegeiből kutatóink létrehoztak egy szófajcímkékkel ellátott és tövesített, összesen 2,25 milliárd szóból álló angol korpuszt (Siklósi 2018). További kísérleteink megmutatták, hogy néhány ismert lexikai erőforrás – pl. a *Roget's Thesaurus* és a *Longman Dictionary of Contemporary English* – kategóriacímkéi szintén megjeleníthetők az angol modell által létrehozott szemantikai térben, amire egy egyszerű példát mutat a 12. ábra.

12. ábra. Angol szavak és kategóriáik egy térben (*turner, teacher, pianist, maid*)

A magyar és az angol szóbeágyazási modellek által definiált szemantikai terekről az is kiderült, hogy jól leképezhetők egymásba: a célnyelvi modellben az eredményvektorhoz közel található szavak az eredeti szó közelítő fordításai (Siklósi–Novák 2016). A cél természetesen nem a pontos fordítások azonosítása volt, hanem a magyar és az angol nyelvű szemantikai tér egymásra illesztése. Sikerült tehát az angol erőforrásokból létrehozott szemantikai kategóriacímkékhez rendelt vektorokat leképezni a magyar nyelvű szóbeágyazási modell terébe. A 13. ábrán egy példa látható, mely néhány magyar szó elhelyezkedését mutatja az angol nyelvi térben.



13. ábra. Magyar szavak az angol szemantikai térben (*esztergályos, tanár, zongorista, takarítónő*)

A 14. ábra alapján képet kaphatunk arról, hogy az olyan nem szótári szavaink, mint a szakszavak vagy a tulajdonnevek, milyen szemantikus, illetve világismereti kategóriákba kerülnek azáltal, hogy a mindenkori szövegkörnyezeteik alapján a magyar térben elfoglalt helyük angol megfelelői az angol fogalmi rendszerekben milyen címkéket kapnak (itt helyszűke miatt csak az első két ilyet mutatjuk). Így lesznek a nyelvi rendszer számára automatikusan elérhetők az olyan – nem feltétlen nyelvi – információk, amik kiolvashatók a táblákból: például, hogy a *Károli,* a *PPKE* és az *ELTE* iskolával és tudással kapcsolatos fogalmak, a *lézer* fénnyel és színnel relációba hozható valami, *Kodály* és *Presser* zenével kapcsolatos személy (de az utóbbiról az is kiderül, hogy aktív zenész is), *Orbán* kormányzati vezető, *Vasarely* festő, a *FIDESZ* kormányalkotó politikai tényező, a *MÜPA* elsősorban zenei szórakoztatással kapcsolatos dolog, az *ÁVÓ* pedig valami olyan, a biztonsággal is kapcsolatba hozható valami, amiről sokaknak a börtön juthat először eszébe.

"Károli"	"PPKE"	"ELTE"
Károli[FN]	PPKE[FN]	ELTE[FN]
568_School_N_2	568_School_ADJ	514_Knowledge_N_2
568_School_ADJ	514_Knowledge_N_2	568_School_ADJ

"lézer"	"Kodály"	"Orbán"	"Vasarely"
lézer[FN]	Kodály[FN]	Orbán[FN]	Vasarely[FN]
441_Luminary_N_6	433_Music_N_1	723_Director_N_2	582_Painting_N_4
447_Color_N_2	433_Music_ADJ	767_Government_N_14	580_Representation_N_9

"Presser"	"FIDESZ"	"MÜPA"	"ÁVÓ"
Presser[FN]	FIDESZ[FN]	MÜPA[FN]	ÁVÓ[FN]
433_Music_N_1	768_Politics_N_2	878_Amusement_N_3	783_Prison_N_7
434_Musician_N_2	767_Government_N_16	433_Music_N_1	693_Safety_N_7
434_Musician_N_1	768_Politics_N_5	665_Store_N_10	777_Servant_N_2
433_Music_N_2	768_Politics_N_1	878_Amusement_N_2	1006_Legality_N_1
433_Music_N_5	776_Master_N_9	198_Abode_N_6	789_Consignee_N_3
623_Poetry_N_2	767_Government_N_14	200_Receptacle_N_4	1008_Jurisdiction_N_1
433_Music_ADJ	954_Pity_N_3	433_Music_N_5	776_Master_N_10
428_Stridor_N_1	725_Council_N_2	921_Ostentation_N_2	552_Information_N_1

14. ábra. Magyar szavak az angol szemantikus tér fogalmi kategóriáival

7. Dinamikus modellek, transzformerek

A szóbeágyazás két hátrányát, a többértelmű szavaknak a különböző környezetekben történő megfelelő értelmezését, valamint a többszavas kifejezések kezelését oldják meg az öt éve megjelent dinamikus reprezentációra épülő ún. transzformer-alapú modellek, melynek két nagy típusa a szövegek elemzését végző enkóder, melynek legismertebb megvalósítása a BERT, azaz a Bidirectional Encoder Representations from Transformers architektúra (Devlin et al. 2018) és a szöveggenerálást végző dekóder, melynek legnevesebb megvalósítása a GPT, azaz a Generative Pre-trained Transformer (Radford et al. 2018). Ezek a modellek képesek a nyelvi kontextus finomabb megértésére és a szemantikai jelentések pontosabb reprezentálására. A terminológiai kutatásban a transzformer-alapú megoldások számos előnyt kínálnak. Először is, a hatalmas mennyiségű szöveges adatokra épített előtanított modellek képesek a korábbiakhoz képest sokkal részletesebb nyelvi tudást működtetni, amelynek eredményeként gazdagabb és pontosabb terminológiai reprezentációkat is képesek előállítani. Figyelemmel tudják kísérni a szókapcsolatokat és jelentős nyelvi környezetre építenek, ezáltal sokkal pontosabban képesek értelmezni a terminusok használatát és összefüggéseit. Finomhangolással vagy a terminológiai adatokkal történő kibővítéssel a modellek testreszabhatók egy-egy konkrét területen előforduló terminológiára. Ez lehetővé teszi a terminológiai adatbázisok és szótárak korábbiaknál hatékonyabb automatikus építését,

valamint a terminológiai definíciók és használati példák automatikus előállítását. A terminológiai fogalmak osztályzása, sőt az egyes osztályok reprezentatív elemeinek a kiválasztása is lényegesen hatékonyabb, mint korábban. Leszögezhető, hogy a közeljövőben a transzformerek által nyújtott megoldások a terminológiai kutatás és fejlesztés területén is hozzá fognak járulni a hatékonyabb és pontosabb munkához.

8. Összefoglalás

Röviden bemutattuk, hogy a szóbeágyazási modellek segítségével az anyanyelvi beszélők által is értelmezhető jegyek állíthatók elő a nyelv egyes szavaihoz, és hogy a magyarra is építhetők ilyen modellek. A morfoszintaktikai, jelentéstani és világismereti hasonlóságok nyers, illetve tövesített korpuszokból megfelelő matematikai módszerek segítségével automatikusan kinyerhetőek. Az adatok klaszterezése után az eredményt a más, nyelvtechnológiailag jobban feldolgozott nyelvekre kialakított erőforrásokkal össze lehet kapcsolni, így olyan nyelvek esetében is sikeres lehet a szemantikai osztályzás, mint a magyar, amelyhez nem állnak rendelkezésre megfelelő lexikai erőforrások. A bemutatott módszerek ráadásul nemcsak a „szokásos" nyelvi elemekre, hanem a tipikus erőforrásokban egyáltalán nem szereplő (nem sztenderd) szóalakokra (amilyenek például a tulajdonnevek vagy a rövidítések) és a normától eltérő alakokra (elütések, helyesírási hibák miatt betű szerint nem is létező szóalakokra) egyaránt jól működnek. Az áttekintés első változatának megjelenése óta létrejöttek az ún. dinamikus nyelvmodellek, melyek számára sem a többértelműség helyes kezelése, sem a több szóból álló kifejezések, vagy akár teljes mondatok hasonló jellegű reprezentáció sem probléma. Ez azt jelenti, hogy ezekkel új lehetőségek jelentek meg a terminológia területén is, hiszen mindezek ott is jól hasznosíthatók, például a fogalmak definícióinak kialakításánál. A kísérletek ebben az irányban is folytatódnak.

Megjegyzés: A tanulmány egy korábbi írás javított, átdolgozott változata. A tanulmány első megjelenésének adatai: Prószéky Gábor 2019. Terminológia és szóbeágyazás. In: Fóris Ágota – Bölcskei Andrea (szerk.) *Terminológiastratégiai kihívások a magyar nyelvterületen.* L'Harmattan Kiadó – OFFI Zrt., Budapest. 47–58. Köszönetemet fejezem ki a kiadóknak, hogy hozzájárulásukat adták a tanulmány bővítéséhez és közléséhez.

Szakirodalom

Devlin, Jacob – Chang, Ming-Wei – Lee, Kenton – Toutanova, Kristina 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805

Mikolov, Tomas – Chen, Kai – Corrado, Greg – Dean, Jeffrey 2013a. Efficient estimation of word representations in vector space. *CoRR,* abs/1301.3781.

Mikolov, Tomas – Yih, Wen-tau – Zweig, Geoffrey 2013b. 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.* ACL, Atlanta. 746–751.

Novák Attila – Novák Borbála 2017. Magyar szóbeágyazási modellek kézi kiértékelése. In: Vincze Veronika (szerk.) *A XIV. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* Szegedi Tudományegyetem, Szeged. 67–77.

Prószéky Gábor 2019. Terminológia és szóbeágyazás. In: Fóris Ágota – Bölcskei Andrea (szerk.) *Terminológiastratégiai kihívások a magyar nyelvterületen.* L'Harmattan Kiadó – OFFI Zrt., Budapest. 47–58.

Radford, Alec – Narasimhan, Karthik – Salimans, Tim – Sutskever, Ilya 2018. *Improving Language Understanding by Generative Pre-Training.* Preprint.

Saussure, Ferdinand de 1967. *Bevezetés az általános nyelvészetbe.* Gondolat Kiadó, Budapest.

Siklósi, Borbála 2018. Using embedding models for lexical categorization in morphologically rich languages In: Gelbukh, Alexander (ed.) *Computational Linguistics and Intelligent Text Processing (Lecture Notes in Computer Science 9623).* Springer, München. 115–126.

Siklósi Borbála – Novák Attila. 2016. Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. In: *A XII. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* Szegedi Tudományegyetem, Szeged. 3–14.

Siklósi Borbála – Novák Attila – Prószéky Gábor 2018. Segíthetnek-e a szóbeágyazási modellek a társadalomtudósoknak? *Magyar Tudomány* 179/7: 945–954.

Wittgenstein, Ludwig 1989. *Logikai-filozófiai értekezés* (ford. Márkus György). Akadémiai Kiadó, Budapest.

Források

Word2Vec, https://deeplearning4j.org/word2vec (Hozzáférés: 2023. július 12.)