Magyar nyelvű szó- és karakterszintű szóbeágyazások

Szántó Zsolt[1], Vincze Veronika[2], Farkas Richárd[1]

[1] Szegedi Tudományegyetem, Informatikai Intézet
Szeged Árpád tér 2.
e-mail:{rfarkas,szantozs}@inf.u-szeged.hu
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat A szóbeágyazási modellek az egyes szavak párszáz dimenziós folytonos térbe való leképezését adják meg úgy, hogy az egymáshoz hasonló szavak közel kerülnek egymáshoz a beágyazási térben. A szóbeágyazások széles körben használatossá váltak az elmúlt években. Jelen cikkben bemutatunk publikusan elérhető magyar nyelvű szóvektorokat, amelyeket 4,3 milliárd szövegszónyi korpuszból építettünk. Az első modellek (word2vec) a szavakat mint alapegységet dolgozták fel. Az utóbbi években több olyan kiterjesztése is született ezen modelleknek, amelyek karakterszintű információkat is ki tudnak aknázni. Ezek a modellek morfológiailag gazdag nyelveken előnyösebbek lehetnek, mint a pusztán szószintű modellek. A cikkben összehasonlítunk ugyanazon adatbázisból épített szó- és karakterszintű szóbeágyazásokat téma- és véleményosztályozási feladatokon kiértékelve.
Kulcsszavak: Szóbeágyazás, karakterszintű szómodell, osztályozás

1. Bevezetés

A szóbeágyazások alkalmazása az utóbbi években nagyban javította egyes természetesnyelv-feldolgozási alkalmazások hatékonyságát. A szóbeágyazások az egyes szavakat egy jellemzően párszáz dimenziós folytonos térbe képezik le, ahol a hasonló jelentésű szavak egymáshoz közel helyezkednek el. A szóbeágyazások nagy ereje abban rejlik, hogy míg az egyes célfeladatokhoz használható annotált adatbázisok mérete általában erősen korlátozott, addig a szóbeágyazások tanítására óriási méretű annotálatlan szövegeket használhatunk. Ennek következtében pedig a célfeladatunk adatbázisában ismeretlen vagy ritkán látott szóalakokat is képesek vagyunk kezelni. A szóbeágyazók tanítása a szavak közvetlen kontextusára épít, azaz hasonló kontextusban előforduló szavak fognak egymáshoz közel elhelyezkedni. A szavakhoz tartozó kontextus alapján kapott vektortérben mind jelentésbeli, mind morfológiai jellemzők is megjelenhetnek.

Jelen cikkben bemutatunk publikusan elérhető magyar nyelvű szóvektorokat, amelyeket 4,3 milliárd szövegszónyi korpuszból építettünk. Az első szóbeágyazási

modellek a szavakat mint alapegységet dolgozták fel. Az utóbbi években több olyan kiterjesztése is született ezen modelleknek, amelyek karakterszintű információkat is ki tudnak aknázni. Ezek a modellek morfológiailag gazdag nyelveken előnyösebbek lehetnek, mint a pusztán szószintű modellek. A cikkben összehasonlítunk ugyanazon adatbázisból épített szó- és karakterszintű szóbeágyazásokat téma- és véleményosztályozási feladatokon kiértékelve.

2. Szóbeágyazási modellek

A szóbeágyazások egy nyelv szavait egy párszáz dimenziós folytonos térben reprezentálják úgy, hogy a hasonló jelentésű szavak egymáshoz közel helyezkednek el térben. Az első modellek az egyes szavakat mint egységeket kezelték. Ebben a megközelítésben ugyanannak a szótőnek két ragozott alakja pontosan úgy különbözik egymástól, mint egy másik szótő (például a *macska* – *kutya* szópár tagjai pontosan ugyanúgy különböznek egymástól, mint a *macska* – *macskát* szópár esetében, azaz ezeket két különálló egységként kezelték a korai modellek). Az elmúlt években több megoldási javaslat is született arra, hogy ezt a problémát orvosolják. A megoldás minden esetben az, hogy karakterszintű információkra támaszkodva építjük fel a szóbeágyazást. Ezek alkalmazása különösen hasznos lehet a morfológiailag gazdag nyelvek esetén, ahol a ragozás miatt a korábban nem látott szavak aránya magasabb az átlagosnál. A ritka szavak problémájára egy másik lehetséges megoldás a szóalakok lemmatizálása [1], aminek mellékhatása, hogy a szóbeágyazás ismeretlen szövegen történő alkalmazásának előfeltétele lesz a szóbeágyazás tanításához használt lemmatizáló lefuttatása, ezzel szemben a karakteralapú módszerek a szóalak ismeretében képesek meghatározni egy új szó helyét a vektortérben. Ennek következtében a karaktersorozatok vizsgálata lehetőséget ad az elgépelésből fakadó hibák kezelésére is, ami különösen fontos lehet például a közösségi médiából származó szövegek esetén.

2.1. Szószintű modellek

Szóbeágyazások tanítására a két legáltalánosabban használt metódus a CBOW és a skip-gram [2]. Mindkét módszer az egyes szavak környezetét veszi alapul. Míg a CBOW esetén a gépi tanulási feladat a környezet alapján a keresett szó predikálása, addig a skip-gram esetén a kiválasztott szó alapján annak környezetére következtetünk.

Cikkünkben a skip-gram modellt követjük, ezt mutatjuk be röviden. Ebben a modelben két mátrixot tanulunk, a szavak és a környezet beágyazását. Az egyes mondatokban minden szóra legyűjtjük annak környezetében előforduló szavakat és a tanulás célfüggvénye a környezetben előforduló szavak megfigyelésének valószínűsége a középső szó feltevése mellett. A valószínűségek egy log lineáris softmax modellel becsülhetőek. A tanítás után a kontextusmátrixot eldobjuk és a szómátrixot használhatjuk szóbeágyazásnak. Az alacsony számítási igénynek köszönhetően a skip-gram modell hatékonyan alkalmazható nagy adatbázisokra is, milliárd szövegszónyi adatbázis feldolgozható egy nap alatt.

2.2. Karakterszintű modellek

A szavak és azok környezetének vizsgálatánál részletesebb reprezentációt kapunk, amennyiben a szavak vizsgálata mellett a szavakban szereplő karaktersorozatokat is figyelembe vesszük. Jelen cikkben a Facebook kutatói által publikált [3] FastTextet alkalmazzuk, ahol a szavak vektorát kiegészítjük a bennük szereplő karakter 3 és 4 gramokkal. Ezen vektorok felett a skip-gram módszerrel tanítunk szóbeágyazásokat. Ennek köszönhetően ha több karaktersorozatot oszt meg két szó, mint például morfémák vagy elgépelt szavak esetén, akkor azoknak is közel kell lenniük a szavak vektorterében.

3. Publikusan elérhető szóbeágyazások magyarra

A szóbeágyazások készítésénél cél volt, hogy minél nagyobb méretű és változatosabb stílusú és forrású szövegeket használjunk fel. Ehhez jó alapot nyújtott a Magyar Nemzeti Szövegtár 2. változata [4,5]. A MNSz2-ben újsághírek, könyvek, Wikipedia és egyéb szerkesztett szövegek mellett találhatók beszélt és közösségi médiából származó anyagok is. Az MNSz2 mellett a Hunglish [6] magyar-angol párhuzamos korpusz magyar nyelvű szövegeit használtuk fel az ismert magyar nyelvű erőforrások közül. Ezt egészítettük ki az origo.hu-ról és az index.hu-ról származó elektronikus újságcikkekkel. Fontos volt, hogy a korpuszban ne csak szerkesztett szövegek, hanem felhasználók által írt, alacsonyabb minőségű, közösségi médiából származó szövegek is megtalálhatók legyenek, hiszen számos alkalmazás dolgozik ilyen szövegeken és igényli az ilyen szövegeken számolt szóbeágyazásokat. Ehhez a gyakorikerdesek.hu oldaról gyűjtöttünk 1,9 milliárd szónyi kérdést és választ. A korpusz ezeken felül tartalmaz az OpenSubtitle [7] oldalról származó 466 millió token terjedelmű magyar rajongói feliratot, ami beszélt diszkurzusok elemzéséhez jelenthet nagy segítséget. Az egyes szövegek szavakra bontására a magyarlanc [8] beépített tokenizálóját alkalmaztuk. Az 1. táblázat tartalmazza az egyes részkorpuszokban szereplő tokenek számát, beleértve az írásjeleket is.

1. táblázat. Felhasznált részkorpuszok méretei.

Korpusz	Tokenek száma (milliárd)
MNSz2	1,53
Hunglish	0,03
Origo	0,15
Index	0,25
gyakorikerdesek.hu	1,87
OpenSubtitles	0,46
Összesen	4,29

Ezen a korpuszon mind szószintű skip-gram, mind karakterszintű szóbeágyazási modelleket tanítottunk. A szóvektorok publikusan és ingyenesen elérhetőek[3] a rgai.inf.u-szeged.hu/w2v oldalon. Legjobb tudomásunk szerint cikkünket megelőzően csak a Polyglot[4] biztosított publikusan elérhető magyar szóbeágyazást, de annak minősége jóval gyengébb, mint az általunk közzétett beágyazásoké.

3.1. Szóbeágyazások kiértékelése

Az elkészült szóbeágyazások kiértékeléséhez két adatbázist használtunk. Véleményosztályozásra az arukereso.hu oldalról letöltött termékértékeléseket használtunk. Az egyes termékekhez megadható előnyöket és hátrányokat alkalmaztuk pozitív és negatív tanító példaként.

Témaosztályozásra videojáték és sport témájú facebook bejegyzésekből készítettünk adatbázist. Ez szolgálhat közösségi médiából származó szövegek témabesorolásának egy megvalósíthatósági tanulmányaként. Forrásnak az alábbi videojátékokkal foglalkozó Facebook-oldalak publikus posztjait használtuk fel: PC Guru, GameStar (Hungary), 576 KByte és a Gameday Iroda, sport témakörben pedig a Nemzeti Sport Online és a FociHíradó oldalakról gyűjtöttünk publikus bejegyzéseket.

Mindkét esetben tehát bináris dokumentumosztályozási problémát fogalmaztunk meg. A témaosztályozásra használt adatbázis 10000 tanító és 2000 kiértékelő példát tartalmaz, a véleménydetekciós adatbázis 5000 tanító és 1000 kiértékelő dokumentumból áll. Mindkét feladatra a tanító és a kiértékelő adatbázison is az egyes címkék 50-50%-ban fordulnak elő.

4. Eredmények

A két szóbeágyazás kiértékelésére az egyes adatbázisokon a FastText rendszer neuronhálókon [9] alapuló dokumentumosztályozóját alkalmaztuk. Az osztályozó legnagyobb előnye, hogy sebességben vetekszik a hagyományos lineáris modelleket alkalmazó algoritmusok sebességével, viszont hatékonyan képes kihasználni a szóbeágyazásokban rejlő lehetőségeket. Az eredményeket a 2. táblázat tartalmazza.

2. táblázat. Szóbeágyazások pontossága téma- és véleményosztályozásra.

	témaosztályozás	véleménydetekció
FastText	90,1	91,3
FastText + Szószintű skip-gram	89,6	90,7
FastText + Karakterszintű skip-gram	93,5	91,7

[3] Maguk a részkorpuszok nyers szövegei semmilyen formában sem érhetőek el és a szóvektorokból nem lehetséges azok visszafejtése sem.

[4] https://sites.google.com/site/rmyeid/projects/polyglot

A FastText dokumentumosztályozója alapértelmezésként a tanító halmazból tanulja meg az egyes szavak szóbeágyazását, de képes korábban – nagyobb adathalmazon tanított – szóbeágyazások alkalmazására is. A szószintű modellre építő nagymennyiségű adaton tanított szóbeágyazásokkal nem sikerült jobb eredményt elérni a külső erőforrást nem használó modellhez képest. Ezzel a karakterszintű modell alkalmazásával a témaosztályozás esetén 3,5, míg véleménydetekció esetén csekélyebb, 0,4 százalékpontos javulást sikerült elérni. Ennek az lehet az indoka, hogy a karekterszintű modell segítségével lehetőségünk volt a nagymennyiségű szövegben nem található szavak reprezentációjának a becslésére is.

5. Összegzés

Ebben a cikkben ismertettünk magyar nyelvű szóbeágyazási modelleket, amelyeket 4,3 milliárd szövegszónyi korpuszból építettünk. Ezek a modellek szó- és karakterszinten is működnek, ami morfológiailag gazdag nyelveken különösen hasznosnak bizonyul az egy szóhoz tartozó lehetséges szóalakok nagy száma miatt. A létrehozott szóbeágyazásokat téma- és véleményosztályozási feladatokon értékeltük ki. A továbbiakban tervezzük a szóbeágyazások más alkalmazásokban való felhasználását is.

A létrehozott szóvektorok szabadon elérhetők a rgai.inf.u-szeged.hu/w2v oldalon.

Köszönetnyilvánítás

Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Hivatkozások

1. Siklósi, B., Novák, A.: Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. In: XII. Magyar Számítógépes Nyelvészeti Konferencia. (2016)
2. 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. Curran Associates, Inc. (2013) 3111–3119
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. Oravecz, Cs., Sass, B., Váradi, T.: Mennyiségből minőséget. Nyelvtechnológiai kihívások és tanulságok az MNSz új változatának elkészítésében. In: XI. Magyar Számítógépes Nyelvészeti Konferencia. (2015)
5. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: LREC. (2014)
6. Varga, D., Halácsy, P., Kornai, A., Nagy, V., Nagy, L., Németh, L., Trón, V.: Parallel corpora for medium density languages. In: Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05. (2007) 247–258
7. Tiedemann, J.: Finding Alternative Translations in a Large Corpus of Movie Subtitles. In: LREC. (2016)

8. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771
9. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)