Szóbeágyazási modellek vizualizációjára és böngészésére szolgáló webes felület

Novák Attila[1,2], Siklósi Borbála[2], Wenszky Nóra[1]

[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 Cikkünkben egy word2vec alapú szóbeágyazási modellek vizualizációjára és böngészésére szolgáló webes felületet mutatunk be, amelybe a modellek lekérdezésén és vizualizációján túl számos komplex funkciót integráltunk. A webes felületen keresztül elérhető funkciók nagy méretű magyar és angol nyelvű korpuszból épített szóbeágyazási modelljeinkre épülnek.

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 ragadni. Angol nyelvre egyre népszerűbb a kézzel gyártott szimbolikus lexikai erőforrások és a nyers szövegből tanulható ritka diszkrét reprezentációk helyett a folytonos vektorreprezentációk alkalmazása, melyek hatékonyságát a neurális hálózatokra alapuló implementációk használatával több tanulmány is alátámasztotta [3,5,1,7]. Ezekben a kísérletekben és alkalmazásokban azonban a leírt módszereket általában a magyarhoz képest jóval kevesebb szóalakváltozattal operáló, kötött szórendű angol nyelvre alkalmazzák. Korábban megmutattuk, hogy az összetett morfológiájú nyelvek esetén is jó minőségű szóbeágyazási modell hozható létre a tanítókorpuszra alkalmazott megfelelő előfeldolgozás (a szavak külön tő- és morfológiaicímke-tokenekre bontása) esetén [8,9].[3]

A beágyazási modellek kiértékelésének egyik módszere az angol nyelvű modellek esetén az analógiatesztek elvégzése [4]. Ezeknél a teszteknél egy szópárosból és egy tesztszóból indulnak ki. A rendszer feladata annak a szónak a megtalálása, ami a 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 felel meg a *király* és a *királynő közötti távolságnak* (aminek oka a *király* szó többértelműsége is).

[3] A nyers szövegekből építettnél jobb.

Így csupán elvétve találtunk olyan analógiapéldákat, melyek helyes eredményt adtak.

A kvantitatív kiértékelés nehézsége ellenére is szerettük volna megvizsgálni a különböző módon létrehozott beágyazási modellek minőségét, illetve szimbolikus szemantikai tudást kinyerni belőlük. Ehhez létrehoztunk egy olyan webes felületet, aminek segítségével a modellek tartalma áttekinthető és könnyen kezelhető formában jelenik meg, illetve amibe további, a szóbeágyazási modellek értelmezhetőségét és felhasználhatóságát támogató megjelenítési formát és eszközt is integráltunk.

Cikkünkben a webes felület funkcióit és az eszköz segítségével feltárható jelenségeket mutatjuk be. A bemutatott demó a nagyközönség számára egyelőre nem érhető el, aminek oka elsősorban technikai, de a jövőben tervezzük a nyílt hozzáférés lehetőségét is megvalósítani.

2. Hasonló szavak lekérdezése

Egy szóbeágyazási modellben a lexikai elemek egy valós vektortér egyes pontjai, melyekben 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 [5,3].

Ezért a beágyazási modellek egyik alapvető funkciója egy tetszőleges szóhoz a modellben legközelebb elhelyezkedő szavak meghatározása a szavakat reprezentáló vektorok koszinusz távolsága alapján. A webes felületen lehetőség van arra, hogy egy tetszőleges szót beírva lekérdezzük a közelében található tetszőleges számú szót egy adott modellben. Jelenleg a felületen több modellt is tesztelni lehet:

- Felszíni szóalakokat tartalmazó modell (hu.surf): a nyers korpuszból tokenizálás után tanított modell, amiben a szavak toldalékolt alakja szerepel (magyar nyelvű modell).
- Tövesített alakokat tartalmazó modell (hu.ana): a modell építése előtt szófaji egyértelműsítést és lemmatizálást alkalmaztunk a korpuszra, majd a morfoszintaktikai információkat külön tokenként, a szótövek kontextusaként tartottuk meg (magyar nyelvű modell). Ilyen modellt tudomásunk szerint elsőként mi készítettünk: [8,9].
- Szófaji egyértelműsített modell (hu.pos): az előző modellhez hasonló, azzal a különbséggel, hogy a fő szófajcímkéket a szótövekre ragasztva tartottuk meg, így az azonos alakú, de más szófajú szavakhoz külön reprezentációt rendelt a modell (magyar nyelvű modell). Szófajcímkékkel annotált korpuszból angol nyelvre készítettek már beágyazási modelleket [10], de sem tövesítést, sem a ragok külön tokenekként való ábrázolását korábban nem alkalmazták.

– Angol nyelvű Wikipédia modell (wikien.pos): az angol Wikipédiából az előző modellnek megfelelő előfeldolgozással létrehozott modell (angol nyelvű modell). Itt is alkalmaztunk tövesítést is az előfeldolgozásnál, a ragokat külön token reprezentálja.
– Szemészeti kifejezéseket tartalmazó modell (szem.ana): az eredeti korpuszból épített magyar modell leszűkítése egy szemészeti korpusz szókincsére (magyar nyelvű modell)
– Lexikai erőforrások szemantikai kategóriáit tartalmazó modellek (4lang, ldocehu, rogethu): három angol nyelvű lexikai erőforrásból (4lang, Longman Dictionary of Contemporary English, Roget's Thesaurus) épített modell [6], mindegyikben az egyes kategóriákhoz felsorolt példaszavakat azok vektorainak átlagolásával és átlagvektor magyar vektortérbe vetítésével létrejött reprezentációval ábrázoljuk (angol nyelvű címkéket a magyar nyelvű vektortérben megjelenítő modellek). Ezeket a modelleket használjuk az 5. és a 6. részben említett funkciók megvalósításánál.

A modellek létrehozásának részletei a következő cikkekben olvashatók: [8,6].

Az egyes lekérdezésekre kapott válaszban a kérdőszóhoz mért távolság alapján rendezve jelenik meg a szólista, amiben szerepel az egyes elemek korpuszbeli gyakorisága és a hasonlóság mértéke (koszinusztávolság) is. Az 1. ábrán két lekérdezés eredménye látható. Az első listában az elemzett hu.ana modellből kérdeztük le az *alma* szóhoz legközelebb eső 10 szót, míg a második listán a nyers korpuszból létrehozott hu.surf modellből a *kenyerek* szóalakhoz tartozó 10 leghasonlóbb szóalak látható. A második esetben a hasonlóság nem csak szemantikai, hanem morfoszintaktikai vonatkozásban is teljesül (a korpuszban ritkábban előforduló szavak esetén azonban az utóbbi modell kevésbé jól használható).

0	alma	1	63906
1	körte	0.8392	13339
2	eper	0.8356	16159
3	banán	0.8222	17732
4	szilva	0.8046	12602
5	őszibarack	0.8011	4698
6	uborka	0.7971	14735
7	répa	0.7937	14107
8	cseresznye	0.7848	11676
9	ananász	0.7820	4827
10	dinnye	0.7689	11428

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
9	sonkák	0.7836	613
10	pogácsák	0.7787	539

1. ábra. Példa hasonló szavak lekérdezésének eredményére a tövesített és a felszíni alakokat tartalmazó modellből

A lekérdezésekre kapott listák elemeit interaktív módon (egérkattintással) is kiválaszthatjuk, ekkor a kattintott szóhoz legközelebbi elemek listáját is megkapjuk a beállított modellből. A lexikai erőforrások szemantikai kategóriáit tartalmazó modellek (4lang, ldocehu, rogethu) kiválasztása esetén a rendszer magyar szavak beírásakor a vektortérben az adott modellben legközelebbi címkéket adja vissza, azokra kattintva pedig fordítva az adott címkéhez legközelebbi szavak jelennek meg.

3. Klaszterezés és vizualizáció

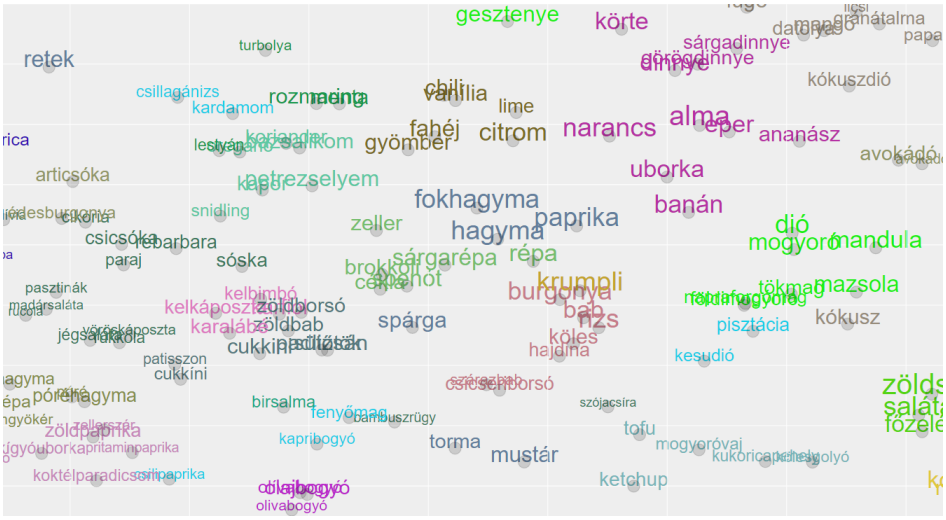
Ha egy szólistában szereplő szavak egymáshoz való viszonyát szeretnénk megjeleníteni, akkor a felület lehetőséget ad arra is, hogy egy listát megadva, az abban szereplő szavakhoz a kiválasztott modell által rendelt vektorok alapján azokat klaszterezve, csoportosítva jelenítsük meg. Így az egymáshoz közel álló szavak azonos, míg a távolabbi szavak külön klaszterben jelennek meg (az elkülönítés érzékenysége állítható paraméter). Az alkalmazott algoritmus részleteit lásd: [8]. A klaszterezésre szánt listát természetesen úgy is előállíthatjuk, hogy az előző pontban ismertetett módon egy kiinduló szóból elindulva az annak a közelében található szavak listáját (vagy akár több ilyen listát egyesítve) csoportosítjuk az eredményt. Ezzel a módszerrel könnyen kiszűrhetjük az esetleg zajként megjelenő találatokat, vagy egy hosszabb listát szemantikailag releváns alcsoportokra bonthatunk.

A fogalmakat reprezentáló vektorok egy szemantikai térben helyezik el az egyes lexikai elemeket, így ez a szerveződés látványosan vizualizálható. Ehhez a listában szereplő szavakhoz tartozó sokdimenziós vektorokat egy kétdimenziós térbe képeztük le a t-sne algoritmus alkalmazásával [2]. 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. A 2. ábra egy ilyen leképezés részletét ábrázolja.

4. Analóg szókapcsolatok megjelenítése

Bár a felületen jelenleg kipróbálható modellekben csupán szavak reprezentációja található meg (természetesen a felület összes funkciója alkalmas többszavas kifejezéseket tartalmazó modellek kezelésére is), néhány olyan lekérdezésre is lehetőség van, ahol több szót együttesen vizsgálhatunk. Egyrészt az egyszavas lekérdezéshez hasonlóan több szót is megadhatunk a lekérdező mezőben, ekkor az egyes szavakhoz tartozó beágyazási vektorokat összeadjuk és az összegvektor-



2. ábra. Gyümölcsök és zöldségek

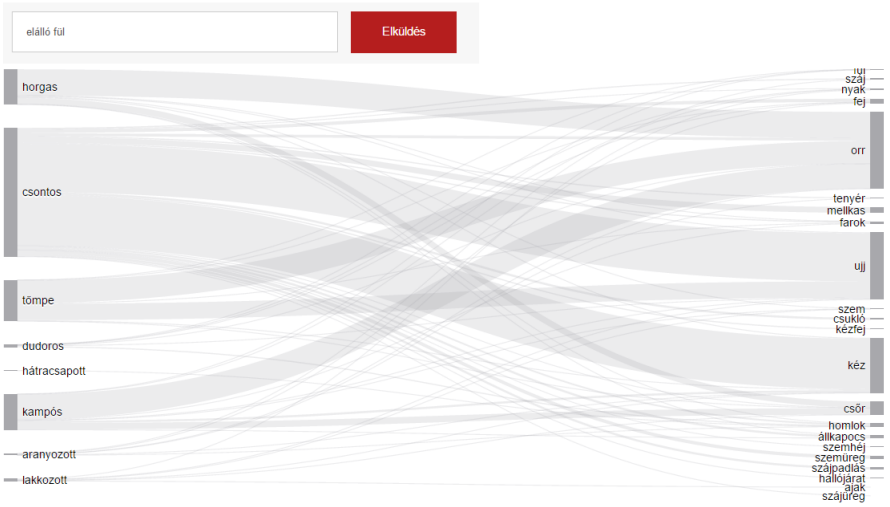
hoz legközelebbi szavakat jelenítjük meg az adott modellből.[4] A lekérdezésben algebrai műveletek is megadhatók (pl. a pozitív és negatív előjellel összegzendő szavak halmaza), így a felület a bevezetésben említett analógiák lekérdezésére is alkalmas. Megszorítást tehetünk arra is, hogy a rendszer a modellből csak egy adott szóhalmaz elemei közül a leghasonlóbbakat adja vissza.

Egy másik funkció használatakor pedig egy több szóból álló kifejezés lekérdezése során mindegyik szóhoz lekérdezzük az annak környékén található hasonló szavakat, majd az összes kombináció szerint párba rendezzük őket az eredeti szórend szerint és az így kapott bigramok korpuszbeli gyakoriságát reprezentáló ábrán jelenítjük meg az eredményt. A 3. ábrán látható *elálló fül* kifejezés esetén tehát az *elálló* szóalakhoz legközelebbi szavak a *horgas, csontos, tömpe, dudoros, hátracsapott, kampós, aranyozott, lakkozott*, tehát az ábra bal oszlopában látható szavak, a *fül* szóhoz hasonlók pedig a jobb oldalon szereplő *száj, nyak, fej, orr, tenyér, mellkas, farok, ujj, szem, csukló, kézfej, kéz, csőr, homlok, stb.*. A két oszlop közötti összeköttetések vastagsága jelöli az összekötött bigramok gyakoriságát. Így látható, hogy az eredeti *elálló fül* kifejezéshez hasonló bigramokat kaptunk, mint pl. *horgas orr, tömpe orr, csontos kéz, stb.*

5. A Dologfelismerő

Egy másik cikkünkben [6] olyan módszert mutatunk be (Dologfelismerő néven), amely egy szóbeágyazási modellhez az eredeti modell értelmezését segítő szemantikai kategóriacímkékkel összerendelt vektorokat ad hozzá. A hozzáadott címké-

[4] Mivel koszinusz-távolságot használunk metrikaként, a vektorösszeghez és az átlagvektorhoz képest számított távolság azonos, lévén ezek iránya azonos.

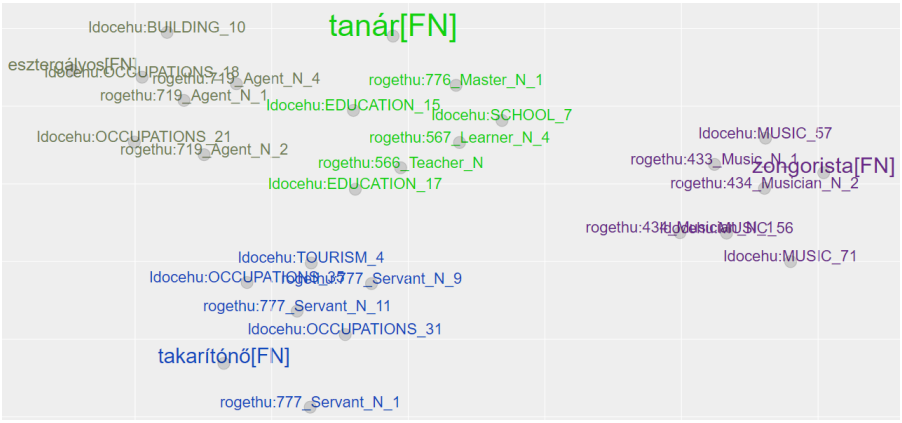


3. ábra. Az *elálló fülhöz* hasonló kifejezések

ket létező lexikai erőforrásokból, azok automatikus transzformációjával illesztjük az eredeti beágyazási térbe. Ennek köszönhetően az eredetileg nagyon sok szót tartalmazó szemantikai térben jóval kisebb számú referenciapontot helyezünk el, és a lekérdezésnél csak ezeket használjuk, 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 (itt angol volt) nem szükségszerűen azonos az eredeti szóbeágyazási modell (itt magyar) nyelvével. Ezzel a módszerrel jöttek létre a 4lang, ldocehu és rogethu modellek.

A webes felületbe is integráltuk ezt a funkciót. Egyrészt lehetőség van a különböző kategorizációs modellekből (4lang, rogethu, ldocehu) egy tetszőleges szóhoz az ahhoz rendelt kategóriacímkéket lekérdezni. Mivel a hozzárendelés során egy köztes lépésben egy angol modellt is felhasználunk, ezért a nyelvek közötti transzformációt is be tudjuk mutatni oly módon, hogy a lekérdezett szóhoz az angol modellben (wikien.pos) legközelebb álló szavakat jelenítjük meg. Az eredmények, akár az angol kapcsolódó szavak, akár a kategóriacímkék esetén, itt is egy-egy, a hasonlóság mértéke szerint rendezett listában jelennek meg.

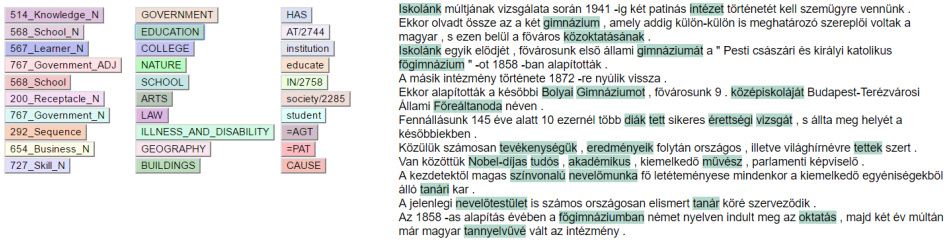
Emellett a kategóriacímkéket a fent bemutatott kétdimenziós ábrán is meg tudjuk jeleníteni. Tetszőleges számú szóhoz tetszőleges címkemodellekből tetszőleges számú címkét azonos szemantikai térbe való transzformáció után egyetlen ábrán jelenítünk meg. Ezáltal az eredeti beágyazási tér egyes területeit a Dologfelismerőben definiált kategóriacímkékkel vizuálisan is annotálni tudjuk. A 4. ábrán négy szóhoz (*zongorista, tanár, esztergályos, takarítónő*) két modellből 3-3 hozzárendelt címke és ezek elhelyezkedése látható a 2 dimenziós térbe leképezve.



4. ábra. Foglalkozások és címkéik

6. Szövegcímkéző

A felületre egy olyan funkciót is integráltunk, ami a Dologfelismerő címkéző funkcióját az ott használt modellekre támaszkodva hosszabb szövegekre valósítja meg (némileg hasonlóan a Wikifierhez[5]). A beadott szöveget elemezzük (lemmatizálás, PoS-taggelés) és a tartalmas szavakhoz (jellemzően főnevek, igék, melléknevek) lekérdezzük az 5 legközelebb álló kategóriacímkét mindhárom modellből, majd ezeket összesítjük a teljes szövegre nézve és gyakoriság szerinti sorrendben jelenítjük meg az eredeti szöveg mellett. Ez a funkció a szöveg tematikus besorolását, egyfajta szövegkivonatolást valósít meg. A megjelenített címkék ebben az esetben interaktív gombként is működnek, egy-egy címke kiválasztásakor a szövegben az olyan címkét kapott szavak kiemelődnek. A 5. ábrán egy ilyen módon felcímkézett szöveg látható.



5. ábra. A „Thingifájer”

[5] http://wikifier.org/

7. Konklúzió

A szóbeágyazás a szavak jelentésének ábrázolására hatékonyan használható reprezentációs módszer, azonban a létrejött modellek minőségének ellenőrzése nehéz feladat nem csak azért, mert a modellek közvetlen kiértékelésére alkalmazható kvantitatív módszerek nyelvenkénti adaptációja nehéz, hanem azért is, mert a szemantikai reprezentáció minőségének meghatározása szubjektív. A bemutatott webes felülettel célunk az volt, hogy a különböző módon létrehozott magyar nyelvű szóbeágyazási modelleket vizsgálni tudjuk, többféle módon jelenítve meg az általuk definiált szemantikai teret. Emellett a felület nagyon hatékonyan használható különböző szemantikai osztályozási, lexikográfiai feladatok elvégzésére. A szóbeágyazási modellek felhasználásával megvalósított komplexebb algoritmusaink egy része szintén elérhető a webes felületről.

Hivatkozások

1. 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)
2. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne (2008)
3. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
4. 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)
5. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. pp. 746–751 (2013)
6. Novák, A., Siklósi, B.: A dologfelismerő. XIII. Magyar Számítógépes Nyelvészeti Konferencia (2017)
7. 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)
8. Siklósi, B., Novák, A.: Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. XII. Magyar Számítógépes Nyelvészeti Konferencia pp. 3–14 (2016)
9. 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)
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