Magyar Tudomány 179(2018)7, 945–954
DOI: 10.1556/2065.179.2018.7.3

SEGÍTHETNEK-E A SZÓBEÁGYAZÁSI MODELLEK A TÁRSADALOMTUDÓSOKNAK?

CAN WORD EMBEDDING MODELS HELP SOCIAL SCIENTISTS?

Novák Attila[1], Siklósi Borbála[2], Prószéky Gábor[3]

[1] PhD, Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar,
MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport, novak.attila@itk.ppke.hu
[2] PhD, Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar,
MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport, siklosi.borbala@itk.ppke.hu
[3] az MTA doktora, Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar, MTA Nyelvtudományi Intézet
proszeky.gabor@itk.ppke.hu

ÖSSZEFOGLALÁS

A nyelvtechnológiában az utóbbi néhány évben előtérbe kerültek az olyan disztribúcióalapú szójelentés-reprezentációs modellek, amelyek a szavak jelentésének a szűken vett grammatikai és szemantikai dimenzióin túl a tágabb stiláris, illetve szociolektális (csoportnyelvi) dimenzióit is meglepő pontossággal megragadják. Ezért ezek a mesterséges neurális hálózatokon alapuló szóbeágyazási modellek nemcsak a nyelvtechnológusoknak, sőt nem is csak a nyelvészeknek érdekesek, hanem mindazon tudományágak képviselőinek gazdag tudásforrást jelenthetnek, akik számára a szövegek alapvető nyersanyagként szolgálnak.

A társadalomtudósok számára különösen érdekesek lehetnek azok a szövegek, amiket a különféle közösségi oldalak felhasználói vagy akár az online sajtóhírekhez fűzött hozzászólások szerzői generálnak. Rengeteg ilyen szöveg áll rendelkezésre ma már elektronikus formában, és ez lehetővé teszi, hogy jó minőségű modelleket hozzunk létre a korábban említett technológia felhasználásával, és azokat különböző dimenziók mentén kereshetővé tegyük. Ízelítőként bemutatunk néhány példát a modell által megfogható jelenségek köréből.

ABSTRACT

Distributional models of word meaning have recently become ubiquitous in language technology. These models represent in remarkable detail the meaning of words encompassing not only the narrow grammatical and semantic but also the wider stylistic and sociolectal dimensions. Thus these word embedding models created using artificial neural networks are not only interesting for NLP researchers or linguists, but they can be rich sources of knowledge also for social scientists, for whom texts serve as essential research material.

Texts generated by users of social media sites and comments on articles published on-line at news portals may be of special interest for social scientists. A great amount of such text is available in a digital form, and this makes it possible for us to create high-quality models using the technology mentioned above, and to make them searchable along various dimensions. As a showcase, the paper presents some examples of the phenomena tackled by the model.

Kulcsszavak: szóbeágyazási modellek, neurális hálózatok, disztribúciós szemantika, csoportnyelv, regiszter

Keywords: word embedding models, neural networks, distributional semantics, sociolects, register

DISZTRIBÚCIÓS MODELLEK

A strukturalista nyelvészek az 1930-as években azt az álláspontot fogalmazták meg, hogy a nyelvi tudás elsődleges forrása a szavak és morfémák disztribúciója. Ennek bizonyítására azonban csak napjaink új tudományos eredményeinek felhasználásával adódott lehetőség. Napjaink digitális társadalma nagyon nagy mennyiségben állít elő újabb és újabb szöveges tartalmakat, melyekben a nyelv alakulása, illetve a nyelvhasználat különböző rétegei jól tetten érhetők.

A disztribúciós szemantika a strukturalisták által korábban megfogalmazott elvet olyan formában fogalmazza újra, hogy a szavak jelentése szorosan összefügg azzal, hogy milyen kontextusban használjuk őket (Firth, 1957). Az egészen a közelmúltig egyeduralkodó hagyományos számítógépes disztribúciós szemantikai modellek létrehozásakor az egyes szavakhoz tartozó reprezentáció ténylegesen az adott szó előre meghatározott méretű környezetében előforduló szavak egy nagy korpuszból számított előfordulási statisztikáit tartalmazta. Ezek a modellek – annak ellenére, hogy bizonyos eredményeket elértek – nem igazán váltották be a hozzájuk fűzött reményeket. Ezzel a fajta reprezentációval az az egyik fő probléma, hogy a legtöbb szó környezetében a legtöbb másik szó soha nem fordul elő, ezért az együttes előfordulásokat ábrázoló mátrix „ritka", ugyanis legtöbb pozíciójában 0 áll.

Az áttörést napjaink nyelvtechnológiai kutatásainak egyik kurrens módszere hozta, amely a szövegek alapján mesterséges neurális hálózatok alkalmazásával folytonos vektortérbeli tömör reprezentációkat, ún. *szóbeágyazásokat* (word embedding) hoz létre. Az alapgondolatot Yoshua Bengio és munkatársai vetették fel a 2000-es évek elején (Bengio et al., 2003), de a hatékony gépi háttér igazán csak a 2010-es években tette lehetővé az igazán nagy méretű modellek betanítását. A tanítás során az egyes szavak fix méretű környezetében szereplő többi szót vesszük figyelembe, az ezekből álló vektor azonban egy neurális hálózat bemenete. A környezetben álló szavak összességét reprezentáló vektorokat használja a hálózat arra, hogy megjósolja az adott környezetben legvalószínűbb célszót. Szemben a hagyományos számolásalapú módszer milliónyi dimenziós ritka mátrixaival, az így létrehozott pár száz dimenziós vektorok mindegyik pozíciójában egy –1 és +1 közötti, szinte minden esetben 0-tól különböző szám szerepel. Az egyes dimenzióknak nincs saját jelentésük, hanem a hálózatot alkotó mesterséges

idegsejtek közötti kapcsolatok erősségét reprezentálják. A tanítás során a rendszer összehasonlítja a hálózat által a környezet alapján jósolt szót az ott ténylegesen szereplővel, és a hiba visszaterjesztésével, illetve ennek megfelelően a környezetet reprezentáló vektorok frissítésével jön létre a tanítás végén 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.

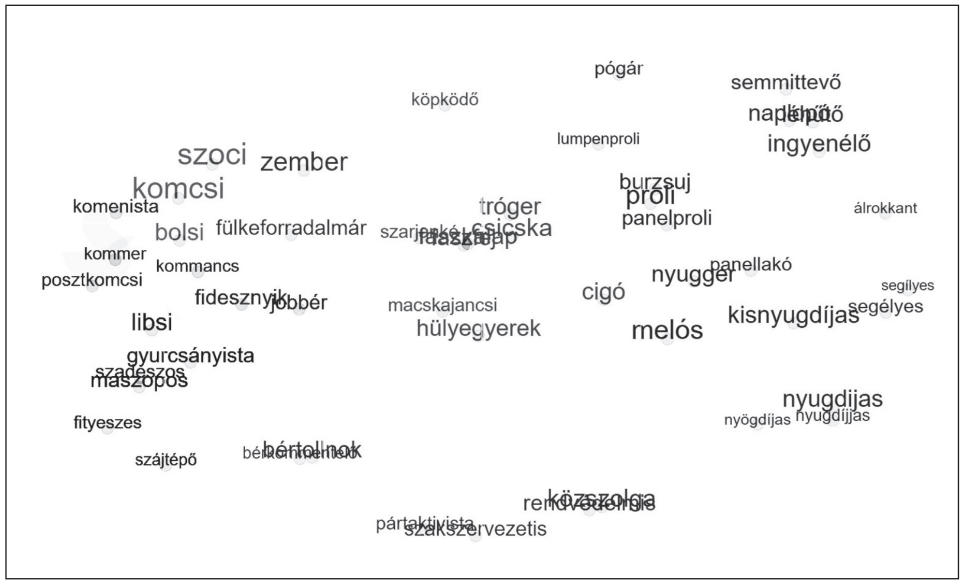
Ebben a rendszerben a lexikai elemeket egy valós vektortér egyes pontjai reprezentálják, melyek konzisztensen helyezkednek el az adott térben, azaz az egymáshoz szemantikailag és/vagy morfológiailag hasonló szavak egymáshoz közel, a jelentésben eltérő elemek egymástól távol esnek. Mindemellett vektoralgebrai műveletek is alkalmazhatók ebben a térben, tehát két elem szemantikai hasonlósága a két vektor távolságaként meghatározható, illetve a lexikai elemek pozícióját reprezentáló vektorok összege jó közelítéssel azok jelentésének összegét határozza meg (Mikolov et al., 2013a, 2013b). 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 (Banea et al., 2014; Iacobacci et al., 2015; Trask et al., 2015).

A szóbeágyazási modellek hatékonyan ragadják meg a szövegekben megjelenő szemantikai információkat, sőt jelentős mennyiségű világismereti tudást is (Mikolov et al., 2013a). Ezek a beágyazási modellek magyar nyelvre is jó eredménnyel működnek kellő méretű és elemzett tanítóanyag alkalmazása esetén (Siklósi–Novák, 2016; Siklósi, 2018).

A KORPUSZ ELŐKÉSZÍTÉSE ÉS A MODELLEK LÉTREHOZÁSA

Egy nagyméretű, több mint egymilliárd szavas, a webről gyűjtött korpuszból hoztunk létre szóbeágyazási modelleket. A korpuszt automatikusan egyértelműsített morfológiai elemzéssel láttuk el. A modell építésekor nem a ragozott szavakat, hanem a szótöveket tartottuk meg, melyek után külön elemként szerepeltek a morfológiai elemző által generált címkék. Mivel ezek a címkék az aktuális szó környezetében maradtak, az általuk hordozott szintaktikai információ továbbra is szerepet kapott az egyes szavakat reprezentáló vektorok létrehozásában. Azonban mivel a modell csak szótöveket tartalmaz, így robusztusabb modell jön létre, mint ha közvetlenül a szövegben szereplő felszíni szóalakokból építenénk a modellt, mert egy-egy szó reprezentációjának kiszámításához annak minden ragozott alakja hozzájárul. Ez a ritka szavak esetében jelentősen javítja a modell minőségét.

Bár a nyelvészeket a keletkezett vektortér grammatikai és szemantikai dimenziói érdeklik leginkább, ám a modell a tágabb stiláris, illetve szociolektális (csoportnyelvi) dimenziókat is meglepő pontossággal ragadja meg. Különböző rétegnyelvek és szaknyelvek speciális szókincse önállóan bejárható régiókként jelenik meg a sokdimenziós térben, amelynek szemléltetésére a vektorokat két dimenzióba leképező vizualizációs eljárást implementáltunk. Az *1. ábrán* a vektortérben a *nyugger* szóhoz legközelebb elhelyezkedő szavak elrendezése látható (a szavak mérete a korpuszbeli gyakoriságukkal arányos). A megjelenített néhány példa is jól szemlélteti egyrészt az érzelmileg túlfűtött politikai kommentelő stílus jellegzetességeit, másrészt az ábrán az is látszik, hogy már ez a néhány szó is jól elkülönülő tematikus-szemantikai csoportokba rendeződik.



1. ábra. A rétegnyelvi *nyugger* szó környezete a vektortérben

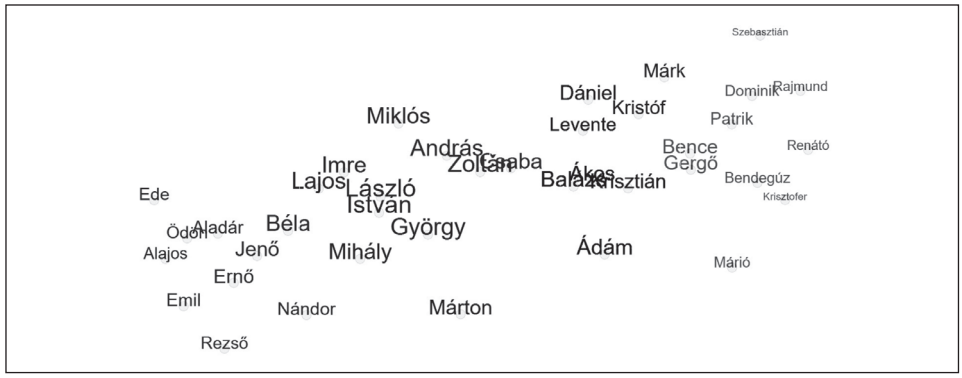
Léteznek olyan technikák is, amelyeknek segítségével a vektortérmodellben a benne szereplő szavak mellett olyan szimbolikus jegyek is elhelyezhetőek, amelyek az adott régióban lévő szavakat valamilyen (szemantikai, stiláris, szociolektális, szakterületi, helyesírási, grammatikai stb.) szempontból jellemzik. Ezeket a jegyeket konkrét szavakhoz, esetleg teljes szövegekhez vagy szövegrészekhez rendelve és ezeket a korpuszhoz külön annotációs szinten hozzáadva a korpusz nemcsak a benne előforduló konkrét szavak, hanem az így hozzáadott annotáció szempontjából is kereshető lesz.

NYELVI RÉTEGZŐDÉS

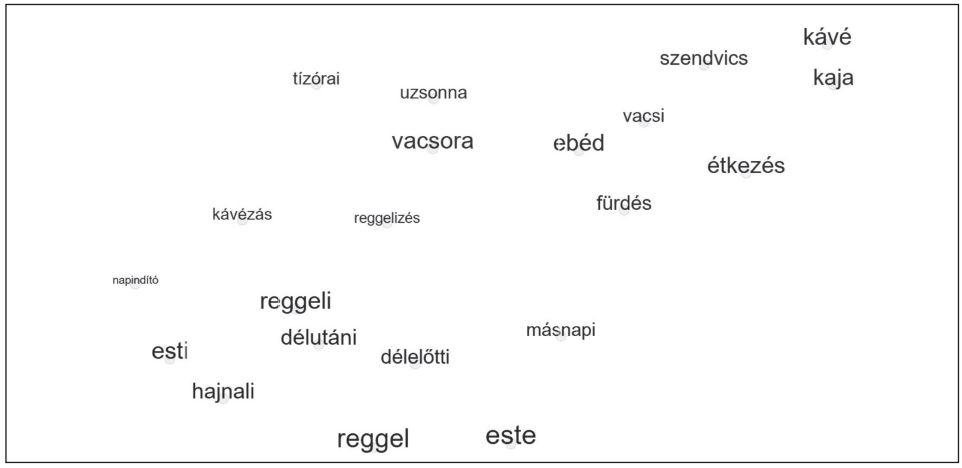
A modellből lekérdezhető a benne szereplő szavakhoz legközelebb elhelyezkedő további szavak listája, az adott szótól való távolság szerint rendezve. Ezt a műveletet a már megjelenített elemek egy részhalmazán folytatva feltérképezhető az adott régió szókincse. A nyelvi rétegek és a rétegnyelvek példátlan gazdagságban és árnyaltságban jelennek meg a lexikai térben, kezdve az *online* játékok rajongóinak zsargonjától a *fanfiction* irodalmat felvonultató fórumok látogatóinak speciális szóhasználatán keresztül a szemészeti szaknyelv rétegein át egészen a vasúti irányítórendszerek szakterminológiájáig. Az így létrejött szólistán automatikus klaszterezési eljárást alkalmazva további tematikus osztályozást végezhetünk, illetve kiszűrhetjük az oda nem illő elemeket. Az *1. táblázat*ban a fenti kategóriákból választott *kempel, ficc, macula* és *balíz* szavakhoz kérdeztük le a modellből a hozzájuk legközelebb eső első néhány szót. A terjedelmi korlátok miatt itt csak a listák elejét van lehetőségünk bemutatni, azonban általánosan elmondható, hogy az ilyen listáknak akár még a többszázadik elemei is releváns kifejezéseket tartalmaznak, amelyek természetesen adott esetben már lazább kapcsolatban állnak az eredeti szóval. A vektortérben olyan típusú kategóriák is elkülönülnek, amilyen típusú megkülönböztetés semmilyen létező szótárban nem szerepel, és sokszor megfelelő elnevezést sem könnyű találni az adott kategória számára. A *2. ábrán* látható például, hogy világosan elkülönülnek a férfi keresztneveken belül az „avítt dzsentrinevek", a hagyományos keresztnevek, az átlagos gimnáziumi osztálynévsorban fellelhető trendi fiúnevek és a roma kiskorúak divatos angolszász–újlatin keresztnevei.

1. táblázat. A különböző rétegnyelvekből való *kempel, ficc, macula, balíz* szavak és a hozzájuk legközelebb eső néhány szó a vektortérben

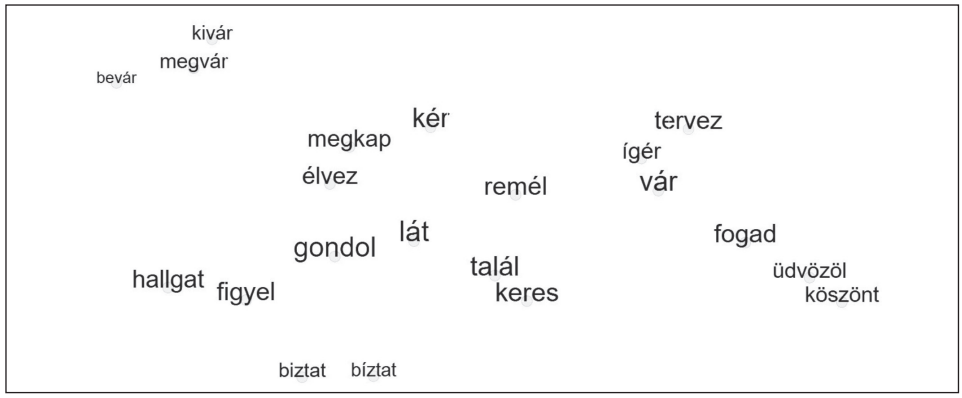
kempel	ficc	macula	balíz
wowozik	fic	sárgafolt	balízcsoport
farmol	fici	degeneratio	vezérlőjel
fearless	fanfic	atrophia	főjelző
healel	törid	glaukóma	transzponder
VF-ezik	ficu	látóidegfő	vágányút
hackel	drarry	szürkehályog	vezérlőegység
maxol	fanfiction	makula	EVC
castol	sztory	ideghártya	jelsorozat
turret	snarry	látóhártya	menetengedély
leöl	SSHG	zöldhályog	kijelzés
sentry	oneshot	centralis	DMI
questel	feji	látóideg	vezérlőközpont
betámad	függővég	glaucoma	riasztóközpont
lewarezol	manga	naevus	komparátor
limpel	dorama	erythema	nyugtázás



2. ábra. Néhány férfinév elrendeződése a vektortérben



3. ábra. A többértelmű *reggeli* szó és környezete a vektortérben



4. ábra. A többértelmű *vár* szó környezete a vektortérben

2. táblázat. Néhány kultúraspecifikus szó képéhez legközelebb eső szavak az angol szóbeágyazási modellben

busó	pörc	cigó
reveler	bacon	thug
reveller	dough	strikebreaker
parade	sauce	racist
re-enactor	sliced	troublemaker
clown	gravy	Palestinians
townspeople	soup	rioter
carnival	curd	hoodlum
festival-goer	steak	Tutsis
townsfolk	stew	Jew
villager	pastry	Arab
onlooker	tortilla	bigot
festivity	lard	whites
mummer	butter	fascist
maypole	flatbread	drunk
procession	mayonnaise	bookie

A webről gyűjtött korpusz gazdagon tartalmaz olyan a „nép" által írt szövegeket, amelyek a különböző webes fórumokon és a cikkekhez írt hozzászólásokban jelennek meg. Ezekben a szövegekben – és következésképpen a szemantikai vektortérben – a szókincs olyan rétegei jelennek meg (vagy egy épp ebből a rétegből vett kifejezéssel: *figyelnek be*), amelyek nyomtatott szótárakban nem szerepelnek. A modell ezeknek a szavaknak az adott szociolektális közegben szokásos jelentését is megragadja, így alkalmas lehet az adott rétegnyelv vizsgálatára, az abban való elmélyülésre (lásd az *1. ábrát*).

DOMÉNADAPTÁCIÓ ÉS -SZELEKCIÓ

Ahhoz, hogy jó minőségű modellek jöjjenek létre, a rendszernek nagy mennyiségű tanítóanyagra van szüksége. Az általunk vizsgált korpusz több milliárd szóból áll. Ha egy adott réteg- vagy szaknyelv szókincsét szeretnénk vizsgálni, akkor nem feltétlenül elegendő a modell betanításához csak az adott nyelvi réteget reprezentáló korpusz, hanem a nagyobb általános korpuszon kapott modellből kiindulva a rendszert az adott szakkorpuszon tovább tanítva létrehozható egy olyan lexikális reprezentáció, amelyben a köznyelvben dominánsan az adott rétegnyelvtől eltérő jelentésben használt szavak reprezentációja a rétegnyelvben domináns jelentéshez közelít. A rendszer tehát arra is használható, hogy egy nagyobb vegyes korpuszból egy adott rétegnyelvet reprezentáló részkorpuszt válasszunk ki annak a rétegnyelvre jellemző lexikai elemei alapján. Ehhez kiindulásként elegendő a jellemző

terminológiának csak néhány elemét megadni, majd az adott vektortérrégió közeli elemeiből automatikusan egy bővebb szakterminológiai szókincset összeállítva és ezt lekérdezve az egész korpuszból kiválaszthatjuk a releváns részkorpuszt.

MATEMATIKAI TRANSZFORMÁCIÓK A VEKTORTÉREN

A disztribúciós modellbeli távolságmérték önmagában általában nem választja el egymástól a hasonló jelentésű, de különböző polaritású elemeket, mint például *jó-rossz, szép-csúnya,* illetve ezek hasonló jelentésű társait, azonban az ellentétpárokra adott példák alapján általában definiálható egy olyan transzformáció a téren, amely olyan forgatást végez, amelyet alkalmazva a vektortér valamelyik dimenziója mentén az ellentétes polaritású elemek szétválnak. Tehát bár az eredeti vektortérmodellben az egyes dimenziókhoz általában nem rendelhető semmiféle jelentés, megfelelő transzformáció után a transzformált vektortérben egy adott dimenzió specifikus jelentést nyerhet.

Egy másik probléma a homonim alakok kezelése. Bár a vektortérmodell a többjelentésű elemekhez egyetlen reprezentáló vektort rendel, ez nem feltétlenül jelenti azt, hogy ne lenne kinyerhető a modellből az egyetlen vektorban reprezentált jelentéshalmaz megfelelő gépi tanulási algoritmusok alkalmazásával. Problémát csak azok az esetek jelentenek, amikor túl sok különböző jelentése van egy szónak, illetve amikor valamelyik jelentés nagyságrendekkel gyakoribb, mint a többi. A *3. ábra* azt szemlélteti, hogy a modell a *reggeli* szónak mind az 'étkezés', mind a 'napszaki' jelentését megragadja, ugyanakkor a *vár* igei használata annyival gyakoribb, mint a főnévi, hogy a főnévi jelentés alig jelenik meg a modellben *(4. ábra)*. Az utóbbi problémára ugyanakkor megoldást jelent, ha morfológiailag annotált korpuszból építjük a modellt: ekkor két különálló vektor reprezentálja a szó igei, illetve főnévi használatát.

TÖBBNYELVŰSÉG

További érdekes lehetőségek nyílnak annak a ténynek a kiaknázásával, hogy a különböző nyelveken készített szóbeágyazási modellek topológiája általában hasonló, ezért akár néhány ezer fordítási szópár megadásával viszonylag pontos leképezés definiálható két különböző nyelvhez készült modell között. Ez lehetővé teszi egyrészt a két nyelv „rokon" lexikális mezői közötti leképezést és az egyik oldalról kiindulva a másik oldal felfedezését, másrészt a kultúraspecifikus szavaknak (például: *busó, pörc, cigó* stb.) a másik nyelven megfelelő terület megvizsgálását. Emellett a leképezés azt is lehetővé teszi, hogy az egyik nyelven hozzáférhető (akár kézzel, sok munkával létrehozott) lexikai erőforrás a másik nyelven is használha-

tóvá váljon. A *2. táblázat*ban a *busó, pörc, cigó* szavaknak megfelelő vektorok által meghatározott pontokhoz legközelebb eső angol szavak láthatóak az angol *Wikipédiá*ból létrehozott szóbeágyazási modellben. Látható, hogy a modell megragadja és leképezi a busójárás fesztiváli hangulatát, vidéki látványosság jellegét, a *pörc* szóról pedig megtudhatjuk, hogy denotátuma étel, míg a *cigó* szó leképezésével kapott listában megjelennek mind a bűnözéssel kapcsolatos, illetve az etnikai intoleranciára utaló szavak, mind a kurrens etnikai ellentétekkel kapcsolatban gyakran felmerülő nemzetiségnevek. Ugyanakkor az angol *Wikipédiá*ból készült korpusz nemigen tartalmaz olyan jellegű csoportnyelvi elemeket, amilyenek közé a magyar *cigó* szó tartozik, ezért az adott esetben a magyar szó és az angol modellbeli képe között nincs pontos regiszterbeli megfelelés. Egy általános angol nyelvű webkorpuszból készült modell esetén azonban nem állna fenn ez a probléma.

ÖSSZEFOGLALÁS

Írásunkban bemutattunk néhány olyan lehetőséget, amelyet a nagyméretű korpuszokból neurális hálózatok segítségével épített szóbeágyazási modellek a szövegekre alapozott kutatásokat végző társadalomtudósok számára megnyitnak. Megpróbáltuk néhány példával illusztrálni, hogy ezek a modellek igen árnyalt módon képesek megragadni a szavak és a hozzájuk kapcsolódó fogalmak tágabb értelemben vett jelentésével kapcsolatos nyelvi szinten tetten érhető tudást, beleértve a stiláris, rétegnyelvi, szakterületi jellemzőket. Az ígéretes lehetőségeknek egy része még csak most körvonalazódik, hiszen a bemutatott megoldások csak néhány éve jelentek meg. Egészen pontosan: a matematikai módszerek nagy része korábban is megvolt, csak a hatékony működtetésükhöz szükséges számítástechnikai háttér nem volt meg. A cikkünkben vázolt modellek alapfogalmai, tehát a vektoros reprezentáció, a neurális hálók vagy a mélytanulás napjainkban a legtöbb területen, így a társadalomtudományi kutatások területén is új lehetőségeket nyitnak. Ezek kiaknázásához időszerű a tanuláselmélet, a nyelvtechnológia és azon társadalomtudományi területek kutatóinak összefogása, ahol a szövegekben megbújó tudás efféle feldolgozása egyre újabb és egyre hasznosabb tudományos megoldások kialakítását teszi lehetővé.

KÖSZÖNETNYILVÁNÍTÁS

A cikkünkben bemutatott eredmények részben az FK 125217 éa a PD 125216 számú projekt keretében a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással az FK 17 és a PD 17 pályázati program finanszírozásában megvalósuló kutatások keretében születtek meg.

IRODALOM

Banea, C. – Chen, D. – Mihalcea, R. – Cardie, C. – Wiebe, J. (2014): Simcompass: Using Deep Learning Word Embeddings to Assess Cross-level Similarity. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014).* Dublin: ACL, 560–565. https://pdfs.semanticscholar.org/4b7b/10ffe383addfc134fb5b10000d085ffd9709.pdf

Bengio, Y. – Ducharme, R. – Vincent, P. – Jauvin, C. (2003): A Neural Probabilistic Language Model. *Journal of Machine Learning Research,* 3, 1137–1155. http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Firth, J. R. (1957): A Synopsis of Linguistic Theory, 1930–1955. *Studies in Linguistic Analysis*, 1–32. http://annabellelukin.edublogs.org/files/2013/08/Firth-JR-1962-A-Synopsis-of-Linguistic-Theory-wfihi5.pdf

Iacobacci, I. – Pilehvar, M. T. – Navigli, R. (2015): 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*). Beijing: ACL, 95–105. http://www.aclweb.org/anthology/P15-1010

Mikolov, T. – Chen, K. – Corrado, G. – Dean, J. (2013a): Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, https://arxiv.org/pdf/1301.3781.pdf

Mikolov, T. – Yih, W. – Zweig, G. (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.* Atlanta: ACL, 746–751. https://www.aclweb.org/anthology/N13-1090

Siklósi B. (2018): 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, Cham, 115–126. https://link.springer.com/chapter/10.1007/978-3-319-75477-2_7

Siklósi B. – Novák A. (2016): Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. In: *A XII. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged: SZTE, 3–14.

Trask, A. – Michalak, P. – Liu, J. (2015): sense2vec - A Fast and Accurate Method for Word Sense Disambiguation in Neural Word Embeddings. *CoRR* abs/1511.06388, https://www.researchgate.net/publication/284476537_sense2vec_-_A_Fast_and_Accurate_Method_for_Word_Sense_Disambiguation_In_Neural_Word_Embeddings