Magyar és angol szavak szemantikai hasonlóságának automatikus kiszámítása

Dobó András, Csirik János

Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
H-6720 Szeged, Árpád tér 2.
{dobo,csirik}@inf.u-szeged.hu

Kivonat: Szavak szemantikai hasonlóságának vizsgálata számos számítógépes nyelvészeti probléma megoldásában fontos szerepet tölt be. Habár már sok különféle módszer létezik e feladatra, az eredményeken még mindig lehetne javítani. Egy korábbi kutatásunk során olyan módszereket fejlesztettünk ki angol szavak szemantikai hasonlóságának automatikus megállapítására, amelyek nagyméretű statikus korpuszokból kinyert statisztikai információ alapján készítenek bináris vagy numerikus tulajdonságvektorokat a szavakhoz, majd a szavak hasonlóságát vektoraik hasonlóságaként számolják ki. Jelen cikkünkben korábbi módszereink továbbfejlesztett változatát mutatjuk be, melyek a korábbiakhoz képest új vektorhasonlóságokat is felhasználnak, továbbá már alkalmasak magyar szavak közötti szemantikai hasonlóság megállapítására is, mely legjobb tudásunk szerint egyedülálló. Az algoritmusok angol és magyar nyelvű tesztadatbázisokon kiértékelve is versenyképes eredményeket érnek el.

1 Bevezetés

Számos számítógépes nyelvészeti probléma megoldásában, mint például az információkinyerésben, helyesírás-javításban és szójelentés-egyértelműsítésben, szavak szemantikai hasonlóságának az ismerete nagy segítséget nyújthat. Ezért az elmúlt nagyjából 20 évben számos kutatás irányult szavak jelentésbeli hasonlóságának automatikus meghatározására. A legtöbb erre a feladatra kialakított módszer webes kereséseket (pl. Google vagy Yahoo!), illetve lexikai adatbázisokat (pl. WordNet vagy Roget's Thesaurus) alkalmaz a hasonlóság kiszámítására. Ugyan ezek használata sok szempontból előnyös és az őket használó algoritmusok általában jól működnek, mint ahogy azt korábban is bemutattuk [1], sok hátránnyal is rendelkeznek.

Ezért korábbi kutatásunk [1] során olyan módszereket készítettünk, melyek sem webes kereséseket sem lexikai adatbázisokat nem használnak, és pusztán statikus korpuszok felhasználásával képesek angol szavak szemantikai hasonlóságának az automatikus kiszámítására[1]. Ezek a módszerek először létrehoznak egy tulajdonságvektort minden szóhoz a felhasznált korpuszban található környezeti szavak vagy

[1] Habár felhasználtuk a WordNetet szavak lemmájának meghatározására, semmi másra nem használtuk. Ez pedig helyettesíthető lenne egyéb módszerekkel.

nyelvtani kapcsolatok és valamely súlyozási módszer segítségével. Ezután szavak hasonlóságát a vektoraik hasonlóságaként számítják ki.

Jelen cikkünkben e korábbi módszerek továbbfejlesztett változatát mutatjuk be. Ezek a módszerek a már korábban használt egy bináris és kettő numerikus vektorhasonlóság mellett további három numerikus hasonlósági mértéket használnak fel. Továbbá, már nem csak angol, hanem magyar szavak közötti szemantikai hasonlóság megállapítására is alkalmasak, mely legjobb tudásunk szerint egyedülálló. A különálló módszerek mellett azok kombinációit is kipróbáltuk, és a korábbi angol nyelvű tesztadatbázisok mellett magyar nyelvű tesztadatbázisokon is kiértékeltük őket.

A következő szakasz a témához kapcsolódó egyéb kutatásokat foglalja röviden össze. Ez után algoritmusaink bemutatása következik, amit az algoritmus eredményeinek prezentálása és a konklúziók levonása követnek.

2 Kapcsolódó munkák

Habár már számos kutatás vizsgálta angol szavak szemantikai hasonlóságának automatikus megállapítását, legjobb tudásunk szerint a miénk az első olyan módszer, mely magyar szavak szemantikai hasonlóságával foglalkozik. Ezért ebben az alfejezetben az eddig publikált, angol szavak szemantikai hasonlóságának kiszámításával foglalkozó módszereket jellemezzük röviden (részletesebb áttekintésük korábbi cikkünkben található meg [1]). Ezeket a felhasznált adatforrások és a működésük alapján három nagy kategóriába sorolhatjuk.

Sok módszer nagyméretű lexikai adatbázisokban tárolt információt használ fel, és a kinyert információk alapján számolja ki szavak szemantikai hasonlóságát. A legtöbb a WordNetet használja, de léteznek olyanok is, melyek a Roget's Thesaurust. Egy nagyon jó példa erre Tsatsaronis et al. [2] módszere, mely egy WordNet alapú hasonlósági pontszámot definiál. Ennek a kiszámításához figyelembe veszi a szavak WordNetbeli távolságát, a közöttük lévő szavak WordNetbeli mélységét és a szavak közti kapcsolatok típusait. Módszerüket kibővítették, hogy ne csak szavak, hanem hosszabb szövegrészek hasonlóságának megállapítására is alkalmas legyen.

Más módszerek szavak hasonlóságának becsléséhez webes kereséseket indítanak a vizsgált szavakkal, és a visszaadott találatok számát, valamint a visszaadott szövegtöredékeket használják fel. Például Higgins [3] webes kereséseket indít a vizsgált szavakkal külön-külön és együtt is, majd a hasonlóságukat a visszaadott találatok számából kiszámított pontonkénti kölcsönös információként adja meg.

Léteznek olyan módszerek is, melyek egy tulajdonságvektort képeznek minden szóhoz a szó egy nagyméretű korpuszban talált környezetei alapján. Habár a mi módszereink hasonlóak ezekhez a módszerekhez, a mieink új tulajdonságokat, súlyozási módszereket és vektorhasonlósági mértékeket használnak a már korábban is alkalmazottak mellett. Egy ebbe a kategóriába tartozó módszer például Rappé [4] is, mely minden szóhoz egy numerikus tulajdonságvektort készít a szó megtalált előfordulási környezetei alapján. Ezekben a vektorokban azok a környezeti szavak találhatók meg, melyek a vizsgált szótól legfeljebb két szó távolságra találhatóak a korpuszban, és a súlyuk olyan jól ismert szókapcsolati mértékeken alapszik, mint a pontonkénti kölcsönös információ. A vektorok által adott mátrixot ezután összetömöríti az SVD mód-

szer segítségével. Végül a szavak hasonlósága a tömörített vektoraik hasonlóságaként kerül kiszámításra.

Mindhárom fő módszertípusnak megvannak az előnyei és a hátrányai, ezért sok kutatás oly módon próbálta meg az addig elért eredményeket tovább javítani, hogy különböző típusú módszereket kombinált, így próbálva azok előnyeit ötvözni. Turney et al. [5] módszere például négy különböző módszer ötvözete. Az első az LSA [6], a második egy webes kereséseken alapuló módszer (PMI-IR), a harmadik egy online fogalomtárban keres (Wordsmyth thesaurus online) és az utolsó webes keresések által visszaadott szövegtöredékeket dolgoz fel. Ezt a négy módszert többféleképpen kombinálták (például a szorzat szabállyal) a végső hasonlóság kiszámításához.

3 Módszereink

Módszereink alapötlete, mint sok egyéb módszer alapötlete, az, hogy a szemantikailag hasonló szavak hasonlóan viselkednek és hasonló szövegkörnyezetekben fordulnak elő. Ezért módszereink minden szóhoz egy tulajdonságvektort képeznek statikus korpuszokból kinyert statisztikai információ alapján. Ezen vektorokban különféle tulajdonságokat, így például a szavak környezetében előforduló úgynevezett környezeti szavakat és a szavakhoz kapcsolódó nyelvtani kapcsolatokat alkalmaznak. Azért, hogy a vektorokon belül a tulajdonságok fontosságát reprezentálni tudják, különféle súlyozásokat alkalmaznak. A szavak hasonlóságát az algoritmusok a létrejött súlyozott vektorok hasonlóságaként definiálják.

A következő alfejezetben korábbi, kizárólag angol szavak szemantikai hasonlóságának számolására alkalmas módszereinket mutatjuk be nagy vonalakban. Ezek a módszerek teljes részletességben már korábbi cikkünkben [1] is bemutatásra kerültek angol nyelven. Ezután rátérünk arra, hogy módszereinket azóta milyen módon fejlesztettük tovább, bővítettük ki.

3.1 Angol szavak szemantikai hasonlóságának kiszámítása

A szavakhoz képzett vektorokban szereplő tulajdonságok kinyerésére két fő változatot használtunk. Az első a szózsák (bag-of-words) alapú megközelítés. Ez a vizsgált szó összes előfordulási helyét megkeresi a felhasznált korpuszban, és az előfordulások környezetében lévő minden, legfeljebb három távolságra szereplő szót belerakja a tulajdonságvektorba, egy távolságalapú súlyozást felhasználva. A másik módja a tulajdonságok kinyerésének a nyelvtani kapcsolatok felhasználása. Ehhez először a korpuszt automatikusan elemeztük a C&C CCG parser [7] segítségével, majd tulajdonságként a vizsgált szóhoz nyelvtanilag közvetlenül kapcsolódó szavakat használtuk a nyelvtani kapcsolatok típusával együtt. Mindkét módszerhez három korpuszt, a British National Corpust (BNC), a Web 1T 5-gram Corpust (csak a 4- és 5-gramokat) és az angol Wikipedia korpuszát használtuk (a Wikipedia korpuszt előfeldolgoztuk Rafael Mudge wikipedia2text_rsm_mods toolkitjével[2]). Mivel tetszőleges korpusz alkalmazható a tulajdonságok kinyeréséhez, ezért a módszereink könnyen adaptálhatók más tárgykörökre és más nyelvekre.

[2] http://blog.afterthedeadline.com/2009/12/04/generating-a-plain-text-corpus-from-wikipedia/

1. táblázat: Módszereink eredménye az angol Miller-Charles adathalmazon (Spearman korreláció). Jelölések: bnc/enwiki/web1t5gram jelöli a korpuszt; bagofwords/parsed jelöli a tulajdonságtípusokat (szózsák vagy nyelvtani kapcsolatok); lin/num jelöli a tulajdonságvektorok létrehozásának és összehasonlításának módszert (Lin [8] módszerén alapuló vagy numerikus vektorokat alkalmazó); cos/dice/pears/spear/zkl jelöli a hasonlósági mértéket; freq/logfreq/ pmi/loglh/qw/pw/rapp jelöli a súlyozást; + jelöli két módszer kombinációját.

Módszer	**Eredmény**
enwiki-parsed-num-zkl-loglh+bnc-bagofwords-num-zkl-loglh	0,773
bnc-bagofwords-num-cos-qw+enwiki-parsed-num-cos-freq	0,773
enwiki-parsed-num-zkl-loglh+enwiki-parsed-num-pears-logfreq	0,754
bnc-bagofwords-num-cos-qw+enwiki-parsed-num-cos-qw	0,750
bnc-bagofwords-num-zkl-loglh	0,744
bnc-parsed-num-pears-qw+enwiki-bagofwords-num-cos-pmi	0,737
bnc-bagofwords-num-zkl-loglh+enwiki-parsed-num-pears-pmi	0,736
bnc-bagofwords-num-zkl-loglh+enwiki-parsed-num-cos-pmi	0,736
enwiki-parsed-num-pears-pmi+enwiki-bagofwords-num-pears-pmi	0,729
enwiki-parsed-num-pears-pmi	0,727
enwiki-parsed-num-cos-pmi	0,727
enwiki-parsed-num-zkl-loglh+enwiki-bagofwords-num-pears-pmi	0,721
enwiki-parsed-num-zkl-loglh+enwiki-bagofwords-num-cos-pmi	0,721
enwiki-parsed-num-zkl-loglh	0,718
bnc-parsed-num-pears-loglh+enwiki-parsed-num-pears-pmi	0,712
bnc-parsed-num-cos-loglh+enwiki-parsed-num-cos-pmi	0,712
enwiki-bagofwords-num-spear-logfreq+enwiki-parsed-num-cos-pmi	0,703
enwiki-bagofwords-num-pears-pmi	0,684
enwiki-bagofwords-num-zkl-loglh	0,548

A tulajdonságvektorok létrehozására és összehasonlítására szintén két különféle szemléletmódot tekintettünk. Először Lin [8] módszerét (azt, amelyik statikus korpuszokkal dolgozik és nem használja fel a WordNetet) újraimplementáltuk néhány módosítással. Ez a módszer bináris tulajdonságvektorokkal dolgozik, melyeket egy Lin [8] által definiált mértékkel hasonlít össze. A másik szemlélet numerikus tulajdonságvektorokkal dolgozik, ahol minden tulajdonsághoz egy súly is tartozik. A súlyok közt szerepelnek egyszerű gyakoriságalapú (gyakoriság - freq, gyakoriság logaritmusa - logfreq), illetve bonyolultabb információelméleti súlyok (pontonkénti kölcsönös információ - pmi, log-likelihood arány - loglh, qw, pw, Rapp-féle [4] - rapp) is. Ez a modell a súlyozott vektorokat különféle vektorhasonlósági mértékekkel (koszinusz hasonlóság - cos, Lin-féle Dice-együttható [8] - dice) hasonlítja össze.

Mivel sok szó többféle szófajt is felvehet, és a különböző szófajú szavakhoz különböző tulajdonságok a fontosak, ezért szavak összehasonlításakor fontos az, hogy

először a szavak szófaját meghatározzuk. Ez módszerünk esetében a tesztszavaknak az adott korpuszban vett előfordulási gyakoriságának felhasználásával történik [1].

Azért, hogy a különféle módszerek előnyeit egyesíteni tudjuk, a módszereket nem csak külön-külön, hanem egymással kombinálva is teszteltük. Két módszer kombinációjakor a szópárok hasonlósága először a két módszerrel külön kerül meghatározásra, majd a kombinált hasonlóság e két hasonlósági pontszámból kerül kiszámításra [1].

3.2 A továbbfejlesztett módszer

Az előző alfejezetben ismertetett módszereinken két fő változtatást hajtottunk végre. Egyrészt a már meglévő három vektorhasonlósági módszer (lin, cos, dice) mellé további három hasonlósági metrikát implementáltunk. Az első a Pearson-féle korrelációs együttható (pears), mely két numerikus változó közti összefüggés erősségét mutatja meg. A másik a Spearman-féle rangkorrelációs együttható (spear), mely a Pearson-együttható olyan speciális esete, ami a numerikus értékek helyett azok rangjával számol. A harmadik megvalósított metrika a Zero-KL metrika [9] inverze (zkl). A Zero-KL metrika a Kullback-Leibler divergencia olyan módosítása, mely már 0 valószínűséget tartalmazó valószínűségi eloszlásokra is értelmezett. Mivel a Zero-KL annál nagyobb értéket vesz fel, minél kevésbé hasonló két valószínűségi eloszlás, és mivel a többi hasonlósági mértékünk pont fordítva működik, ezért mi az inverzét alkalmaztuk.

Az új hasonlósági mértékek alkalmazása mellett még egy nagyon lényeges részét fejlesztettük tovább az algoritmusainknak. Módszereink eddig pusztán angol szavak közötti szemantikai hasonlóság kiszámítására voltak képesek. A továbbfejlesztett változatok már képesek magyar szavak közötti szemantikus hasonlóság automatikus kiszámítására is, melyre legjobb tudásunk szerint jelenleg egyetlen másik módszer sem képes. Magyar tesztszavak esetén módszereink az összehasonlítást pillanatnyilag csak a szózsák modell alapján végzik, vagyis minden tesztszóhoz megkeresik a felhasznált (magyar nyelvű) korpuszban a szó előfordulási helyeit, és az ott talált környezeti szavakat használják fel tulajdonságként, a nyelvtani kapcsolatok figyelembe vétele nélkül. Korpuszként a magyar Wikipédia korpuszát használtuk fel (szintén előfeldolgoztuk Rafael Mudge wikipedia2text_rsm_mods toolkitjével). A jövőben majd szeretnénk megvalósítani a nyelvtani kapcsolatokat alkalmazó modellt is.

4 Eredmények

Az elkészült módszereket mind angol, mind magyar tesztadatbázisokon kiértékeltük. Angol szavak esetén két gyakran alkalmazott adathalmazt használtunk fel. Az első a 30 szópárból álló Miller-Charles adathalmaz (MC), melynél minden szópárhoz 38 egyetemi hallgató rendelt hasonlósági pontszámot. Mivel a korábbi WordNet-verziók nem tartalmaztak két szót e szavakból, ezért rendszerint csak a maradék 28 szópárt használták fel a kiértékelésben, és mi is így tettünk. A másik adathalmaz a 80 kérdésből álló TOEFL szinonimakérdések halmaza, ahol minden kérdés egy tesztszót és négy lehetséges megoldást tartalmaz, a feladat pedig annak eldöntése, hogy melyik szó a leghasonlóbb a tesztszóhoz. A kiértékelési metrika az MC adathalmaz esetén az átlagos pontszámokkal vett Spearman-korreláció, míg a TOEFL adathalmaz esetén a helyes válaszok százaléka volt.

2. táblázat: Módszereink eredménye az angol TOEFL-kérdéseken (helyes válaszok százaléka).

Módszer	**Eredmény**
bnc-parsed-num-pears-loglh+enwiki-parsed-num-pears-pmi	88,75%
bnc-parsed-num-cos-loglh+enwiki-parsed-num-cos-pmi	88,75%
enwiki-parsed-num-pears-pmi+enwiki-bagofwords-num-pears-pmi	87,50%
enwiki-parsed-num-zkl-loglh+enwiki-bagofwords-num-pears-pmi	87,50%
enwiki-parsed-num-zkl-loglh+enwiki-bagofwords-num-cos-pmi	87,50%
bnc-parsed-num-pears-qw+enwiki-bagofwords-num-cos-pmi	86,25%
bnc-bagofwords-num-zkl-loglh+enwiki-parsed-num-pears-pmi	85,00%
bnc-bagofwords-num-zkl-loglh+enwiki-parsed-num-cos-pmi	85,00%
enwiki-parsed-num-zkl-loglh+enwiki-parsed-num-pears-logfreq	83,75%
enwiki-bagofwords-num-pears-pmi	83,75%
enwiki-parsed-num-pears-pmi	82,50%
enwiki-parsed-num-cos-pmi	82,50%
enwiki-bagofwords-num-spear-logfreq+enwiki-parsed-num-cos-pmi	82,50%
enwiki-bagofwords-num-zkl-loglh	81,25%
enwiki-parsed-num-zkl-loglh	80,00%
enwiki-parsed-num-zkl-loglh+bnc-bagofwords-num-zkl-loglh	80,00%
bnc-bagofwords-num-cos-qw+enwiki-parsed-num-cos-qw	77,50%
bnc-bagofwords-num-zkl-loglh	72,50%
bnc-bagofwords-num-cos-qw+enwiki-parsed-num-cos-freq	72,50%

Mivel magyar szavakra tudomásunk szerint nem létezik még olyan algoritmus, mely szavak szemantikai hasonlóságának megállapítására képes, ezért még nincs általánosan használt tesztadatbázis sem a kiértékeléshez. Ennek hiányában arra a következtetésre jutottunk, hogy legegyszerűbben oly módon tudjuk módszereinket kiértékelni, hogy az angol szavakat tartalmazó két tesztadatbázist lefordítjuk magyarra. Ugyan tudjuk, hogy a legtöbb angol szóhoz nem létezik olyan magyar szó, mely pontosan ugyanazzal a jelentéskörrel rendelkezik, mégis úgy gondoljuk, hogy kezdeti kiértékelésre megfelelőek ezek az adatbázisok, és hogy segítségükkel algoritmusaink teljesítménye jól becsülhető. Így végül magyarra az MC adathalmaz magyar fordítását (MC-Hu), illetve a TOEFL adathalmaz magyar fordítását (TOEFL-Hu) használtuk fel, az angollal megegyező kiértékelési metrikák használatával. A fordításnál igyekeztünk, hogy a magyar tesztek minél jobban tükrözzék angol verzióik tulajdonságait.

Az algoritmusok angol tesztszavakon adott eredményeit az 1. és 2. táblázat foglalják össze. Az algoritmusaink által elért legjobb eredmény az MC adathalmaz esetén 0,773, míg a TOEFL-kérdések esetén 88,75% volt. Ha összehasonlítjuk az új vektorhasonlóságokat alkalmazó módszerek eredményét a régiekével, akkor jól látható, hogy az újabb verziók hasonlóan jó eredményt értek el, mint korábbi társaik, sőt néhol a korábbiaknál jobbat. A legtöbb olyan algoritmus, mely jól teljesített az egyik adat-

halmazon, az jó eredményt ért el a másikon is. Néhányat kiemeltünk azok közül, melyek a két adathalmazt együttesen figyelembe véve a legjobb eredményt érték el:

a. enwiki-parsed-num-zkl-loglh+bnc-bagofwords-num-zkl-loglh:
   (MC: 0,773, TOEFL: 80,00%)
b. enwiki-parsed-num-zkl-loglh+enwiki-parsed-num-pears-logfreq:
   (MC: 0,754, TOEFL: 83,75%)
c. bnc-parsed-num-pears-loglh+enwiki-parsed-num-pears-pmi:
   (MC: 0,712, TOEFL: 88,75%)
d. enwiki-parsed-num-pears-pmi+enwiki-bagofwords-num-pears-pmi:
   (MC: 0,729, TOEFL: 87,50%)
e. bnc-parsed-num-pears-qw+enwiki-bagofwords-num-cos-pmi:
   (MC: 0,737, TOEFL: 86,25%)

3. táblázat: Eredményeink összehasonlítása más módszerek eredményeivel az angol Miller-Charles adathalmazon (Spearman-korreláció).

Módszer	**Eredmény**	**Felhasznált adatforrások**
Emberi felső korlát [11]	0,934	
Agirre et al. [10]	0,92	WordNet, korpusz
Patwardhan és Pedersen [12]	0,91	WordNet
Jarmasz és Szpakowicz [13]	0,87	Roget's Thesaurus
Tsatsaronis et al. [2]	0,856	WordNet
Kulkarni és Caragea [14]	0,835	Webes keresés
Lin [8]	0,82	WordNet, korpusz
Resnik [11]	0,81	WordNet, korpusz
enwiki-parsed-num-zkl-loglh+ bnc-bagofwords-num-zkl-loglh	0,773	korpusz
enwiki-parsed-num-zkl-loglh+ enwiki-parsed-num-pears-logfreq	0,754	korpusz
bnc-parsed-num-pears-qw+ enwiki-bagofwords-num-cos-pmi	0,737	korpusz
bnc-bagofwords-num-zkl-loglh+ enwiki-parsed-num-pears-pmi	0,736	korpusz
enwiki-parsed-num-pears-pmi+ enwiki-bagofwords-num-pears-pmi	0,729	korpusz
enwiki-parsed-num-pears-pmi	0,727	korpusz
Gabrilovich és Markovitch [15]	0,72	korpusz
bnc-parsed-num-pears-loglh+ enwiki-parsed-num-pears-pmi	0,712	korpusz
Milne és Witten [16]	0,70	Wikipedia linkek, Webes keresés
Sahami és Heilman [17]	0,618	Webes keresés

Eredményeinket mások módszereivel a 3. és 4. táblázatban hasonlítottuk össze. Ez azt mutatja, hogy módszereink az MC adathalmazon általában közepes eredményt értek el, míg a TOEFL adathalmazon összességében harmadik legjobban teljesítettek. Azonban, ha csak azokat a módszereket tekintjük, melyek a mi módszereinkhez hasonlóan csak statikus korpuszokat használnak fel adatforrásként, akkor több módszerünk is (például d. és e.) az MC és a TOEFL adathalmazon rendre első és második legjobb eredményt ért el más kutatások eredményeihez hasonlítva.

Az 5. és 6. táblázat foglalja össze algoritmusaink eredményét a magyar tesztadatbázisokon. Az MC-Hu adatbázis esetén elért legjobb eredmény 0,637, míg a TOEFL-Hu kérdések esetén 60,00%. Ebben az esetben azonban korábbi eredmények hiányában nem tudjuk eredményeinket másokéval összehasonlítani. Viszont, ha ezeket az eredményeket az angol tesztadatbázisokon elért eredményekkel vetjük össze, akkor az figyelhető meg, hogy magyar tesztszavakon átlagosan lényegesen rosszabb eredményt értek el, mint az angol tesztek esetén. Véleményünk szerint ez több tényezőnek tudható be. Egyrészt a magyar nyelv nyelvtana lényegesen bonyolultabb az angolénál. Másrészt a felhasznált magyar korpusz mérete lényegesen kisebb az alkalmazott angol korpuszokénál. Harmadrészt, mivel a magyar nyelv szabad szórendű, ezért a nyelvtani kapcsolatok sokkal több információval szolgálnának egy szóról, mint a környezeti szavak. Tehát a nyelvtani kapcsolatokat is felhasználó modell véleményünk szerint az eddigieknél jobb eredményeket érhetne el.

A magyar nyelv esetén is azok az algoritmusok, melyek az egyik adathalmazon jól teljesítettek, általában jó eredményt értek el a másikon is. A következő algoritmusok teljesítettek legjobban mindkettő adatbázist figyelembe véve:

f. huwiki-bagofwords-num-zkl-loglh+huwiki-bagofwords-num-pears-pmi:
(MC: 0,637, TOEFL: 58,75%)
g. huwiki-bagofwords-num-zkl-pmi+huwiki-bagofwords-num-pears-pmi:
(MC: 0,629, TOEFL: 57,50%)
h. huwiki-bagofwords-num-zkl-loglh:
(MC: 0,622, TOEFL: 60,00%)

Megvizsgáltuk azt is, hogy melyek azok a módszerek, melyek a felhasznált korpusztól és a nyelvtől függetlenül jól teljesítenek. Mivel a különböző nyelvekhez más korpuszok tartoznak, ezért a korpuszokat sem vettük figyelembe. Az találtuk, hogy mind kombinált, mind különálló módszerből létezik számos olyan, mely jól teljesít mindkét nyelv mindkét tesztadatbázisa esetén, vagyis nyelvtől és tesztadatbázistól függetlenül jól tud működni. Ezek közül néhány:

i. num-zkl-loglh+num-pears-pmi:
(MC: 0,736, TOEFL: 87,50%, MC-Hu: 0,637, TOEFL-Hu: 58,75%)
j. num-zkl-loglh+num-cos-pmi:
(MC: 0,736, TOEFL: 87,50%, MC-Hu: 0,611, TOEFL-Hu: 58,75%)
k. num-zkl-loglh:
(MC: 0,744, TOEFL: 81,25%, MC-Hu: 0,622, TOEFL-Hu: 60,00%)
l. num-pears-pmi:
(MC: 0,727, TOEFL: 83,75%, MC-Hu: 0,617, TOEFL-Hu: 58,75%)

A felsorolt négy algoritmus mindegyike jól teljesít mind a négy tesztet tekintve. Ha csak azokat az algoritmusokat vesszük figyelembe, amelyek kizárólag statikus korpuszokat használnak fel adatforrásként, akkor az i. és j. algoritmus által elért eredmények például az MC és TOEFL adathalmazon tesztelve rendre az első és második legjobbak más kutatások eredményeihez hasonlítva, továbbá az MC-Hu és TOEFL-Hu adathalmazokon elért eredményeik is saját módszereink eredményeit tekintve a legjobbak között vannak.

4. táblázat: Eredményeink összehasonlítása más módszerek eredményeivel az angol TOEFL kérdéseken (helyes válaszok százaléka).

Módszer	**Eredmény**	**Felhasznált adatforrások**
Turney et al. [5]	97,5%	Webes keresés, fogalomtár
Rapp [4]	92,5%	korpusz
bnc-parsed-num-pears-loglh+ enwiki-parsed-num-pears-pmi	88,75%	korpusz
enwiki-parsed-num-pears-pmi+ enwiki-bagofwords-num-pears-pmi	87,50%	korpusz
enwiki-parsed-num-zkl-loglh+ enwiki-bagofwords-num-pears-pmi	87,50%	korpusz
Tsatsaronis et al. [2]	87,5%	WordNet
bnc-parsed-num-pears-qw+ enwiki-bagofwords-num-cos-pmi	86,25%	korpusz
Matveeva et al. [18]	86,25%	korpusz
enwiki-parsed-num-zkl-loglh+ enwiki-parsed-num-pears-logfreq	83,75%	korpusz
enwiki-parsed-num-pears-pmi	82,50%	korpusz
Higgins [3]	81,3%	Webes keresés
enwiki-parsed-num-zkl-loglh+ bnc-bagofwords-num-zkl-loglh	80,00%	korpusz
Jarmasz és Szpakowicz [13]	78,7%	Roget's Thesaurus
Átlagos nem angol anyanyelvű, amerikai egyetemre felvételiző diák [6]	64,5%	
Landauer és Dumais [6]	64,3%	korpusz
Lin [8]	24,0%	WordNet, korpusz
Resnik [11]	20,3%	WordNet, korpusz

5. táblázat: Módszereink eredménye a magyar Miller-Charles adathalmazon (Spearman-korreláció).

Módszer	**Eredmény**
huwiki-bagofwords-num-zkl-loglh+ huwiki-bagofwords-num-pears-pmi	0,637
huwiki-bagofwords-num-zkl-pmi+ huwiki-bagofwords-num-pears-pmi	0,629
huwiki-bagofwords-num-zkl-loglh	0,622
huwiki-bagofwords-num-zkl-logfreq+ huwiki-bagofwords-num-pears-pmi	0,621
huwiki-bagofwords-num-pears-pmi	0,617
huwiki-bagofwords-num-zkl-loglh+ huwiki-bagofwords-num-cos-pmi	0,611
huwiki-bagofwords-num-cos-pmi	0,610
huwiki-bagofwords-num-pears-pmi+ huwiki-bagofwords-num-cos-freq	0,588

6. táblázat: Módszereink eredménye a magyar TOEFL-kérdéseken (helyes válaszok százaléka).

Módszer	**Eredmény**
huwiki-bagofwords-num-zkl-loglh	60,00%
huwiki-bagofwords-num-pears-pmi+ huwiki-bagofwords-num-cos-freq	60,00%
huwiki-bagofwords-num-pears-pmi	58,75%
huwiki-bagofwords-num-zkl-logfreq+ huwiki-bagofwords-num-pears-pmi	58,75%
huwiki-bagofwords-num-zkl-loglh+ huwiki-bagofwords-num-pears-pmi	58,75%
huwiki-bagofwords-num-zkl-loglh+ huwiki-bagofwords-num-cos-pmi	58,75%
huwiki-bagofwords-num-zkl-pmi+ huwiki-bagofwords-num-pears-pmi	57,50%
huwiki-bagofwords-num-cos-pmi	57,50%

5 Konklúzió

Cikkünkben olyan módszereket mutattunk be, melyek alkalmasak magyar és angol szavak közötti szemantikai hasonlóság automatikus megállapítására. Ezek statikus korpuszokból kinyert statisztikai információk alapján egy tulajdonságvektort képeznek minden szóhoz, majd a szavak hasonlóságát vektoraik hasonlóságaként számolják ki. Több variációt kipróbáltunk, melyek különféle tulajdonságtípusokat, vektortípuso-

kat, súlyozásokat, valamint vektorhasonlósági mértéket alkalmaznak, továbbá a különálló módszerek kombinációit is teszteltük.

Minden módszert nyelvenként két különböző adathalmazon értékeltünk ki, angol esetén a Miller-Charles adathalmazon (MC) és a TOEFL szinonimakérdéseken, magyar esetén pedig ezek magyarra fordított változatán (MC-Hu és TOEFL-Hu). Angol szavak esetén legjobb módszereink közepes eredményt értek el az MC adathalmazon, míg harmadik legjobban teljesítettek a TOEFL-kérdéseken. Azonban, ha kizárólag azokat a módszereket tekintjük, melyek csak statikus korpuszokat alkalmaznak, akkor algoritmusaink a két adathalmazon rendre első és második legjobb eredményt értek el.

Az algoritmusok angol tesztszavakon lényegesen jobb eredményt értek el, mint magyar változataikon. Ezt részben annak tudjuk be, hogy a magyar nyelv nyelvtana lényegesen bonyolultabb az angolénál és hogy a felhasznált magyar korpusz mérete lényegesen kisebb az alkalmazott angol korpuszokénál. Továbbá, mivel a magyar nyelv szabad szórendű, ezért a nyelvtani kapcsolatok sokkal több információval szolgálnának egy szóról, mint az általunk jelenleg használt környezeti szavak. Ezért véleményünk szerint a nyelvtani kapcsolatokat is felhasználó modell az eddigieknél lényegesen jobb eredményeket érhetne el.

Az eredmények alapján úgy gondoljuk, hogy módszereink sikeresen alkalmazhatóak lennének valós problémákon is. Megfigyelhető, hogy az algoritmusok (főként az angol nyelv esetén) jobb eredményt érnek el a TOEFL-kérdéseken, mint a MC adathalmazon. Ez azt sugallja, hogy alkalmasabbak arra, hogy egy tesztszóhoz kiválasszák a leghasonlóbb szót egy listából, mint arra, hogy két szó pontos hasonlóságát megállapítsák.

Úgy gondoljuk, hogy a jövőben érdemes lenne módszereinket további, még nagyobb korpuszok segítségével kipróbálni, különösen a magyar verzió esetén (például Agirre et al. [10] egy 1,6 Terawordös angol korpuszt használtak, és algoritmusukat 2000 CPU magon futtatták). Továbbá mindenképpen szeretnénk a nyelvtani kapcsolatokat is alkalmazó modellt magyar nyelvre is implementálni, amivel reményeink szerint eredményeinket tovább tudnánk javítani. Ezen felül úgy véljük, mint azt a 2. fejezetben is említettük, hogy különböző típusú módszerek kombinálásával azok előnyeit ötvözhetjük. Ezért véleményünk szerint még jobb eredményeket tudnánk elérni, ha módszereinket kombinálnánk más, webes kereséseket vagy lexikális adatbázisokat felhasználó módszerekkel.

Hivatkozások

1. Dobó, A., Csirik, J.: Computing Semantic Similarity Using Large Static Corpora. In: van Emde Boas, P. et al. (eds.): SOFSEM 2013. LNCS, Vol. 7741. Springer, Heidelberg (2013, forthcoming) 491–502
2. Tsatsaronis, G., Varlamis, I., Vazirgiannis, M.: Text Relatedness Based on a Word Thesaurus. Journal of Artificial Intelligence Research, Vol. 37 (2010) 1–39
3. Higgins, D.: Which Statistics Reflect Semantics? Rethinking Synonymy and Word Similarity. In: Kepser, S., Reis, M. (eds.): Linguistic Evidence: Empirical, Theoretical and Computational Perspectives. Mouton de Gruyter, Berlin, New York (2005) 265–284
4. Rapp, R.: Word Sense Discovery Based on Sense Descriptor Dissimilarity. In: 9th Machine Translation Summit. Association for Machine Translation in the Americas, Stroudsburg (2003) 315–322

5. Turney, P.D., Littman, M.L., Bigham, J., Shnayder, V.: Combining Independent Modules to Solve Multiple-choice Synonym and Analogy Problems. In: 4th Conference on Recent Advances in Natural Language Processing. John Benjamins Publishers, Amsterdam (2003) 482–489
6. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, Vol. 104 (1997) 211–240
7. Clark, S., Curran, J.R.: Parsing the WSJ using CCG and log-linear models. In: 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg (2004) 103–110
8. Lin, D.: An information-theoretic definition of similarity. In: 15th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco (1998) 296–304
9. Hughes, T., Ramage, D.: Lexical Semantic Relatedness with Random Graph Walks. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007) 581–589
10. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A.: A study on similarity and relatedness using distributional and WordNet-based approaches. In: 10th Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies. Association for Computational Linguistics, Stroudsburg (2009) 19–27
11. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: 14th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco (1995) 448–453
12. Patwardhan, S., Pedersen, T.: Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In: 11th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg (2006) 1–8
13. Jarmasz, M., Szpakowicz, S.: Roget's Thesaurus and Semantic Similarity. In: 4th Conference on Recent Advances in Natural Language Processing. John Benjamins Publishers, Amsterdam (2003) 212–219
14. Kulkarni, S., Caragea, D.: Computation of the Semantic Relatedness between Words using Concept Clouds. In: International Conference on Knowledge Discovery and Information Retrieval. INSTICC Press, Setubal (2009) 183–188
15. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: 20th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco (2007) 1606–1611
16. Milne, D., Witten, I.H.: An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In: 23rd AAAI Conference on Artificial Intelligence. AAAI Press, Menlo Park (2008) 25–30
17. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: 15th international conference on World Wide Web. ACM Press, New York (2006) 377–386
18. Matveeva, I., Levow, G.-A., Farahat, A., Royer, C.: Term Representation with Generalized Latent Semantic Analysis. In: 5th Conference on Recent Advances in Natural Language Processing. John Benjamins Publishers, Amsterdam (2005) 45–54