CBOW/A: módosított CBOW algoritmus annotált szövegekből készített vektortérmodellek létrehozására

Novák Attila, Laki László János, Novák Borbála

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

Kivonat Cikkünkben a szóbeágyazási modellek készítésére alkalmas fastText könyvtár CBOW algoritmusának egy olyan módosított változatát mutatjuk be, amellyel a felszíni szóalakok és az azokhoz tartozó annotációk reprezentációját egyszerre tartalmazó vektortérmodell hozható létre. Bemutatunk egy konkrét modellt is, amelyet morfológiai és szintaktikai függőségi annotációt tartalmazó angol nyelvű korpuszon tanítottunk be, és amely alkalmas olyan lekérdezések hatékony megválaszolására, mint hogy *mit eszünk, mit csinálunk egy csontvázzal, mit csinálunk még azzal, amit eszünk,* stb.

1. Bevezetés

A szakirodalomból ismert, hogy számos alkalmazásban hasznos lehet grammatikai annotációt tartalmazó korpuszból épített szóbeágyazási modelleket használni, mert ezek bizonyos feladatokban jobban teljesítenek, mint az annotálatlan felszíni szóalakokból épített beágyazási modellek [1,2]. Ugyanakkor a legtöbb gyakorlati nyelvtechnológiai feladatban szükség van a felszíni szóalakok vektor reprezentációjára. Ebben a cikkben egy olyan vektortérmodellt mutatunk be, amely egyszerre tartalmazza a felszíni szóalakok, a lemmák és a szavak közötti grammatikai viszonyok reprezentációját, amelyben tehát triviális módon értelmezhető az ilyen különböző típusú objektumok közötti távolság, és így használható olyan jellegű kérdések megválaszolására, hogy tipikusan milyen elemek állnak egymással adott típusú kapcsolatban. Például, az *eat* ige tárgyaként ételek listáját várjuk eredményül. A cikkben bemutatott modellt egy angol nyelvű korpuszból hoztuk létre, de az alkalmazott módszer nyelvfüggetlen.

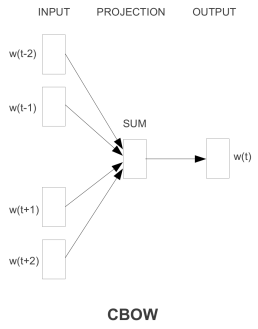
2. Folytonos disztribúciós szemantikai modellek

A disztribúciós szemantika lényege, hogy a szavak jelentése szorosan összefügg azzal, hogy milyen kontextusban használjuk őket. A hagyományos disztribúciós

szemantikai modellek létrehozásakor az egyes szavak előre meghatározott méretű környezetét az azokban előforduló szavak nagy korpuszból számított előfordulási statisztikái alapján határozzuk meg.

Ezzel szemben a nyelvtechnológiai kutatások egyik kurrens módszere a folytonos vektoros reprezentációk alkalmazása (*word embedding*), melyek nyers szöveges korpuszból szemantikai információk kinyerésére alkalmazhatók. Ebben a rendszerben a lexikai elemek egy valós vektortér egyes pontjai, 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, 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 [3,4].

Ennek a modellnek a tanítása során is az egyes szavak fix méretű környezetét vesszük figyelembe, az ezekből álló vektor azonban egy neurális hálózat bemenete. A környezetet reprezentáló vektorok együttesét használja a hálózat arra, hogy megjósolja a célszót. A tanítás során a hiba visszaterjesztésével és ennek megfelelően a környezetet reprezentáló vektorok frissítésével jön létre 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. Az erre a feladatra felépített neurális hálózat a CBOW (*continous bag-of-words*) modellt implementálja, ami az 1. ábrán látható és az egyik legnépszerűbb implementációja a word2vec[1]. Egy másik lehetőség az ú.n. skip-gram modell alkalmazása, amikor a hálózat bemenete a célszó, az optimalizálás célja pedig e szó környezetének megjósolása.



1. ábra. A CBOW (*continous bag-of-words*) modell

[1] `https://code.google.com/archive/p/word2vec/`

A fastText algoritmus [5] a word2vec implementációját elsősorban azzal egészítette ki, hogy a szavak mellett az azokat alkotó karakter n-gramok reprezentációit is létrehozza, illetve a szó környezetében szereplő szavak karakter n-gramjait is szövegkörnyezetnek tekinti.

Ebben a cikkben a fastText algoritmus CBOW modelljének egy olyan módosított változatát mutatjuk be, amely egy modellen belül egyszerre hozza létre egy elemzett korpusz alapján a felszíni szóalakok és a hozzájuk rendelt akár több különböző típusú annotáció vektorreprezentációját.

3. A korpusz előkészítése

A jelen cikkben bemutatott kísérletek kiinduló anyagául a 2,25 milliárd token méretű angol Wikipedia korpusz[2] szolgált. A korpuszt a SpaCy keretrendszerbe[3] integrált angol neurális taggerrel és függőségi elemzővel elemeztük, amely lemmát, szófajcímkét és a szavak közötti függőségi viszonyokat rendelt az egyes szóalakokhoz. A SpaCy elemzéseit a feldolgozás első lépésében a CONNL-U formátum módosított változatában íratjuk ki (l.2. ábra). Majd további feldolgozás után egy olyan reprezentáció születik, amelyben a felszíni szóalakot annotációs címkék sorozata követi, melyek közül az első a lemma és a szófaj, és ezt az igevonzatok és a szabad határozók esetében az igei fej és az adott összetevőt a fejhez kapcsoló reláció címkéje követi. Az utóbbi típusú címkéből több is lehet, ha az adott szó több predikátummal is vonzatviszonyban áll (l.3. ábra).

A függőségi fa szerkezetű alapelemzéseket kiterjesztett függőségi reprezentációvá alakítjuk át. Az átalakítás során számos transzformációt végzünk, illetve számos új függőségi viszonyt veszünk fel. Azonos reprezentációt kap például egy adott aktív igealak tárgya, ugyanazon ige passzív változatának alanya, illetve egy befejezett melléknévi igenév vagy egy passzív vonatkozó mellékmondat által módosított főnév. A tagmondatok fejéhez kapcsolódó tartalmas szavak így olyan annotációt is kapnak a lemmájuk és a szófajcímkéjük mellett, amely explicit módon tartalmazza, hogy milyen igékhez milyen függőségi viszony kapcsolja őket. Ebben az annotációban az ú.n. *phrasal verb*-ök, a prepozíciós vonzatok és a kopulás szerkezetek összevontan tartalmazzák az igét és a prepozíciót, illetve a kopulát és a névszói állítmányt.

A 2. ábrán a *Bryozoa* egyrészt a *phylum* névszói állítmány alanya, másrészt egy olyan jelzői mellékmondat módosítja, amelynek feje a *know* 'ismer' ige. A *know* ige *as* prepozíciós vonzata pedig a *Polyzoa*, illetve az *animals* 'állatok'. A 3. ábrán látható, hogy a feldolgozás második lépése után már az eredetileg a jelzői mellékmondat által módosított *Bryozoa* a *know* ige tárgyaként szerepel. Bár a feldolgozás során a koordinált elemekre átterjesztjük a koordináció „fejének" vonzatviszonyait, egy elemzési hiba folytán az *Ectoprocta* a példában szereplő annotációban nem lett a *know* prepozíciós vonzata. Ugyan ebben a mondatban ez csak hibás elemzés eredménye, de egyébként feltehetőleg érdemes

[2] A `https://dumps.wikimedia.org/` linkről letölthető 2016. májusi verzió
[3] `https://spacy.io/`

lenne az appozitív szerkezetekre is elvégezni a koordinációra alkalmazott műveletet. Ugyancsak érdemes lenne a `compound` viszony mentén az angol összetett szavakat is egy elemmé összevonni.

```
#The Bryozoa, also known as the Polyzoa, Ectoprocta or commonly as moss animals, are a phylum
of aquatic invertebrate animals.
0       The     the     DET     DT      det     1       bryozoa PROPN
1       Bryozoa bryozoa PROPN   NNP     nsubj   16      be      VERB    _know#VB@<acl
                                                                        _be_phylum#VB@nsubj
2       ,       ,       PUNCT   ,       punct   1       bryozoa PROPN
3       also    also    ADV     RB      advmod  4       know    VERB    _know#VB@advmod
4       known   know    VERB    VBN     acl     1       bryozoa PROPN
5       as      as      ADP     IN      prep    4       know    VERB    _know#VB@prep
6       the     the     DET     DT      det     7       polyzoa PROPN
7       Polyzoa polyzoa PROPN   NNP     pobj    5       as      ADP     _know#VB@prep_as@pobj
8       ,       ,       PUNCT   ,       punct   7       polyzoa PROPN
9       Ectoprocta      ectoprocta      PROPN   NNP     appos   7       polyzoa PROPN
10      or      or      CCONJ   CC      cc      9       ectoprocta      PROPN
11      commonly        commonly        ADV     RB      advmod  12      as      ADP
                                                                        _know#VB@prep_as@advmod
12      as      as      ADP     IN      prep    4       know    VERB    _know#VB@prep
13      moss    moss    NOUN    NN      compound        14      animal  NOUN
14      animals animal  NOUN    NNS     pobj    12      as      ADP     _know#VB@prep_as@pobj
15      ,       ,       PUNCT   ,       punct   1       bryozoa PROPN
16      are     be      VERB    VBP     ROOT    16      be      VERB
17      a       a       DET     DT      det     18      phylum  NOUN
18      phylum  phylum  NOUN    NN      attr    16      be      VERB    _be_phylum#VB@attr
19      of      of      ADP     IN      prep    18      phylum  NOUN
20      aquatic aquatic ADJ     JJ      amod    22      animal  NOUN
21      invertebrate    invertebrate    ADJ     JJ      amod    22      animal  NOUN
22      animals animal  NOUN    NNS     pobj    19      of      ADP     _of@pobj
23      .       .       PUNCT   .       punct   16      be      VERB
```

2. ábra. A felhasznált korpusz egy mondatának annotációja a kiegészített CONLL-U formátumban a feldolgozás első lépése után

4. A módosított CBOW algoritmus

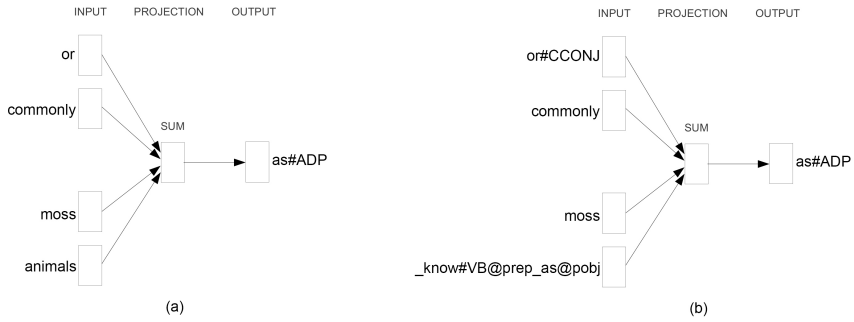
A szóbeágyazási modell építéséhez használt annotált korpuszban a szó (illetve írásjel) típusú tokeneket tetszőleges számú speciális, a 3. ábrán szereplő példában ■ jellel kezdődő címke típusú token követi. A fastText könyvtár CBOW algoritmusát úgy módosítottuk, hogy az ismertetett formájú bemeneti korpuszból olyan modellt építsen, amelyben a felszíni szóalakok és a hozzájuk tartozó annotációs címkék egyszerre vannak reprezentálva.

Az algoritmus első változatában a modell építése során csak a felszíni szóalakokat használtuk a betanítandó neurális hálózat bemeneteként megjelenő szövegkörnyezetként. Célszóként azonban a felszíni alakok és a hozzájuk tartozó címkék is megjelentek (4.(a) ábra). Ez a konfiguráció azonban olyan modellt hozott létre, amely a legcsekélyebb mértékben sem hasonlított ahhoz, amit kapni szerettünk volna. A címkék modell által generált vektorreprezentációja nemhogy hasonlított volna az adott címkével annotált szó reprezentációjához, hanem éppen ellenkezőleg, a lehető legnagyobb mértékben különbözött tőle (gyakorlatilag

```
The ■the#DET Bryozoa ■bryozoa#PROPN ■_know#VB@dobj ■_be_phylum#VB@nsubj
, ■,#PUNCT also ■also#ADV known ■know#VERB as ■as#ADP the ■the#DET
Polyzoa ■polyzoa#PROPN ■_know#VB@prep_as@pobj , ■,#PUNCT Ectoprocta
■ectoprocta#PROPN or ■or#CCONJ commonly ■commonly#ADV as ■as#ADP moss
■moss#NOUN animals ■animal#NOUN ■_know#VB@prep_as@pobj , ■,#PUNCT are
■be#VERB a ■a#DET phylum ■phylum#NOUN of ■of#ADP aquatic ■aquatic#ADJ
invertebrate ■invertebrate#ADJ animals ■animal#NOUN . ■.#PUNCT
```

3. ábra. A felhasznált korpusz egy mondatának annotációja a feldolgozás második lépése után

merőleges volt rá). Ennek az volt az oka, hogy a negative sampling algoritmus kizárólag negatív példaként látta bármely szó szövegkörnyezetében a címkéket és ezért a hálózat minden címke reprezentációját igyekezett a lehető legmesszebb juttatni a pozitív szövegkörnyezetként is előforduló felszíni szóalakok reprezentációjától (minden címke minden szótól a lehető legtávolabb helyezkedett el). Hogy valóban ez történik, azt úgy tettük egyértelművé, hogy egy olyan korpuszon tanítottuk be a modellt, amelyben minden szóalaknak pontosan egy címkéje volt, amely azonos volt magával a szóalakkal. Az így betanított modellben gyakorlatilag minden szó és a hozzá tartozó címke koszinusz távolsága (hasonlósága) lényegében nulla volt.



4. ábra. A módosított CBOW modell architektúrák

Ezt az anomáliát úgy küszöböltük ki, hogy a szövegkörnyezetben egyenletes eloszlással mintavételeztük a felszíni alakokat és címkéiket (4.(b) ábra). Így a címkék és a felszíni alakok egyaránt megjelentek pozitív és negatív tanítópéldaként, és így a kapott modell már sokkal inkább hasonlított ahhoz, amit vártunk.

A modell tanítása során 300 dimenziós vektorokat építettünk, és nem használtuk a fastText karakter-n-gram alapú modelljét (a -minn 0 -maxn 0 kapcsolókat használtuk). Egyénként az alapbeállításokkal futtattuk a tanítást: 5 token suga-

rú ablak, min. 5 előfordulás a szavakra és a címkékre, negatív mintavételezés 5 példával, stb.

5. Mire jó ez a modell

A modell egyszerre tárolja rendkívül kompakt formában a szavak felszíni alakjára, azok lemmájára és szófajára, illetve a közöttük tipikusan fennálló függőségi viszonyokra jellemző reprezentációkat. Az, hogy egyetlen modellen belül jelennek meg ezek az információk, lehetőséget ad arra, hogy a modellnek olyan kérdéseket tegyünk fel, hogy például *Mit isznak?, Mit bányásznak?, Miben hiszünk? Ki eszik? Mit csinálunk egy csontvázzal?*, stb. Annak kiértékelésére, hogy a modell az ilyen jellegű kérdésekre mennyire jó választ ad, sajnos nem állt rendelkezésünkre megfelelő gold standard erőforrás. A modellt jellemző átfogó kvantitatív kiértékelés helyett ezért kénytelenek vagyunk egy viszonylag szűk lexikai elemkészletre vonatkozó lekérdezéslista eredményeként kapott válaszok kézi kiértékelésére, illetve azokra a megfigyelésekre szorítkozni, amelyeket a [6] cikkben leírt szóbeágyazási modellek vizualizációjára szolgáló felületen keresztül a modellel kapcsolatban tettünk.

A felület képes arra, hogy a lekérdezésként megadott szóhoz megjelenítse a vektortérben hozzá legközelebb álló elemeket azok koszinuszhasonlóságával és gyakoriságával együtt. Lehetőség van arra, hogy szűrőket definiáljunk az így megjelenített legközelebbi szomszédok alakjára vonatkozóan. Ez ad lehetőséget például arra, hogy a *Mit isznak?* jellegű kérdésekre a rendszer által adott választ megkaphassuk. Ehhez egy olyan lekérdezést fogalmazunk meg, amelyben a „*drink* ‘iszik’ ige tárgya” objektum legközelebbi szomszédait keressük azzal a feltétellel, hogy a modellben szereplő elemeket megszűrjük, és csak a `NOUN` szófajcímkét tartalmazóakat tartjuk meg. Egy ilyen lekérdezés eredménye látható az 5. ábrán.

Ha címkét nem tartalmazó elemhez (felszíni szóalakhoz) indítunk lekérdezést, akkor a rendszer automatikusan olyan lekérdezésekkel egészíti ki az eredeti lekérdezést, amelyben az adott szót lemmának feltételezve hozzáfűzi ahhoz a korpuszban az adott lemmával előforduló szófajcímkéket. Így például a *can* lekérdezéshez megkapjuk válaszként egyrészt a *can* szóalak, másrészt a *can* ’tud, képes’ segédige, harmadrészt a *can* ‘konzerv’ főnév mint lemma, illetve az annotációhoz használt SpaCy tagger által hibásan más szófajúként címkézett elemek reprezentációjához legközelebbi elemeket (6. ábra). Valamely lemmával indítva a lekérdezést, a válaszban általában az első találatok között megkapjuk a szó ragozott alakjait, ragozott alakhoz pedig valahol az első találatok között lesz a szó lemmája.

5.1. Az elemzőrendszer hibái

A lekérdezőrendszer által a modellből visszaadott válaszok viszonylag koncentrált módon elénk tárják, hogy a korpusz annotációjához használt elemzőrendszer milyen változatos jellegű hibákat vezet be az annotációba. Ez már azon a szinten is megjelenik, hogy a generált kimenetben látunk olyan szófajcímkéket, illetve

0	_drink#VB@dobj	1	18814
1	drink#NOUN	0.7048	36413
2	drinking#NOUN	0.6033	27063
3	juice#NOUN	0.5734	14046
4	beer#NOUN	0.5699	41032
5	bottle#NOUN	0.5634	32186
6	drinker#NOUN	0.5593	3204
7	brandy#NOUN	0.5588	2957
8	champagne#NOUN	0.5379	3691
9	alcohol#NOUN	0.5374	54060
10	pint#NOUN	0.5339	2581
11	vodka#NOUN	0.5320	3045

5. ábra. A „*drink* ige tárgya” objektum legközelebbi szomszédai

0	can	1	2176270
1	can	1.00000001471	2176270
2	can#VERB	0.9967	2314044
3	may	0.8407	1367560
4	may#VERB	0.8396	1624916
5	could#VERB	0.8212	963212
6	could	0.8167	943772
7	must	0.8164	389083
8	must#VERB	0.8126	395927
9	will	0.7895	1220730
10	will#VERB	0.7882	1303870
11	should#VERB	0.7329	594608

0	can#VERB	1	2314044
1	can#VERB	1.00000009725	2314044
2	can	0.9967	2176270
3	may#VERB	0.8300	1624916
4	may	0.8295	1367560
5	could#VERB	0.8230	963212
6	could	0.8172	943772
7	must	0.8057	389083
8	must#VERB	0.8041	395927
9	will#VERB	0.7876	1303870
10	will	0.7840	1220730
11	should#VERB	0.7292	594608

0	can#NOUN	1	9460
1	can#NOUN	1.00000004784	9460
2	cans	0.9386	5678
3	bottle#NOUN	0.7667	32186
4	bottles	0.7545	12914
5	bag#NOUN	0.6997	33140
6	bags	0.6971	12483
7	bottle	0.6969	19000
8	tins	0.6909	783
9	carton#NOUN	0.6866	1506
10	bag	0.6566	19570
11	containers	0.6511	10963

0	can#PROPN	1	1980
1	can#PROPN	1.0000001992	1980
2	CAN	0.6573	2764
3	Can	0.4954	54105
4	Llauder	0.4206	6
5	tap#PROPN	0.3849	1660
6	SPAM	0.3815	416
7	be#PROPN	0.3780	1724
8	jerrycan#PROPN	0.3636	11
9	llauder#PROPN	0.3602	6
10	pan#PROPN	0.3542	40251
11	Pan	0.3481	38327

6. ábra. A *can* különböző előfordulásainak legközelebbi szomszédai

olyan függőségi relációkat a kérdésként megadott szóhoz mint lemmához rendelve, amelyről tudjuk, hogy hibás.

5.2. Szemantikailag jól behatárolható vonzatú igék – mit eszünk

Az elemzőrendszer által bevezetett hibák ellenére a modell válaszainak túlnyomó része elég meggyőző, különösen azokban az esetekben, ahol például az adott ige adott vonzatviszonyában szemantikailag jól behatárolható körbe tartozó lexikai elemek jelennek meg. A 7. ábrán az „*eat* ‘eszik’ ige tárgya” viszonylatában például valóban azt látjuk, hogy a listában túlnyomórészt ételek jelennek meg, köztük viszonylag előkelő helyen számos olyan különleges étel is, amely csak néhány alkalommal fordul elő a Wikipedia-korpuszban, ugyanakkor minden esetben mint az étkezés tárgya. Ezek az elemek így a modellben szorosabban kapcsolódnak

az evéshez, mint sok számunkra talán prototipikusabbnak tűnő étel, amelyeknek azonban számos egyéb aspektusával például az elkészítésük vagy feldolgozásuk módjával kapcsolatban rengeteg információ fordul elő a korpuszban, és így a reprezentációjuk távolabb esik az evés tárgya objektum vektorreprezentációjától. Az *eat* alanyára vonatkozóan nem jön létre a modellben ennyire jól körülhatárolható reprezentáció. A főleg enciklopédikus ismereteket tartalmazó korpuszban eleve nagyságrendekkel ritkábban jelenik meg az evés alanya testes lexikai elemmel kitöltve. Ugyanakkor a *leak* 'szivárog' ige lehetséges alanyai jobban körülhatárolható jelentésű csoportokba tagolódnak. Mint „a *leak* ige alanya" objektum 100 legközelebbi főnévi szomszédját klaszterezve ábrázoló 8. ábrán látható, a modell erre jól vissza is adja, hogy folyadékok, gázok, azok szállítására, tárolására, az áramlás és a nyomás szabályozására stb. szolgáló eszközök és konténerek, valamint információ (titkok, feljegyzések stb.) szokott (ki)szivárogni.

0	_eat#VB@dobj	1	78427
1	meat#NOUN	0.5810	50211
2	meal#NOUN	0.5749	33003
3	eating#NOUN	0.5535	3159
4	food#NOUN	0.5519	254592
5	manjuu#NOUN	0.5519	5
6	flesh#NOUN	0.5492	15264
7	carrot#NOUN	0.5461	4247
8	gebrocht#NOUN	0.5435	21
9	diet#NOUN	0.5392	35798
10	φαγεινν#NOUN	0.5374	6
11	taiyaki#NOUN	0.5181	28

7. ábra. Az „*eat* ige tárgya" objektum legközelebbi főnévként elemzett szomszédai

5.3. Más igék hasonló vonzatai – mit csinálunk még azzal, amit eszük

Ugyan arra a kérdésre, hogy *Mit eszünk?*, a választ megkaphatnánk pusztán az elemzett korpuszban az *eat* ige tárgyaként megjelölt lexikai elemek lekérdezésével is, a modell természetes módon lehetőséget ad az olyan kérdések megfogalmazására és megválaszolására is, hogy például *Milyen igék tárgya, vagy milyen igék valamilyen prepozíciós tárgya szokott olyasmi lenni, mint az eat 'eszik' ige tárgya?*. (9. ábra) Ha az így kapott igék listáját összevetjük a pusztán az *eat* igéhez közeli igék listájával, akkor jól látható, hogy az első kérdésre válaszként kapott

0	dilbit#NOUN dispersant#NOUN downflow#NOUN overfilling#NOUN csf#NOUN permeate#NOUN filtrate#NOUN fluid#NOUN liquid#NOUN leachate#NOUN slurry#NOUN seawater#NOUN coolant#NOUN
1	turbopump#NOUN oxidizer#NOUN oxidiser#NOUN antifreeze#NOUN pressurant#NOUN freon#NOUN
2	naphtha#NOUN flammable#NOUN combustible#NOUN gallon#NOUN avgas#NOUN gas#NOUN fuel#NOUN
3	methane#NOUN ammonia#NOUN fume#NOUN vapor#NOUN vapour#NOUN
4	protodermis#NOUN ampule#NOUN tetrafluoroethylene#NOUN styrol#NOUN hexafluoride#NOUN sulphide#NOUN trichloroethene#NOUN
5	caulking#NOUN drywall#NOUN penetrant#NOUN gasket#NOUN sealant#NOUN lubricant#NOUN
6	envelope#NOUN duct#NOUN pump#NOUN injector#NOUN stokehold#NOUN compartment#NOUN bilge#NOUN boiler#NOUN feedwater#NOUN pipework#NOUN pipe#NOUN
7	bubbler#NOUN diverter#NOUN deaerator#NOUN aspirator#NOUN thermosiphon#NOUN scrubber#NOUN dehumidifier#NOUN
8	manhole#NOUN downpipe#NOUN standpipe#NOUN
9	venting#NOUN pressurisation#NOUN umbilical#NOUN ductwork#NOUN ducting#NOUN
10	ballonet#NOUN gasbag#NOUN arcing#NOUN corium#NOUN preventer#NOUN calandria#NOUN drywell#NOUN pressurizer#NOUN
11	anthrax#NOUN sarin#NOUN secret#NOUN memo#NOUN leaker#NOUN bugging#NOUN wikileak#NOUN
12	slick#NOUN _leak#VB@nsubj leaking#NOUN leak#NOUN seepage#NOUN leakage#NOUN spill#NOUN spillage#NOUN
13	malfunction#NOUN overheating#NOUN shutoff#NOUN scram#NOUN firedamp#NOUN blackdamp#NOUN rupture#NOUN explosion#NOUN bursting#NOUN

8. ábra. A „*leak* ige alanya” objektum 100 legközelebbi főnévi szomszédja klaszterezve

listáról hiányoznak az *eat* ige lexikai reprezentációjához egyébként közeli, az étkezéstől különböző testi szükségletekre és élvezetekre vonatkozó igék, ugyanakkor megjelennek a pusztítással kapcsolatos igék, amely jól szemlélteti, hogy amit megeszünk, azt elpusztítjuk. A prepozíciós vonzatokra vonatkozó kérdés pedig egészen új evés- és fogyasztásigéket hoz be a listára, amelyek a szintaktikai disztribúció különbözősége miatt nem jelentek meg az előbbi halmazokban.

0	_eat#VB@dobj	1	78427
1	_eat#VB@dobj	1.0000	78427
2	_consume#VB@dobj	0.6521	34141
3	_drink#VB@dobj	0.6255	18814
4	_ingest#VB@dobj	0.5984	3965
5	_cook#VB@dobj	0.5980	10631
6	_swallow#VB@dobj	0.5414	5390
7	_eat_out#VB@dobj	0.5316	116
8	_devour#VB@dobj	0.5242	3080
9	_digest#VB@dobj	0.4930	2547
10	_eat_up#VB@dobj	0.4925	472

0	eat#VERB	1	109537
2	drink#VERB	0.6370	39160
3	consume#VERB	0.6151	44994
4	cook#VERB	0.6138	25875
5	devour#VERB	0.6011	4366
6	chew#VERB	0.5581	6142
7	vomit#VERB	0.5555	3111
8	ingest#VERB	0.5551	5736
9	sleep#VERB	0.5460	45645
10	swallow#VERB	0.5455	10402
11	munch#VERB	0.5348	210

0	_eat#VB@dobj	1	78427
1	_feed#VB@prep_on@pobj	0.5829	40039
2	_feast#VB@prep_on@pobj	0.5103	588
3	_taste#VB@prep_like@pobj	0.4741	577
4	_subsist#VB@prep_on@pobj	0.4582	1393
5	_regurgitate#VB@prep_by@pobj	0.4516	9
6	_dine#VB@prep_on@pobj	0.4512	163
7	_consume#VB@prep_as@pobj	0.4512	1050
8	_cook#VB@prep_like@pobj	0.4508	145
9	_be_rice#VB@prep_along@prep_with@pobj	0.4491	11
10	_forage#VB@prep_for@pobj	0.4439	1648

9. ábra. Az (1) *eat* ige tárgyához leghasonlóbb tárggyal rendelkező igék listája, (2) az *eat* ige tárgyához leghasonlóbb főnevek listája, és az (3) *eat* ige tárgyához leghasonlóbb prepozícióval rendelkező igék listája

5.4. A vonzatok irányából induló lekérdezések

Ha nem az igék, hanem a vonzatok irányából indulva teszünk fel kérdéseket a rendszernek, akkor arra kaphatunk választ, hogy egy-egy főnév tipikusan mi-

lyen igékkel áll valamilyen meghatározott viszonyban, például mi szokott történni vagy miket szoktunk csinálni az adott dologgal. Ha például megkérdezzük a rendszertől, hogy milyen igék tárgya a *skeleton* 'csontváz' főnév (10. ábra), akkor a számtalan ásatással, temetéssel, ravatalozással, rekonstrukcióval kapcsolatos ige mellett megjelenik a *char* 'elszenesedik' ige is, amelynek nem tárgya, hanem alanya az, ami elszenesedik. Ezt a hibát az annotálórendszerünkben alkalmazott azon feltételezés vezeti be, hogy a befejezett melléknévi igenevek által módosított főnév eredetileg az ige tárgya, azonban a páciens alanyú igékből is képezhető befejezett melléknévi igenév. Hasonlóképpen kerül a *Mit eszünk?* kérdésre kapott válasz elemei közé néhány idegen nyelvű 'enni' jelentésű szó, pl. az ógörög $\phi\alpha\gamma\epsilon\iota\nu$ vagy a finn *syödä*, amelyek az angol Wikipediában szereplő etimológiai fejtegetéseknek az elemzőrendszer általi félreelemzéséből jöttek létre (pl. a többször előforduló *Greek "*$\phi\alpha\gamma\epsilon\iota\nu$*" to eat* olyan alakú szerkezet, mint a *some food to eat*) (7. ábra).

0	skeleton#NOUN	1	18934
1	_unearth#VB@dobj	0.4950	5492
2	_discover#VB@dobj	0.4943	156506
3	_excavate#VB@dobj	0.4924	12528
4	_find#VB@dobj	0.4643	882721
5	_disarticulate#VB@dobj	0.4513	12
6	_fossilize#VB@dobj	0.4504	68
7	_uncover#VB@dobj	0.4426	18932
8	_derive#VB@dobj_that@prep_of@pobj	0.4402	12
9	_excavate_up#VB@dobj	0.4284	6
10	_mummify#VB@dobj	0.4098	144

10. ábra. A *skeleton* tárgyú igék listája

5.5. Eredmények

Idő és gold standard adatok hiányában sajnos csak egy viszonylag szűk lekérdezéslista eredményeként kapott válaszok pontosságának kiértékelésére volt módunk. Ötféle lekérdezés eredményét teszteltük:

1. adott ige tárgyaként milyen főnevek jelennek meg
2. az adott ige tárgyaként megjelenő főnevek milyen más igék tárgyaként jelennek meg
3. az adott ige tárgyaként megjelenő főnevek milyen más igék prepozíciós vonzataként jelennek meg
4. adott ige alanyaként milyen főnevek jelennek meg
5. adott főnév milyen igék tárgyaként jelenik meg

Az tárgyra vonatkozó első három lekérdezés eredményét a következő igékre értékeltük ki: *eat* 'eszik', drink 'iszik', mine 'bányászik' prove 'bizonyít' build

‘épít’ excavate ‘kiás, ásatásokat folytat’ terminate ‘megszüntet, befejez’ expect ‘vár’. Az alanyra vonatkozó lekérdezést a következő igékre: eat ‘eszik’, leak szivárog, kiszivárogtat’, explode ‘felrobban’, prove ‘bizonyít’, flow ‘folyik’, dry szárít’. Az utolsó „milyen igék tárgya” lekérdezést pedig a következő főnevekre futtattuk: skeleton ‘(csont)váz’, rice ‘rizs’, toy ‘játék’, key ‘kulcs’, lamp lámpa’, paper ‘papír, cikk’, lamb ‘bárány’.

Minden lekérdezéshez az első 40 jelöltet értékeltük ki. Helyesnek értékeltünk egy választ, ha az adott szó az adott viszonylatban helyes (evéshez étel, bányászathoz ásvány vagy ahonnan bányásznak – ez is lehet a *mine* ige tárgya, csontvázhoz kiásás, elásás stb.), illetve ha van olyan helyes és tipikus vonzat, amivel a másik ige által megnevezett tevékenységet tényleg szokták csinálni. Ha nem a megfelelő vonzatviszonyban jelent meg egy szó, azt nem fogadtuk el (pl. a *folyó-kanyarulatban, kanyonban* stb. *folyik víz*, de nem maga a *kanyon* folyik). Nem anyanyelvi beszélőként egyébként a válaszok első ránézésre zajosabbnak tűntek, mint amilyennek végül bizonyultak: utánanézve az eredményül kapott igéknek és főneveknek, a gyanús és számunkra ismeretlen szavak nagyobb részéről az derült ki, hogy valóban jó találat.

Az 1 táblázatban látható, hogy mit kaptunk az egyes lekérdezések eredményének illetve az összes lekérdezés aggregált eredményének pontosságára. Látható, hogy az első benyomásoknak megfelelően az alany viszonylatában kaptuk a leggyengébb eredményt, legjobban pedig a „melyik másik igéknek vannak hasonló tárgyai” kérdésre válaszolt a rendszer.

típus	pontosság
tárgy>főnév	0.85
tárgy>másik ige tárgya	0.95
tárgy>másik ige prep. vonzata	0.76
alany>főnév	0.71
főnév>milyen ige tárgya	0.82
all	0.83

1. táblázat. A rendszer válaszainak pontossága a tesztelt lekérdezéstípusokra

6. Konklúzió

A cikkben bemutattunk egy algoritmust és egy azzal generált konkrét modellt, amely egy közös vektortérmodellben ábrázolja felszíni szóalakok és az azokhoz rendelt annotációk disztribúcióalapú reprezentációját. A bemutatott modellben morfoszintaktikai és részleges szintaktikai függőségi annotációt használtunk. Mivel a reprezentáció nagyon kompakt, a modelltől olyan viszonylag komplex kérdésekre, mint hogy *Mit szoktunk még azokkal a dolgokkal csinálni, amit inni szoktunk?* is nagyon egyszerű formában és rendkívül gyorsan értelmes választ

kapunk. Az algoritmus természetesen bármilyen más annotáció és az annotált elemek közös disztribúciós modellbe gyúrását és az annotáció formai jegyeire alkalmazott szűrők segítségével a különbözőféleképpen annotált elemek, illetve az annotálatlan nyers adatok közötti disztribúciós hasonlóságok feltárását is lehetővé teszi.

Köszönetnyilvánítás

Jelen kutatás az FK 125217 és a PD 125216 számú projekt keretében az FK 17 és a PD 17 pályázati program finanszírozásában a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással valósult meg.

Hivatkozások

1. Ebert, S., Müller, T., Schütze, H.: LAMB: A good shepherd of morphologically rich languages. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA (2016)
2. Novák, A., Novák, B.: POS, ANA and LEM: Word embeddings built from annotated corpora perform better. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2018, Hanoi, Vietnam, Springer International Publishing, Cham. (2018)
3. Mikolov, T., Yih, W.t., Zweig, G.: 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, Georgia, Association for Computational Linguistics (2013) 746–751
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. CoRR **abs/1607.04606** (2016)
6. Novák, A., Siklósi, B., Wenszky, N.: Szóbeágyazási modellek vizualizációjára és böngészésére szolgáló webes felület. In Tanács, A., Varga, V., Vincze, V., eds.: XIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2017) 355–362