Lexikális behelyettesítés magyarul

Takács Dávid[1], Gábor Kata[2]

1 Prezi, e-mail: takdavid@gmail.com
2 INRIA, e-mail: kata.gabor@inria.fr

Kivonat Cikkünkben a lexikális behelyettesítési feladat (lexical substitution) magyarra adatptálását és két különböző megoldásának tesztelését tárgyaljuk. A lexikális behelyettesítés célja olyan algoritmus megalkotása, mely képes egy lexikális egység egy-egy mondatbeli előfordulását másik egységgel helyettesíteni olyan módon, hogy a mondat eredeti jelentését a lehető legjobban megőrízze. A feladat általunk kipróbált változatában az algoritmusnak kell elvégeznie a behelyettesítésre javasolt jelöltek generálását, valamint a szövegkörnyezetbe legjobban illeszkedő lexikális egység kiválasztását. A kiértékelés során a rendszer által javasolt jelölteket annotátorok által adott válaszokkal vetjük össze. A behelyettesítési feladat magyarra alkalmazásának célja, hogy felmérjük a disztribúciós szemantikai módszerek működésének hatékonyságát, valamint - a más nyelveken végzett kísérletekkel összevetve - képet kapjunk az esetlegesen felmerülő magyar-specifikus kihívásokról: a rendelkezésre álló erőforrásokról, illetve a nyelvi jellegzetességekből adódó problémákról.

Kulcsszavak: lexikális behelyettesítés, lexikális szemantika, disztribúciós szemantika

1. Bevezetés

A lexikális szemantikai kutatások, ezen belül a disztribúciós szemantika egyre nagyobb teret nyer a számítógépes nyelvészet különböző ágaiban (pl. szinonimadetektálás, szemantikai relációk tanulása, ontológiák/lexikai adatbázisok automatikus építése, dokumentum-kategorizálás). A korpuszból kinyert vektoriális reprezentációk kiértékelésének egyik lehetséges módja az eredmények intergálása valamilyen nyelvtechnológiai alkalmazásba, ám erre nem minden esetben nyílik közvetlen lehetőség. Ennek megfelelően többféle kiértékelési feladat és gold standard létezik a témában (l. SemEval kampányok). A vektoros szemantikai reprezentációk lehetővé teszik, hogy a szavak jelentése/szemantikai tartalma közötti hasonlóságot, vagy éppen a szisztematikus eltéréseket számszerűsítsük. Egyes kiértékelési szabványok az annotátorok által megadott (szintén numerikus) szemantikai hasonlósági értékeket [21] vagy plauzibilitási ítéleteket [19] használnak. A lexikális behelyettesítés előnye az előbbi kiértékelési módszerekkel szemben, hogy az annotátorok számára természetesebb, a nyelvi tudást közvetlenebbül mozgósító feladatot jelent, és nem támaszkodik előre meghatározott jelentéstárakra vagy nyelvészeti definíciókra (szemben például a hagyományos WSD

feladattal).

A lexikális behelyettesítés [14,5] célja olyan algoritmus megalkotása, mely képes egy lexikális egység (egyszerű szó, többszavas kifejezés) egy-egy mondatbeli előfordulását másik egységgel helyettesíteni olyan módon, hogy a mondat eredeti jelentését a lehető legjobban megőrizze. A feladat általunk kipróbált változatában az algoritmusnak kell elvégeznie a behelyettesítésre javasolt jelöltek (elsősorban, de nem kizárólag szinonimák) generálását, valamint a szövegkörnyezetbe legjobban illeszkedő lexikális egység kiválasztását. A kiértékelés során a rendszer által javasolt jelölteket annotátorok által adott válaszokkal vetjük össze. A behelyettesítési feladat magyarra alkalmazásának célja, hogy felmérjük a lexikális/disztribúciós szemantikai módszerek működésének hatékonyságát, valamint a más nyelveken végzett kísérletekkel összevetve képet kapjunk az esetlegesen felmerülő magyar-specifikus kihívásokról: a rendelkezésre álló erőforrásokról, illetve a nyelvi jellegzetességekből adódó problémákról.

A lexikális behelyettesítés jellemzően két részfeladtra osztható. Az első lépés a jelöltek kinyerése egy erre alkalmas jelentés- vagy szinonima adatbázisból (általában WordNetből), illetve korpuszból disztribúciós módszerekkel, pl. vektoriális közelség szerint. Bár sok kritika fogalmazódott meg a WordNet alkalmasságát illetően (elsősorban jelentésegyértelműsítési kontextusban [25,10] illetve a magyarra [9]), az angol nyelvű lexikális behelyettesítési verseny (SemEval 2007) során a legjobbnak bizonyult módszerek mégis mind támaszkodnak a WordNetre [8,13]. A második lépés a jelöltek rangsorolása aszerint, hogy melyik illeszkedik legjobban az adott szövegkörnyezetbe. Ez a feladat közel áll a jelentésegyértelműsítéshez, ám annotált szinonima-tár hiányában nem támaszkodhatunk felügyelt tanítási módszerekre. Lesk szótári definíciókat [11], Aguirre és Rigau WordNet alapú távolsági mértékeket [1], Carrol és McCarthy szemantikai szelekciós információkat [4] használ az egyértelműsítéshez. A disztribúciós szemantikában használt vektoriális szó-reprezentációk is alkalmasak rá, hogy szavak vagy nagyobb szövegegységek közötti hasonlósági mértékeket számítsunk belőlük. Egyes kutatások látens szemantikai dimenziókat alkalmaznak a szójelentések automatikus elkülönítésére és kontextusbeli egyértelműsítésére [12,23]. A szavak elosztott reprezentációján (*distributed lexical representations* vagy *word embedding*) alapuló nyelvmodellek [16] által generált vektoriális reprezentációk is alkalmasak arra, hogy rajtuk értelmezhető közelségi metrikák alapján döntsünk a szavak szemantikai közelségéről. Ezek a módszerek több SemEval versenyen - szóhasonlósági és szóanalógiás feladatok esetében - jól teljesítettek (Semeval 2012, 2014). A word2vec [16] és a GloVe [20] módszerek a szavakhoz vagy tetszőleges nagyobb egységekhez egy valós vektortérbeli vektort rendelnek úgy, hogy az így létrejött reprezentációra két tulajdonság jellemző: egyrészt az egymáshoz közel eső szavak szemantikai, illetve morfológiai értelemben is közeliek, másrészt a vektorok közötti vektoriális különbségek konzisztensek, és egyik szópárról a másikra átvihetők. Jellegzetes példa a szópárok között kinyerhető analógiás hasonlóságra: $\mathrm{v}(king) \sim \mathrm{v}(queen) = \mathrm{v}(man) \sim \mathrm{v}(woman)$. Ez a két tulajdonság indokolja a

módszerek közvetlen használhatóságát a szószemantikai feladatokban. A behelyettesítéses feladaton legújabban Ferret [6] végzett kísérletet francia nyelvre a word2vec által generált reprezentáció felhasználásával.

Kísérletünkben létrehozunk egy ilyen vektoros reprezentációt magyar szavakra, és ennek használhatóságát mindkét részfeladatra kipróbáljuk. Másodsorban egy WordNet alapú módszerrel próbálkozunk [7], mely a WordNet-beli lemmákat, illetve a köztük definiált hierarchikus kapcsolatokból származó információt kombinálja a disztribúciós szemantika és a dokumentumkategorizálás területén használt eljárásokkal. A célszó különböző jelentéseit és az ezekhez tartozó lexikai egységeket a WordNetből nyerjük ki. A WordNet-jelentések klaszterezése után a jelentéseket körülvevő releváns csomópontok körbejárásával tematikus kategóriákat képezünk, melyekhez ezután a korpuszból gyűjtünk kategória-specifikus kontextusokat. Az egyértelműsítés során a jelöltek vektoros reprezentációját vetjük össze a kontextus szavaival. Végül egy hibrid módszert is kipróbálunk, mely a WordNetből kinyert jelölteket kizárólag korpusz alapú disztribúciós információ felhasználásával rangsorolja.

2. Erőforrások

2.1. Magyar Nemzeti Szövegtár

A disztribúciós információ kinyeréséhez a Magyar Nemzeti Szövegtár [24] (MNSZ, továbbiakban: korpusz) első, kibővített, elemzett változatát használtuk [24,18]. A korpusz ezen változata 260 millió szót tartalmaz, egyértelműsítése az MNSZ egyétrelműsítő eszközlánc segítségével állt elő [17]. A kísérleteinkhez a korpusz lemmásított változatát használtuk.

2.2. WordNet

A WordNet olyan elektronikus lexikális szemantikai adatbázis, melyben a nyelvi fogalmak hálózatba szerveződnek. A fogalmakat szinonimahalmazok (synsetek), a közöttük lévő kapcsolatokat szemantikai relációk (hipernima, meronima, antonima stb.) reprezentálják. A WordNet alapegysége a szavakból álló szinonimaosztályok, úgynevezett synsetek.
A magyar WordNet [15] mintegy 40.000 synsetet tartalmaz, melyek nagy része meg van feleltetve ekvivalens angol WordNet synseteknek, így implicit módon más nyelvek wordneteinek is.

3. Jelöltek kinyerése a WordNet-ből

A célszó behelyettesítésére szánt jelölteket csoportosan nyerjük ki a WordNetből, ahol egy csoport a célszó egy jelentésének felel meg. Módszerünk célja egyfelől, hogy a különböző (és valóban megkülönböztetendő) jelentések mindegyikére találjunk jelöltet, másfelől, hogy az így kapott jelentések később hatékonyan felhasználhatók legyenek az egyértelműsítés során. Fontos azonban megjegyezni,

hogy nem közvetlen célunk a synsetek/jelentések közül választani a mondatbeli behelyettesítés során: a jelentések megkülönböztetése csak a jelölt-kinyerés és a jelentés-specifikus kontextusok kiválasztása során kerül előtérbe.

A szinonimák kinyerésének első lépéseként tehát azonosítjuk azokat a synseteket, melyek tartalmazzák a célszót. Mivel a korábbi, angol nyelvű kísérletekhez hasonlóan [8] a magyar WordNet esetében is előfordul, hogy a célszó az adott synsetből kinyerhető egyetlen szó, a keresést ilyen esetekben kiterjesztettük a hiperonimákra is[1].

Közismert, hogy a WordNet nagyon részletes jelentés-megkülönböztetésekkel operál [10]: számos olyan megkülönböztetést tartalmaz, mely az adott feladat kontextusában nem releváns, sőt, akár kifejezetten megnehezítheti a jelentés-egyértelműsítést [25]. Mivel a WordNet szerkezete önmagában nem feltétlenél nyújt információt a synsetek közti szemantikai távolságról, úgy döntöttünk, hogy a synsetek lexikális tartalmát felhasználva próbáljuk meg kiszűrni az irreleváns megkülönböztetéseket. Az egymással megegyező lexikális tartalmú synseteket tehát összevontuk, csakúgy, mint azokat a synset-párokat, melyek közül a kisebb synset részhalmazát alkotja a nagyobbnak. Az összevont synseteket a későbbiekben nem különböztetjük meg az eredeti synsetektől: valamennyit különböző jelentésként fogjuk kezelni.

4. Vektor alapú megközelítés

A word embedding előnye, hogy a szavak között többféle szemantikai viszonyt képezhetünk le egy valós vektortérben, és ezeket egyszerű numerikus, lineáris módszerekkel tárhatjuk fel. Megfigyelhető például, hogy a célszavaknak egy adott jelentéshez tartozó szinonimái többnyire egymás közelében fordulnak elő. Emiatt a tulajdonság miatt lehetőség van arra, hogy a lexikális behelyettesítési feladatot egy lépésben oldjuk meg: azokat a szavakat választjuk ki, amelyek tetszőlegesen választott szemantikai közelségmérték szerint a legközelebb esnek a célszóhoz, hiszen ezek a legalkalmasak jelöltnek. Ez a naiv megoldás egyszerű és intuitív, de nagyon jól szerepel a jelöltek generálásában - amint látni fogjuk az *oot* értékelésben (1 táblázat) - ezért baseline-nak tekinthetjük. Mi a skalárszorzatos megoldást implementáltuk *cosine* néven.

Megfigyelhetjük azt is, hogy az egy adott szemantikai mezőbe tartozó szavak szintén közelebb esnek egymáshoz. Mivel gyakran előfodul, hogy a célszóval egy mezőbe tartozó további szavak is vannak a kontextusban, ezért megkísérelhetünk a jelöltek közül aszerint választani, hogy mennyire esnek közel a kontextus szavaihoz. Precízebben, minden jelöltre kiszámolunk egy megfelelőségi mértéket úgy,

[1] A tesztadatok létrehozásakor a SemEval 2007 feladathoz hasonlóan a magyar behelyettesítési feladatban is megengedtük az általánosabb értelmű fogalommal való helyettesítést.

hogy összegezzük a jelölt és minden egyes kontextusszó közelségi mértékét, és eszerint rendezve a legjobb jelölteket adjuk vissza. Így [27] megközelítését implementáljuk. A közelségi mérték, többek között, lehet skalárszorzat és euklideszi távolság, ezeknek az alkalmazott mérték szerint *bestcosinecontext* és *bestl2context* a neve a kísérletünkben.

Lehetőségünk van arra is, hogy a célszóhoz közel eső jelöltek közül azokat részesítsük előnyben, amelyek a kontextushoz közelebb vannak, azaz amelyek a célszó vektorától a kontextus eredő vektora irányába esnek. Ez a módszer több paramétert is elfogad: a kontextus eredő vektorának kiszámításakor az egyes elemeket különféleképpen súlyozhatjuk, és megválaszthatjuk azt is, hogy milyen távolságban keressük a helyettesítő szavakat az eredetitől, azaz a kontextus és a célszó milyen lineáris kombinációját számítjuk ki. A kísérletezés során azt találtuk, hogy ha nagy súlyt adunk a kontextusnak, azaz távolabb keresgélünk az eredeti célszótól, rátalálhatunk egy-egy távolabbi szinonimára is, de nagyobb számban találunk használhatatlan (nem behelyettesíthető) szavakat is. Ezt a tulajdonságot a kiértékelő adatok is igazolják: az összes kísérleti konfigurációból az *averagecontext* szerepel az oot-kiértékelésben a legjobban (azaz nagyon jó jelölteket is talál), de a best-értékelése nem jó (azaz zajos: sok jelöltje nem megfelelő).

Ugyanezeket a számításokat elvégezhetjük tetszőleges szavakra, amelyeknek ismerjük a vektortérbeli reprezentációit. Ez lehetőséget ad egy egyszerű hibridizációra: más módszerek által generált jelölteket tudunk a fent leírt módszerekkel kiértékelni és sorbarendezni. A kiértékelésben látható, hogy a *hybrid bestcosinecontext* konfiguráció ötvözi a különböző megközelítések előnyeit, és összességében a legjobb eredményeket adja. Ebben a konfigurációban a WordNet-ből kinyert jelölteket rangsoroltuk a *bestcosinecontext* érték alapján.

5. WordNet alapú megközelítés

A WordNet alapú megközelítés motivációja a WordNet szerkezetében rejlő információ kihasználása és ötvözése a korpusz alapú megközelítés előnyeivel. A folyamat első lépésében a jelölteket a célszót tartalmazó synsetekből nyertük ki. Ezután a célszó synsetjeit csoportosítjuk, és az így kapott jelentésekhez a synsetek tartalmát felhasználva keresénk olyan kontextusokat a korpuszban, melyek az adott jelentésre specifikusak, és így megkülönböztető erővel bírnak az egyértelműsítés során. Ezek a kontextusok fogják képezni a célszó specifikus egyértelműsítő vektoterét, melyen valamennyi jelöltet elhelyezzék. Végül az utolsó lépésben a mondat szavait vetjék össze a jelölteknek az egyértelműsítő kontextusokra felvett értékeivel, és eszerint rangsoroljuk őket.

5.1. Jelentések megkülönböztetése

A WordNet nemcsak azt teszi lehetővé, hogy releváns behelyettesítési jelölteket generáljunk. A munkafolyamat következő lépéseinek célja, hogy a WordNet szerkezetét és lexikális tartalmát kihasználva összegyűjtsük azokat a kontextusokat,

melyek vélhetően releváns információt hordoznak a jelentések, és ezen keresztül a behelyettesítési jelöltek kontextusbeli egyértelműsítéséhez. Feltételezésünk szerint e megközelítés egyik előnye lehet, hogy egy tetszőleges korpuszban a szavak ritkább jelentései meglehetősen alulreprezentáltak, ezért a tisztán disztribúciós, illetve nyelvmodell alapú egyértelműsítési módszerek ezen jelentések előfordulásait nehezen tudják azonosítani. Ha azonban rendelkezésünkre áll egy lista a szó lehetséges jelentéseiről, valamint a különböző jelentésekkel kapcsolatba hozható szavakról - melyek a wordnet-hierarchiából könnyedén kinyerhetők - lehetőségünk nyílik arra, hogy olyan kontextusokat is figyelembe vegyünk az egyértelműsítés során, melyek egyébként a célszóval, illetve a behelyettesítési jelöltjeivel nem, vagy csak kevésszer fordulnak elő a korpuszban. A korábbiakban bemutatott tisztán disztribúciós alapú megközelítéssel szemben tehát a második kísérlet célja megvizsgálni, hatékony lehet-e egy külső erőforrás bevonása az egyértelműsítési folyamatba.

A jelentések megkülönböztetéséhez a 3 pontban bemutatott synset-klaszterekből indulunk ki. Célunk, hogy minden jelentést megfelelő mennyiségű lexikális elemmel tudjunk reprezentálni. Ezen lexikális elemek részben az előző pontban kinyert behelyettesítési jelöltek (szinonimák, illetve ezek hiányában hiperonimák). A következő lépésben azokhoz a jelentésekhez, melyek háromnál kevesebb lexikális elemmel hozhatók kapcsolatba, újabb szavakat kerestünk a WordNet környező synsetjeiben: minden olyan synsetben, mely közvetlenül hiperonim, category_domain, illetve mero_part relációban áll az adott synsettel (összevonás esetén a kiinduló synsetek valamelyikével). Ezen kapcsolódó synsetek lexikális tartalmát a jelentéshez csatoltuk [2].

A behelyettesítési jelölteket a későbbiekben úgy kívánjuk kiválasztani, hogy a mondatbeli kontextust összevetjük az egyes jelentésekre jellemző kontextusokkal. Ehhez létre kell hoznunk egy olyan egyértelműsítési vektorteret, mely a célszó összes különböző jelentésének leginkább jellemző kontextusait tartalmazza, és a lehető legkevésbé favorizálja a gyakori jelentéseket. A célszó jelentéseihez társított valamennyi szó minden előfordulását figyelembe véve kinyertük a korpuszból a szavak kontextusait. Szintaktikai elemzés híján kontextusként kezeltünk minden, a szó környezetében előforduló lemmát, kettő, illetve öt szóból álló kontextus-ablakok használatával (ezek a szó-ablakok bizonyultak a legeredményesebbnek [2] részletes összehasonlító vizsgálatában). Ezután minden S synset-klaszterre kikerestük azokat a w szavakat, melyek az adott synset-klaszter s szavaira leginkább jellemzőek az alábbi képlet szerint :

$$spec_{w,S} = \sum_{s \in S} weight_{s,w} \tag{1}$$

[2] Fontos megjegyezni, hogy az így kinyert új szavak már nem számítanak behelyettesítési jelöltnek, kizárólag a jelentések jellemző kontextusainak feltérképezéséhez használtuk őket.

ahol a *weight* súly a PMI (Pointwise Mutual Information) egy normalizált változata. A PMI kollokációk és jellemző kontextusok kinyerésének bevett módja:

$$PMI_{w_1,w_2} = log\frac{p(w_1,w_2)}{p(w_1)p(w_2)} \tag{2}$$

A PMI hátránya azonban, hogy erősen favorizálja a ritka kontextusokat. Esetünkben ez az egyértelműsítési feladatot megnehezíti, ezért két normalizálási eljárást is kipróbáltunk, hogy kivédjük a ritka kontextusok felülreprezentálását [3,22]:

$$NPMI_{w_1,w_2} = log\frac{p(w_1,w_2)}{p(w_1)p(w_2)}/-log\ \ p(w_1,w_2) \tag{3}$$

$$squaredPMI_{w_1,w_2} = log\frac{p^2(w_1,w_2)}{p(w_1)p(w_2)} \tag{4}$$

A jelentés-specifikus kontextusokat a fenti értékek szerint rangsoroltuk. A célszó egyértelműsítő vektotere a célszó jelentéseire jellemző kontextusok uniójából áll elő: jelentésenként a legjellemzőbb 200, illetve 500 kontextust tartottuk meg. Előfordulhat, hogy egy kontextus több jelentésre is jellemző, ezt is hasznos információnak tekintettük. A vektorterek mérete így a célszó jelentései számának, és a jelentések átfedésének függvényében változó.

Mivel feladatunk nem közvetlenül a jelentésegyértelműsítés, hanem a legmegfelelőbb jelentés kiválasztása, ezért a célszóhoz tartozó összes behelyettsítési jelöltet elhelyezzük a fenti vektortérben. Ehhez három különböző reprezentációt használtunk: a célszó és a kontextus együttes előfodulásainak számát ($freq(c,w)$), a célszó kontextus melletti relatív gyakoriságát:

$$\frac{freq(c,w)}{freq(c)} \tag{5}$$

illetve a relatív gyakoriság normalizált értékét:

$$\frac{\frac{freq(c,w)}{freq(c)} - \mu}{\sqrt{\frac{\sigma^2}{N}}} \tag{6}$$

Az együttes előfordulásokat ugyanolyan kontextus-ablakot használva számoltuk, ahogyan az egyértelműsítő vektotereket előállítottuk.

5.2. Egyértelműsítés

A mondatbeli kontextusba illeszkedő jelölt kiválasztásakor a jelölteknek az egyértelműsítő vektortérben alkotott reprezentációját vetjük össze a mondat szavaival. Ehhez először is lemmatizáltuk a teszmondatokat az MNSZ egyértelműsítő eszközlánc [17] segítségével. A mondat p vektora úgy áll elő, hogy a mondat i

szavait is ráképezzük a célszó egyértelműsítő vektorterére egy karakterisztikus függvénnyel:

$$p_i = \begin{cases} 1, & \text{ha i előfordul a mondatban} \\ 0 & \text{egyébként} \end{cases}$$

A jelölteket eztuán a p mondat-vektor és a jelölt c egyértelműsítő vektora közti kompatibilitás szerint rangsoroljuk, melyet a következő képlet szerint számolunk ki:

$$compatibility(c, p) = c \cdot p = \sum_{i=1}^{n} c_i \times p_i \tag{7}$$

A mondat szavai közül tehát csak azokat vesszük figyelembe, melyek a célszó valamelyik jelentéséhez specifikus kontextusként lettek társítva, és azzal a súllyal esnek latba, amit az adott jelölt hozzájuk rendel a korpuszbeli előfordulásai alapján.

6. Kiértékelés és perspektívák

Az eredmények kiértékeléséhez használt adatok elkészítésekor a McCarthy és Navigli (Semeval 2007), illetve a Fabre és tsai (SemDis 2014) által követett módszert vettük alapul. Tíz poliszém főnevet választottunk, melyek minden jelentésükben rendelkeznek egytagú szinonimával. Feltétel volt továbbá, hogy maga a főnév, valamint szinonimái (jelentésenként legalább egy) is kellő mértékben reprezentálva legyenek a rendelkezésre álló korpuszban. Minden célszóhoz 10-10 példamondatot kerestünk oly módon, hogy minden szónak minden jelentése reprezentálva legyen. Az adatokat és az instrukciókat a Qualtrics online felmérés-készítő alkamazás segítségével osztottuk meg. A célszó mondatbeli előfordulásaihoz legalább 3-3 önkéntes annotátor javasolt mondatonként legfeljebb négy behelyettesíthető lexikai elemet. A rendszer által javasolt megoldásokat ezekkel a kiértékelési adatokkal vetettük össze, figyelembe véve azt is, hogy a rendszer megoldása hány annotátor javaslatai között szerepel.

Hasonlóan a korábbi lexikális behelyettesítési feladatokhoz, kétféle mértéket használtunk a gold sztenderddel való összevetéshez [14] : a *best* mérték a rendszer elsőnek rangsorolt javaslatát veszi figyelembe, míg az *oot* (*out of ten*) azt méri, hogy az első tíz javaslat között hány jó jelölt szerepel, a sorrendre való tekintet nélkül. Mivel a WordNet alapú módszerek gyakran ennél kevesebb (és csak nagyon ritkán több) jelöltből indulnak ki, ezért ebben az esetben az *oot* érték inkább a WordNet mint forrás lefedettségének indikátora. Az egy mondatra adott *best* érték azt mutatja meg, hogy a rendszer által javasolt legjobb jelölt hányszor szerepel az annotátorok megoldásai között (minél többen javasolták, annál valószínűbb, hogy erős jelölt), elosztva az annotátorok által javasolt összes megoldás számával. A rendszer *best* mutatója az összes mondatra kapott best értékek átlaga. Az *oot* érték számításakor a rendszer által javasolt első tíz jelölt

pontszámait (azaz szintén az egyes annotátorokéval egyező javaslatok pontszámát) osztjuk el az összes annotátor által tett javaslatok számával.

Módszer	**BEST**	**OOT**
veconly bestcosinecontext	**0.06702**	0.267739
veconly bestl2context	0.05913	0.25895
veconly averagecontext	0.02997	**0.29371**
veconly cosine	0.02806	0.28349
wnet.lemma2.size200.NPMI.rawcount.txt	**0.11064**	**0.23881**
wnet.lemma5.size200.NPMI.rawcount.txt	0.09560	0.23881
wnet.lemma2.size200.NPMI.relfreqnorm.txt	0.09451	0.22743
wnet.lemma2.size500.NPMI.rawcount.txt	0.09423	0.23881
wnet.lemma5.size500.NPMI.rawcount.txt	0.08731	0.23881
wnet.lemma5.size200.NPMI.relfreqnorm.txt	0.08717	0.22410
hybrid bestcosinecontext	**0.11029**	**0.24003**
hybrid bestl2context	0.07988	0.23741

1. táblázat: Eredmények módszerenkénti bontásban

Amint az 1 táblázat mutatja, az *oot* értékek elég konzisztensek a WordNet alapú jelölt-generálás esetében, ami nem meglepő, hiszen a WordNet csak kevés szó esetében adott tíznél több jelöltet. Érdekes azonban, hogy a vektor alapú megközelítések minden esetben túlszárnyalták a WordNet alapú jelölt-generálást, némileg több jó jelöltet állítva az első tízben. Összességében módszereink az esetek 40-45 százalékában képesek legalább egy jó jelöltet állítani, ami körülbelül megfelel a nemzetközi eredményeknek [5]. Ugyanakkor a tisztán vektor alapú módszerek a WordNetre támaszkodó megközelítésnél gyengébben teljesítettek a legjobb jelölt kiválasztásában, az átlagot és a legkedvezőbb beállításokat tekinve is.

A WordNet-alapú módszerek eredményei meglehetősen nagy szórást mutatnak. A figyelembe vett paraméterek közül a legnagyobb jelentősége a kontextusok kiválasztásához használt specifikussági mértéknek van: az NPMI sokkal jobban teljesít, mint a squaredPMI. A kontextusok fajtái közül a kétszavas ablak pontosabbnak bizonyult, mint az ötszavas, és az egyértelműsítő vektortér méretének növelése csökkentette az egyértelműsítés pontosságát. Összességében tehát a kevesebb, specifikusabb és közvetlenebb kontextusokból képzett információ bizonyult a leghasznosabbnak. A legjobb eredményt azonban, várakozásunknak megfelelően, a hibrid módszerrel értük el.

Eddigi munkánk természetes folytatása lehet az MNSZ2 teljes anyagának felhasználása a disztribúciós modellek számításakor. Igéretes lehetőség a hibrid megközelítés további kombinációinak kiértékelése, továbbá az optimális vektoros reprezentáció megkeresése a paraméterek finomhangolásával. További annotátori munkával lehetséges lenne a tesztanyag összekötése a már elérhető magyar jelentésegyértelműsítő korpusszal (hunwsd [26]), illetve az általunk gyűjtött goldstandard annotálása jelentésekkel.

A kísérletek során kézi és gépi munkával létrehozott adatokat szabadon elérhetővé tesszük.

7. Köszönetnyilvánítás

Ezúton köszönjük Oravecz Csabának az MNSZ-egyértelműsítő eszközlánc rendelkezésünkre bocsátását és a használatában nyújtott segítségét. Köszönetet mondunk továbbá minden önkéntesnek, akik közreműködtek a kiértékelési adatok létrehozásában.

Hivatkozások

1. Aguirre E., Rigau G.: Word Sense Disambiguation using Conceptual Density. In Proceedings of COLING'96, 16–22. (1996)
2. 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 ACL Conference. (2014)
3. Bouma G.: Normalized (Pointwise) Mutual Information in Collocation Extraction. In: From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference. 31–40 (2009),
4. Carroll J., McCarthy D.: Word Sense Disambiguation Using Automatically Acquired Verbal Preferences. In Computers and the Humanitie. vol.34 109–114. (2000)
5. Fabre C., Hathout N., Ho-Dac L., Morlane-Hondère F., Muller P., Sajous F., Tanguy L., Van de Cruys T.: Présentation de l'atelier SemDis 2014 : Sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In Proceedings of the TALN 2014 Conference, Marseille, France. (2014)
6. Ferret O.: Using a generic neural model for lexical substitution (Utiliser un modèle neuronal générique pour la substitution lexicale) In TALN-RECITAL 2014 Workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle, 218–227 (2014)
7. Gábor K.: The WoDiS System - WOlf and DIStributions for Lexical Substitution (Le système WoDiS - WOLF et DIStributions pour la substitution lexicale) In TALN-RECITAL 2014 Workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle, 228–237 (2014)
8. Hassan S., Csomai A., Banea C., Sinha R., Mihalcea R.: Unt : Subfinder : Combining knowledge sources for automatic lexical substitution. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic : Association for Computational Linguistics. (2007)

9. Héja E., Kuti J., Sass B. Jelentésegyértelműsítés - egyértelmű jelentésítés? In: MSZNY2009, VI. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged. 348–352. (2009)
10. Ide N., Wilks Y.: Making sense about sense. In Word Sense Disambiguation : Algorithms and Applications, vol. 33 of Text, Speech and Language Technology, 47–74. Dordrecht, The Netherlands : Springer. (2006)
11. Lesk M.: Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from a ice cream cone. In Proceedings of SIGDOC-1986 (1986)
12. Lin D., Pantel P.: Concept discovery from text. In Proceedings of the 19th International Conference on Computational Linguistics (COLING) (2002)
13. Martinez D., Kim S. N., Baldwin T.: Melb-mkb : Lexical substitution system based on relatives in context. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, CzechRepublic : Association for Computational Linguistics (2007)
14. McCarthy D., Navigli R.: The English Lexical Substitution Task. in Language Resources and Evaluation, 43(2). 139–159 (2009).
15. Miháltz M., Hatvani Cs., Kuti J., Szarvas Gy., Csirik J., Prószéky G., Váradi T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary, 311–321. (2008)
16. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.: Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS. (2013)
17. Oravecz Cs., Dienes P.: Efficient Stochastic Part-of-Speech tagging for Hungarian. In Proceedings of the Third International Conference on Language Resources and Evaluation, pages 710-717. (2002)
18. Oravecz Cs., Váradi T., Sass B.: The Hungarian Gigaword Corpus. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC) European Language Resources Association. (2014).
19. Padó U.: The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Sentence Processing. Dissertation, Saarland University, Saarbucken. (2007)
20. Pennington J., Socher R., Manning C.: Glove: global vectors for word representation Empirical Methods in Natural Language Processing (EMNLP), 2014 (to appear)
21. Rubenstein H., Goodenough J.: Contextual correlates of synonymy. in Communications of the ACM, 8(10), 627–633 (1965).
22. Thanopoulos A., Fakotakis N., Kokkinakis G.: Comparative Evaluation of Collocation Extraction Metrics. In Proceedings of the Third International Conference on Language Resources and Evaluation. (2002)
23. van de Cruys T., Poibeau T., Korhonen A.: Latent vector weighting for word meaning in context. In Proceedings of the EMNLP 2011 Conference, 1012–1022 (2011)
24. Váradi T.: The Hungarian National Corpus. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC) European Language Resources Association.385–389. (2002)
25. Véronis J.: Sense tagging : does it make sense ? In Corpus Linguistics by the Lune : a festschrift for Geoffrey Leech. Frankfurt : Peter Lang (2003)
26. Vincze V., Szarvas Gy., Almási A., Szauter D., Ormándi R., Farkas R., Hatvani Cs., Csirik J.: Hungarian Word-sense Disambiguated Corpus. In: Proceedings of

6th International Conference on Language Resources and Evaluation, Marrakech, Morocco. (2008)
27. Zweig G., Platt J. C., Meek C., Burges C. J., Yessenalina A., Liu Q.: Computational approaches to sentence completion. In 50th Annual Meeting of the Association for Computational Linguistics (ACL), p. 601–610, Jeju Island, Korea. (2012)