Bu-Bor-éK: grafikus címkenormalizáló eszköz

Novák Attila[1,2], Novák Borbála[1,2]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat A webes portálokon megjelenő tartalmakat gyakran tematikus címkékkel látják el, amelyeket jelenléte többek között hatékonyabb kereshetőséget, a webes keresőkben jobb találatokat eredményez, illetve a kapcsolódó vagy személyre szabott tartalmak megjelenítéséhez is használható. A kulcsszavakat gyakran manuálisan és nem egységesen rendelik a tartalmakhoz, ez gyakran a címkekészlet nemkívánatos elburjánzásához vezet. Cikkünkben egy olyan grafikus eszközt mutatunk be, amelyet az említett probléma kezelésére címkebeágyazási modellek kétdimenziós megjelenítéséből kiindulva többek között címkekészletek normalizálására és szerkesztésére lehet használni. A címkekészlet szerkesztésére szolgáló eszközben a vizuális modell bejárható, a címkék kereshetőek, szerkeszthetőek, címkeosztályokba sorolhatóak, a szinonim címkék összevonhatóak.
Kulcsszavak: információkinyerés, annotáció, lexikai erőforrások, kulcsszónormalizálás, grafikus eszköz

1. Bevezetés

Az utóbbi években lezajlott paradigmaváltás eredményeképpen mára nem túlzás azt állítani, hogy a nyelvtechnológiában előforduló szinte minden feladatra neurális hálózatok alkalmazásán alapuló megoldásokkal érhető el a legjobb eredmény. Míg kezdetben a szóbeágyazási modellek önmagukban is lenyűgöző szemantikai reprezentációt produkáltak, addig mára a világ élmezőnyébe tartozó kutatói egyre bonyolultabb architektúrákat alkalmaznak egy-egy feladat megoldására. Ezeknek az összetett hálózatoknak a belső működése sok esetben már teljesen értelmezhetetlen. Az azonban még mindig igaz, hogy a neurális modellekben a szavak, illetve egyre inkább a szavaknál kisebb lexikai egységek nem szimbolikus formában, hanem néhányszáz dimenziós vektorokként jelennek meg.

Az általában köztes reprezentációként, de akár végeredményként létrejövő vagy éppen egy hálózat bemeneteként szolgáló sokdimenziós vektorok értelmezése és azok minőségének ellenőrzése nehéz feladat. A szakirodalomban elterjedt közvetlen kiértékelési módszerek a hasonlósági listák, illetve analógiák vizsgálatával ellenőrzik a szóbeágyazások minőségét (l. pl. Faruqui és mtsai (2016), Schnabel és mtsai (2015)), vagy valamilyen a beágyazási modellt egy ráépülő feladat megoldásához használó komplexebb modell teljesítményének változásán

keresztül próbálják közvetetten jellemezni a beágyazási modellek minőségét. Egy másik megközelítés a sokdimenziós vektorok terét két-három dimenzióba képezi le, ami már könnyen vizualizálható, így ránézésre is áttekinthetővé teszi a modellben szereplő elemek reprezentációját, azok egymáshoz képesti elhelyezkedését a modell alkotta térben. Ez utóbbi módszer kvantitatív kiértékelésre kevésbé alkalmas, viszont nagyobb rálátást, jobban áttekinthető megjelenítést tesz lehetővé.

Ezek a kiértékelésre és elemzésre szolgáló módszerek azonban statikusak, a modellben létrejött reprezentációnak csupán a megjelenítésére szolgálnak. Ebben a cikkben egy olyan eszközt mutatunk be, amely a többszáz dimenziós beágyazások kétdimenziós leképezéséből kiindulva lehetővé teszi a megjelenített elemek mozgatását, szerkesztését és összevonását. Az eszköz hatékonyan alkalmazható többek között zajos címkekészletek kézi tisztítására beágyazásalapú címkemodellből kiindulva. A tisztított címkekészlettel a modell újrataníthatóvá, és pontosabb, egységesebb eredményt adó modell nyerhető.

2. Motiváció

A bemutatásra kerülő eszközt két motivációs példán mutatjuk be. Az első példában egy szövegcímkéző rendszer tanításakor használt címkekészlet normalizálása a feladat, a másodikban pedig egy a szavakhoz szemantikai osztályokat rendelő rendszer osztályrendszerének átalakítása.

2.1. Szövegcímkézés

A webes hírportálokon megjelenő szövegeket gyakran különböző tematikus címkékkel látják el, melyek lehetővé teszik a látogatók számára, hogy kifejezetten valamilyen számukra érdekes témával, személlyel, eszközzel stb. kapcsolatos cikkeket vagy egyéb tartalmakat jelenítsék meg. Másrészt a kulcsszavakat az adott cikkhez kapcsolódó egyéb cikkek vagy tartalmak megjelenítéséhez is használják, illetve szerepet játszanak a címkék a keresőmotorok (pl. a Google) találatrangsorolási algoritmusaiban is.

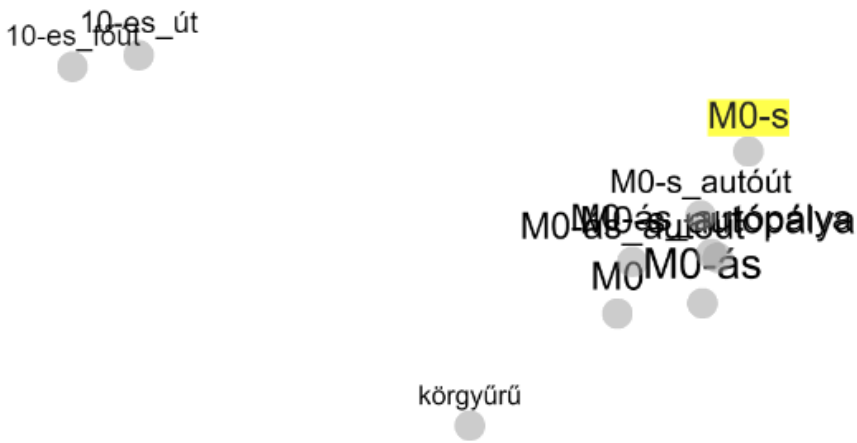
Egy szöveghez az annak tartalmához kapcsolódó tematikus kulcsszavak automatikus hozzárendelésére számos algoritmikus megoldás létezik. Egy ilyen kifejezetten magyar nyelvű sajtószövegek címkézésére szolgáló eszközt mutat be például Farkas (2009). Ennek ellenére sok online is megjelenő szövegarchívumban a cikkekhez a szerző/szerkesztő által egyedileg kézzel hozzárendelt kulcsszavak szerepelnek (pl. Farkas (2009) sem említi, hogy az ott bemutatott algoritmust újonnan születő cikkek címkézésére (vagy annak segítésére használták volna). A kézi címkézést néha erre szakosodott (általában könyvtáros végzettségű) szakember végzi, azonban sokszor inkább maguk a szerzők végzik el ezt a feladatot is.

Ebből kifolyólag az egy archívumon belül használt címkekészlet gyakran nem egységes, a szerzők ugyanannak a címkének különböző (gyakran elírt) formáit használhatják: *M0-ás autópálya, M0-ás, M0-s autópálya, M0-s, M0-ás autóút,*

M0, M0-s autóút. Bár a kézzel címkézett szövegek jól használhatóak egy automatikus címkéző rendszer tanításához, a kulcsszavak változatossága miatt a rendszer mért pontossága alacsonyabb lesz az elvártnál.

Egy folyamatban levő projekt keretében sajtószövegek automatikus címkézésére vállalkoztunk, amelyre a fastText programcsomag (Joulin és mtsai, 2017) címkézőalgoritmusát használjuk. Az osztályozóhálózat(ok) bemenetén az adott szöveg tokenjeinek, illetve token-n-eseinek reprezentációja jelenik meg (a bennük szereplő különböző hosszú karakter-n-gramok reprezentációjának átlagaként), és az osztályozó ehhez a szövegreprezentációhoz és az egyes lehetséges címkékhez rendel illeszkedési értéket multinomiális logisztikus regresszió alkalmazásával. Megfelelő küszöbérték választása mellett az adott szövegre jól illeszkedő kulcsszavak elválaszthatóak a kevésbé jól illeszkedőktől. Bár megjelenése óta a fastText modellnél jobban teljesítő szövegosztályozó modellek is készültek (a cikk írásának idején az ilyen jellegű feladatokban a mélyneurális XLNet architektúra nyújtja a legjobb teljesítményt több angol nyelvű adatbázison (Yang és mtsai, 2019)), ezeknek komplexitása, hardver- és futásiidő-igénye a pontosságbeli teljesítménykülönbséget jóval meghaladó mértékben nagyobb, mint a fastTexté.

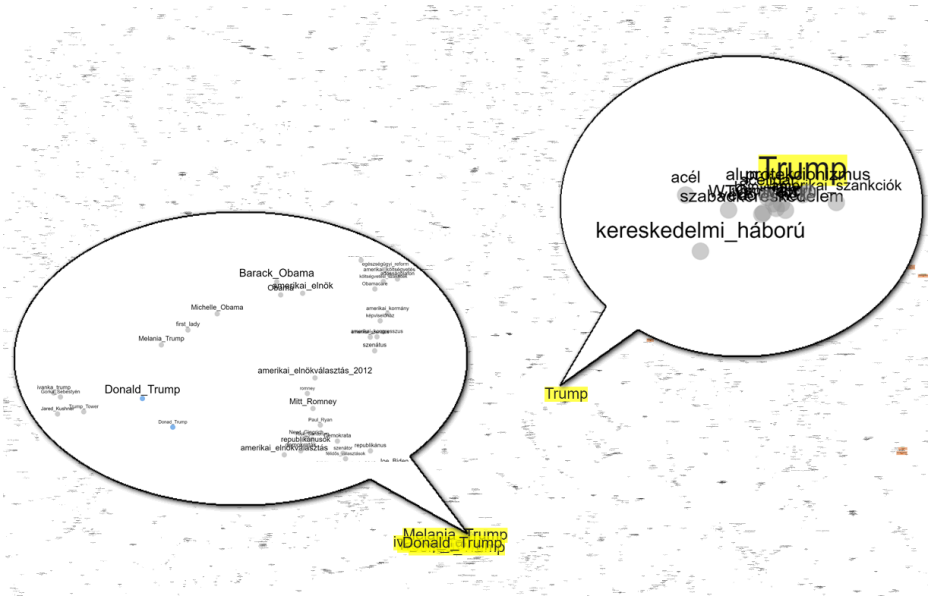
A betanított modell címketerében megfigyelhető, hogy egy címke különböző írásváltozatainak a reprezentációja a beágyazási térben egymáshoz közel helyezkedik el, mert hasonló témájú cikkeket címkéznek ugyanannak a kulcsszónak a különböző változataival (1. ábra).



1. ábra: Az M0-s címke írásváltozatainak elhelyezkedése a címketérben

Időnként megfigyelhetőek eltérések ettől az alapvető mintázattól, de ennek mindig a szövegkorpuszra, a címkehasználat egyedi sajátosságaira, illetve a címkék többértelműségére visszavezethető magyarázata van. Modellünk például egy-

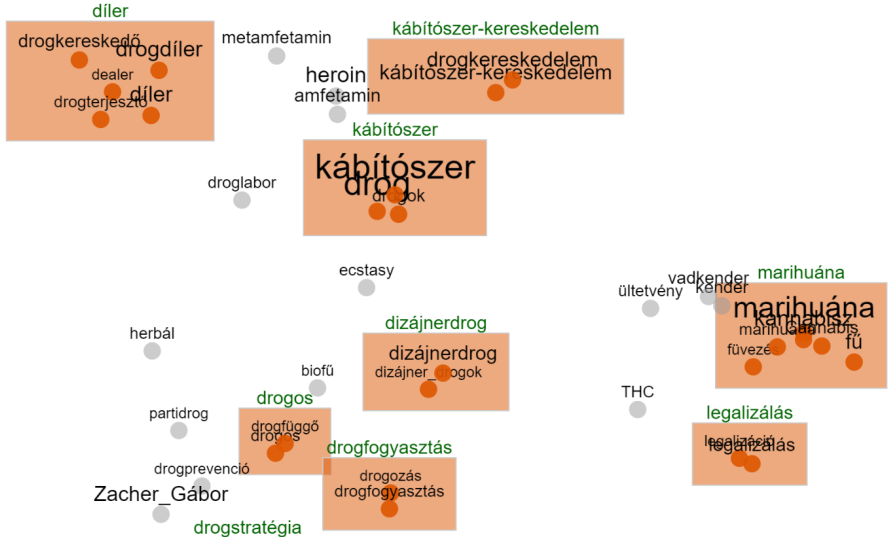
értelműen megragadta azt a sajátosságot, hogy az adott korpuszban a gazdasági témájú cikkek szerzői az amerikai elnököt következetesen *Trump*-nak címkézik, míg a politikai cikkek szerzői a keresztnevét is használva (azt időnként elírva) *Donald (Donad, Donal) Trump*-nak. Így a keresztnév nélküli Trump címke a keresztnevesektől viszonylag távol a kereskedelmi háborúval kapcsolatos címkék között szerepel (2. ábra). Többértelműségből fakadóan került pl. a *magyar csapat* címke viszonylag távol a sporttól (annak ellenére, hogy olimpiai témájú cikkek is viselik ezt a címkét) és közel Németh Szilárd rezsibiztoshoz a 2014 eleji politikai jellegű „Magyar Csapat"-kezdeményezésről szóló cikkek hatására. Számos esetben a szinonim címkék nemcsak különböző írásváltozatokat, hanem lényegében ugyanannak a fogalomnak különböző eredetű/stílusú megnevezéseit ölelik fel, pl. *fű, marihuána, kannabisz* stb. (3. ábra).



2. ábra: Trump címkéi egymástól távol

2.2. Szemantikai osztályozás

A másik feladatban a Dologfelismerő (Novák és Siklósi, 2017) által használt címkerendszer normalizálása volt a cél. A Dologfelismerő létező szemantikai erőforrásokból (Roget's Thesaurus (Chapman, 1977), Longman (Summers, 2005), 4lang (Kornai és mtsai, 2015)) szóbeágyazások segítségével készített modell alapján rendel szemantikai kategóriákat, illetve tulajdonságokat bármilyen (magyar vagy angol) szóhoz. A Dologfelismerő létrehozásakor is ismert probléma volt a



3. ábra: fű, marihuána, kannabisz

felhasznált erőforrások címkekészletének régiessége (Roget's Thesaurus), illetve a különböző erőforrások címkekészletének összehangolása. Bár a Dologfelismerő sokszor jól meg tudja ragadni egy szó szemantikai tulajdonságait és kategóriáit, a megjelenített címkékből ez nem mindig látszik. Az értelmetlennek tűnő címke reprezentációjának létrehozásához használt szavak csoportját vizsgálva sok esetben azonban kiderül, hogy csupán az eredeti erőforrásokból átvett megnevezés a téves. Erre korábban az eredeti címkék klaszterezéssel történő felbontása volt a megoldás. 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. Ez a megoldás azonban csupán egy technikai megoldás volt, a címkék elnevezése továbbra sem feltétlenül tükrözte az általuk reprezentált tartalmat.

Célunk volt a Dologfelismerő címkekészletének normalizálása, de mivel a modell címkekészletének mérete miatt a címkék kézzel való átnézése nem könnyű feladat, ráadásul sok esetben nemcsak a címkék átnevezése, hanem több címke összevonása is szükségesnek tűnt.

3. A címkekészletek normalizálására szolgáló eszköz

A tipikusan néhány száz dimenziós címkebeágyazási modellben[1] megjelenő címkereprezentációkat a vizualizációhoz és a grafikus szerkesztéshez először két dimenzióba vetítjük. Bár a tavalyi MSZNY-en bemutatott beszédfelismerők vizualizációjával kapcsolatos eredményeken (Grósz és Tóth, 2019), és különösen az egyik szerzővel folytatott későbbi beszélgetésen felbuzdulva kísérleteztünk autoencoder alapú vizualizációval is, az eredmények a mi esetünkben nem bizonyultak használhatónak, így a lokális kapcsolatokat jobban megőrző klasszikus t-SNE vizualizációs algoritmus (van der Maaten és Hinton, 2008) alkalmazása mellett maradtunk.

A javascript-alapú cytoscape.js gráfvizualizációs és -szerkesztő csomag (Franz és mtsai, 2015) felhasználásával készítettük el a címketér elemeinek szerkesztésére, illetve az azonos szerepű címkék összevonására szolgáló testre szabott böngészőalapú szerkesztőeszközünket.

A címketér t-SNE algoritmussal kapott 2 dimenziós képét[2] a Cytoscape-pel kompatibilis json formátumba konvertáljuk, és ehhez hozzáadjuk a gyakorisági adatokat (illetve a megjelenítéshez a csomópontoknak a gyakoriság logaritmusával arányos méretét). A megjelenített címketérkép egérrel/trackpaddel navigálható, zoomolható, az egyes címkék mozgathatóak.

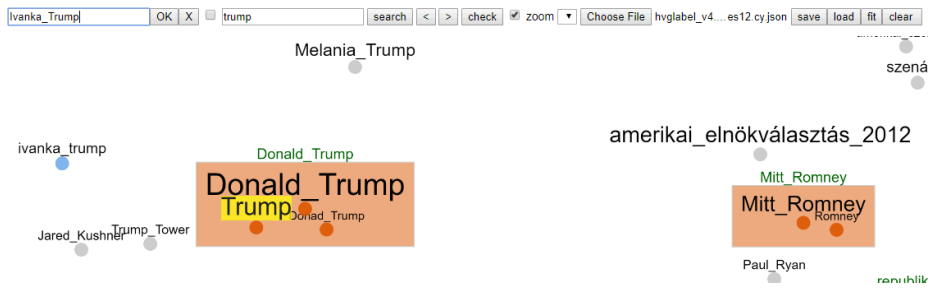
A címkéket egymás közelébe mozgatva vagy az ekvivalens címkék kijelölése után a megfelelő billentyű-egérgombkombináció megnyomásával azok szinonimacsoportba csoportosíthatóak. A csoportot reprezentáló narancsszínű téglalapként megjelenő szülőcsomópont eredő címkéje automatikusan a csoportban szereplő leggyakoribb címke lesz (4. ábra: *Donald Trump, Mitt Romney*), de más címke is kiválasztható, illetve a címkék szerkeszthetőek. Szerkesztéskor, illetve összevonáskor mindig megfelelően nyomon követjük az eredeti címkéket is, hiszen az eszköz célja éppen az, hogy az eredeti túlságosan változatos, illetve hibás címkéket az adatbázisban javítani, illetve egységesíteni tudjuk.

A megjelenítésre és szerkesztésre szolgáló felület felett helyeztük el a címkék szerkesztésére, a keresésre, a modell betöltésére és elmentésére és különböző statisztikai információk megjelenítésére szolgáló vezérlőelemeket (4. ábra fölül). Lehetőség van a szerkesztendő/javítandó címke egy billentyűlenyomásra történő automatikus kis/nagybetűsítésére is. Erre viszonylag gyakran van szükség a hibásan csupa kisbetűvel írt nevek miatt (4. ábra: itt éppen az *ivanka trump* címke nagybetűsítése történik a felül baloldalt látható címkeszerkesztő mezőben).

Az eszköz lehetőséget ad arra, hogy címkékre és címkerészletekre keressünk. Ilyenkor az illeszkedő címkéket (sárgával) kiemelve és kinagyítva jeleníti meg az eszköz (2. ábra), illetve lehetőség van csak az illeszkedő címkéket magába foglaló területre való automatikus ráközelítésre is. A kinagyított/kiemelt címkéket

[1] A cikkben említett tematikuscímke-beágyazási modell dimenziószáma 100, a Dologfelismerő modellé 300.

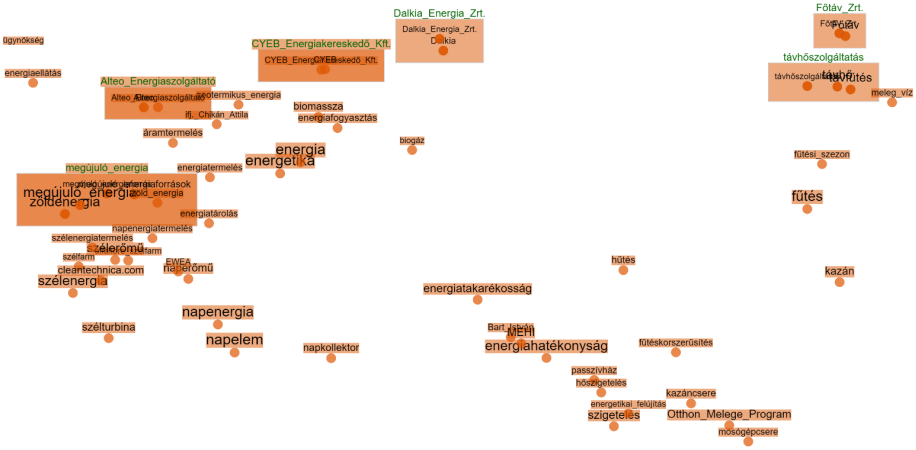
[2] A t-SNE perplexitásparamétereként 50-es értéket használtunk. Ha a megjelenítendő elemszám kisebb, mint a perplexitásparaméter háromszorosa, akkor a preplexitásértéket a megjelenítendő elemek harmadára állítjuk be.



4. ábra: A grafikus szerkesztő vezérlőelemei – Címke automatikus nagybetűsítése szerkesztéskor (*ivanka trump*)– Szinonimacsoportba rendezett címkék az eredő címkéjükkel (*Donald Trump, Mitt Romney*) – Trump címkéi itt már együtt

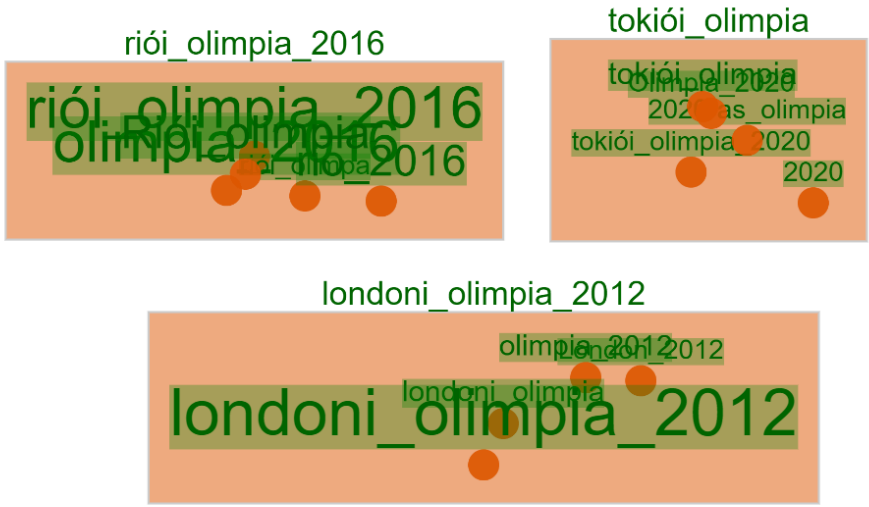
abban az esetben is könnyen a megfelelő helyre lehet mozgatni, ha azok az eredeti térben esetleg a velük ekvivalens címkéktől messzebb találhatóak (mint a korábban említett *Trump* címke esetében). Lehetőség van az illeszkedő címkéken való végiglépkedésre is mindig automatikusan az épp soron következő címke környezetére ráközelítve (l. 4. ábra: *Trump* sárgával kiemelve).

A már átnézett nem összevonandó címkéket megjelölhetjük ellenőrzöttként. Ezek címkéje az összevont kulcsszavakat reprezentáló dobozokhoz hasonló narancsszínű hátteret kap (5. ábra). Így adminisztrálhatjuk a munka előrehaladását.



5. ábra: Az átnézett, kijavított címkéket készként megjelölhetjük.

Az eszköz lehetővé teszi speciális címkeosztályok kezelését is. Pl. a tematikus címkézésnél megkülönböztethetünk olyan címkéket, amelyek egy-egy időben jól körülhatárolt eseményt jelölnek (pl. egy konkrét sportverseny, fesztivál, kiállítás, díjátadó, baleset vagy választás). Ezek reprezentációja nagyon hasonlít bármelyik másik hasonló eseményéhez (pl. a 2018-as Oscar-gála legjobban a többi Oscar-gálához hasonlít), azonban ezek hosszú távon a címkézőrendszer szempontjából valószínűleg nem hasznos címkék. A szerkesztő lehetővé teszi az ilyen címkék megjelölését, és a hosszú távon őket helyettesítő általános címkékhez kapcsolását.



6. ábra: Az egyedi eseményeket jelölő címkék megjelölése.

Lehetőség van a címketérkép aktuális állapotával kapcsolatos statisztikai adatok megtekintésére is (feldolgozottnak jelölt, még feldolgozásra váró, átnevezett, speciálisnak jelölt (pl. egyedi eseményeket jelölő), illetve az összevont, valamint az összevontak fölé rendelt (szülő-) címkék száma).

A Dologfelismerő címkéinek szerkesztéséhez kiegészítettük az eszközt egy olyan funkcióval, ami lehetővé teszi a címkéhez kapcsolódó példaszavak megjelenítését (egyszerűen a címke kiválasztásával), ami alapján egyrészt a címke által jelölt halmaz szemantikai koherenciája felmérhető (és a címke megjelölhető, ha a szóhalmaz nem koherens), másrészt a címke a halmazt ténylegesen fedő fogalomra átnevezhető. Illetve ugyanez a funkció segíti a címkék összevonását is. A címketér alapján való megjelenítés és a konkrét példaszavak címkék alá rendelése azt is lehetővé teszi, hogy – ellentétben magával a szóbeágyazási térrel, ahol a

többértelmű szavak nem jelennek meg több példányban – a konkrét példaszavak több különböző címke alatt megjelenhetnek, akár különböző értelemben.

Az eszköz használatával olyan a természetes nyelvhasználatból adódó szemantikai csoportok is feltárultak, amelyek egyébként nem merültek volna fel bennünk: pl. heraldikai elemek, szabász-varrászati eljárások, stb. Emellett az egyébként hasznos és az eredeti címkekészletből mindenképp megtartandónak látszó címkéken/csoportokon belül (pl. betegségek) tapasztaltuk olyan jellegű alcsoportok megjelenését, amelyek inkább egy mindennapi 'józan ész' jellegű ontológiára utalnak (pl. a betegségek szétválása veszélyes–nem veszélyes betegségekre). Ezeket az átnevezett címkékben tükröztettük.

4. Összefoglalás

Cikkünkben egy címkekészletek normalizálására és szerkesztésére szolgáló grafikus eszközt mutattunk be. Címkebeágyazási modellből a t-SNE algoritmussal nyert 2D vizualizációból kiindulva lehet egyszerű műveletekkel összevonni a szinonim címkéket, átnevezni a nem megfelelő nevet viselőket, és ily módon jobb minőségű tanítóanyagot hozni létre a nyelvi modellek építéséhez.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 és a PD 125216 számú projekt keretében az FK 17 és a PD 17 pályázati program, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

Chapman, R.: Roget's International Thesaurus. Harper Colophon Books, Crowell (1977), https://books.google.hu/books?id=9VhQAAAAMAAJ

Farkas, R.: Az origo automatikus címkézési projekt tapasztalatai. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2009). pp. 84–92. Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2009)

Faruqui, M., Tsvetkov, Y., Rastogi, P., Dyer, C.: Problems with evaluation of word embeddings using word similarity tasks. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 30–35. Association for Computational Linguistics, Berlin, Germany (Aug 2016), https://www.aclweb.org/anthology/W16-2506

Franz, M., Lopes, C.T., Huck, G., Dong, Y., Sümer, S.O., Bader, G.D.: Cytoscape.js: a graph theory library for visualisation and analysis. In: Bioinformatics (2015)

Grósz, T., Tóth, L.: Mély neuronhálós beszédfelismerők működésének értelmező elemzése. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 287–298. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2019)

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (2017), https://www.aclweb.org/anthology/E17-2068

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)

van der Maaten, L., Hinton, G.E.: Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)

Novák, A., Siklósi, B.: A Dologfelismerő. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017). pp. 25–36. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2017)

Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 298–307. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015), https://www.aclweb.org/anthology/D15-1036

Summers, D.: Longman Dictionary of Contemporary English. Longman Dictionary of Contemporary English Series, Longman (2005), https://books.google.hu/books?id=4zktAAAACAAJ

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. CoRR abs/1906.08237 (2019), http://arxiv.org/abs/1906.08237