Eloszlás alapú szemantikai modellek tulajdonságvektorainak készítése és összehasonlítása során használt paraméterek átfogó elemzése több nyelvre

Dobó András

Témavezető
Prof. Dr. Csirik János

Informatika Doktori Iskola
Természettudományi és Informatikai Kar
Szegedi Tudományegyetem

A doktori értekezés összefoglalója

Szeged
2019

Bevezetés

Számos számítógépes nyelvészeti (NLP) problémához, többek között információ visszakereséshez (Hliaoutakis et al., 2006), helyesírás-javításhoz (Budanitsky and Hirst, 2001) és összetett szó értelmezéshez (Dobó and Pulman, 2011), fontos hogy meg tudjuk határozni szavak szemantikai hasonlóságának vagy kapcsolatának mértékét. Míg a szemantikai kapcsolat számos, szavak között fennálló relációt (többek között a hasonlóságot is) számításba vesz, addig a szemantikai hasonlóság csak a szavak által jelölt fogalmak tényleges egyformaságát veszi figyelembe. Ezáltal a hasonlóságból következik a kapcsolat, de ez fordítva nem igaz. Például, a "bicikli" és a "motorkerékpár" szavak hasonlóak, mivel mindkettő kétkerekű járművet jelöl, így kapcsolódnak is egymáshoz. Ezzel szemben a "postás" és a "levél" szavak közeli kapcsolatban állnak, mivel általában a postás kézbesíti a leveleket, de mégsem hasonlítanak egymásra, mert meglehetősen különböző fogalmakat jelölnek. Továbbá, a "kemence" és a "hajóút" szavak egyáltalán nem hasonlítanak egymásra és nem is kapcsolódnak egymáshoz.

Motiváció

A legtöbb modell a jelentés eloszlási hipotézisére (Harris, 1954) alapszik, és ezáltal a szemantikai hasonlóság vagy kapcsolat mértékét nagyméretű korpuszból kinyert eloszlási adatok alapján számolja. Ezeket a modelleket gyűjtőnévvel eloszlás alapú szemantikai modelleknek (DSM) szokás nevezni (Baroni and Lenci, 2010; Baroni et al., 2014). Ezekben a modellekben először a lehetséges tulajdonságok kerülnek megállapításra, általában szövegkörnyezeti szavak formájában, ami után a modellek súlyokat rendelnek minden szó-tulajdonság párhoz komplex módszerek segítségével, ezáltal tulajdonság-vektorokat készítve minden szóhoz. A szavak szemantikai hasonlóságának vagy kapcsolatának a mértékét ezt követően a szavak tulajdonság-vektorainak az összehasonlításával számítják ki. Habár a DSM-ek számos lehetséges paraméterrel rendelkeznek, e paraméterek igazán átfogó elemzése, ami a paraméterek egymástól való függését is figyelembe veszi, még hiányzik és szükséges lenne, mint ahogy azt Levy et al. (2015) is sugallja.

A legtöbb DSM-mel foglalkozó kutatás a problémának csak egy vagy két aspektusára fókuszál, és a modell többi paraméterét adottnak veszi valamilyen standard beállítással. Például, a kutatások nagy része megszokásból egyszerűen koszinuszt használ vektorhasonlósági mértékként (pl. Bruni et al., 2013; Baroni et al., 2014; Speer et al., 2017; Salle et al., 2018) és/vagy (pozitív) pontonkénti kölcsönös információt súlyozási sémaként (pl. Islam and Inkpen, 2008; Hill et al., 2014; Salle et al., 2018). És még a figyelembe vett paraméterek esetén is általában csak néhány lehetséges

beállítást tesztelnek. Továbbá, vannak olyan paraméterek is, amiket a legtöbb tanulmány teljesen figyelmen kívül hagy, és nem is lettek még igazán elemezve a múltban, még külön-külön sem (pl. simítás, vektor-normalizáció vagy a tulajdonságok gyakoriságára minimum limit). Sőt mi több, mivel ezek a paraméterek nagyban befolyásolni tudják egymást, a külön-külön, egyenkénti elemzésük nem is elegendő, mivel az nem veszi figyelembe azok egymásra hatását.

Van néhány olyan kutatás ami több paramétert is tesztel többfajta lehetséges beállítással, mint például Lapesa and Evert (2014) és Kiela and Clark (2014), de ezek is messze vannak attól, hogy igazán átfogó képet adjanak, és szintén nem tesztelik teljes mértékben a különféle paraméterek között fellépő kölcsönhatásokat. Tehát, habár fontos lenne a paramétereket és azok kombinációját részletesen kielemezni, mint ahogy azt Levy et al. (2015) is megemlíti, még mindig nem létezik ezeknek igazán átfogó tanulmánya. Továbbá, annak ellenére, hogy a legjobb paraméterbeállítások a különféle nyelvek esetén különbözőek lehetnek, a tanulmányok döntő többsége általában pusztán egyetlen nyelvvel foglalkozik (legtöbbször az angollal), vagy figyelembe vesz több nyelvet is, de a konklúziók nyelvek közötti részletes összehasonlítása nélkül. Ebben az értekezésben ezeket a kutatási hiányokat szeretnénk betölteni.

Feladat és célkitűzés

A DSM-ek rendszerint két egymástól különálló fázissal rendelkeznek. Az első fázisban statisztikai információt (pl. nyers gyakoriságokat) nyernek ki nyers adatokból (pl. egy nagyméretű nyers szöveges korpuszból), statisztikai eloszlási adatok formájában. A második fázisban tulajdonságvektorokat készítenek a kinyert információból minden szóhoz, majd ezeket a vektorokat hasonlítják egymáshoz a szavak hasonlósági vagy kapcsolati mértékének a megállapításához. Mi a kutatásunk során az első fázisban kinyert információt már adottnak vesszük, és egy szisztematikus, párhuzamos elemzését végezzük el a tulajdonságvektorok készítése és összehasonlítása során használt tulajdonságoknak, miközben a tulajdonságok egymásra hatását is figyelembe vesszük.

Azért döntöttünk úgy, hogy csak a DSM-ek második fázisát elemezzük, mivel a két fázis egymástól meglehetősen különálló és független, és a második fázis minden egyes lehetséges paraméter-érték kombinációjának tesztelése már így is lehetetlen a lehetséges kombinációk óriási száma miatt. Ezért egy teljes analízis helyett már így is egy heurisztikus módszert kellett alkalmaztunk. Tehát ezen felül még az első fázis különféle paramétereit (pl. használt korpusz, szövegkörnyezeti típus (ablak-alapú vagy dependencia-alapú) és szövegkörnyezeti méret) is tesztelni ésszerűtlennek és megvalósíthatatlannak tűnt, és így e kutatás hatókörén kívülre esett. Ezért ennek a fázisnak a vizsgálatát teljes egészében kihagytuk, egy kivétellel.

A statikus korpuszokból kinyert információkon alapuló DSM-eknek két jelentős csoportja van az első fázisuk alapján: gyakorisági-vektor-alapú (CVBM) és prediktív modellek (PM; más névvel szóbeágyazási modellek) (Baroni et al., 2014). A prediktív modellek elmúlt évekbeli nagy népszerűsége miatt, továbbá azért, hogy még teljesebb képet kapjunk, a gyakorisági-vektor-alapú modellek által korpuszokból kinyert információk mellett elvégeztünk néhány kísérletet prediktív modellek által kinyert információkkal is az angol nyelv esetén. Továbbá, a későbbiekben még kiegészítettük az elemzésünket egy olyan modellel is, ami tudás-gráfból kinyert szemantikai vektorokon alapszik. A megérzésünk az volt, hogy lesz egy olyan konfiguráció ami a legjobb eredményt fogja elérni mindhárom típusú modell esetén. Azt azonban meg kell jegyezzük, hogy a prediktív modellek esetén a figyelembe vett paramétereknek csak egy részét lehetett tesztelni e modellek jellegzetességei miatt. Részben ezért is fókuszáltunk inkább a gyakorisági-vektor-alapú modellekre.

A kutatásunk során összességében 10 fontos paramétert azonosítottunk a gya-korisági-vektor-alapú DSM-ek második fázisában, mint például a vektorhasonlósági mértéket, a súlyozási sémát, a tulajdonság-transzformációt, a simítást és a dimenzió-csökkentést. Ezek közül azonban összesen 4 érhető el prediktív illetve tudás-gráf-alapú szemantikai vektorok használata esetén, mivel ilyen inputok használatakor a nyers gyakoriságok már nem érhetőek el, a súlyozott vektorok már elkészültek és általában már a dimenzió-csökkentés is végrehajtásra került rajtuk.

Elemzésünk során e paramétereket párhuzamosan értékeltük ki számos beállítással annak érdekében, hogy megtaláljuk a legjobb konfigurációt, amit a lehető legmagasabb pontszámokat éri el a standard tesztadatbázisokon. Az átfogó elemzésünket angolra, spanyolra és magyarra különkülön is megcsináltuk, majd a különböző nyelvek esetén levont konklúziókat összehasonlítottuk.

Néhány paraméterre nagy mennyiségű, akár több ezer lehetséges beállítást is teszteltünk, ami több milliárd lehetséges paraméter-beállítási kombinációt eredményezett. Amellett, hogy természetesen minden paraméter konvencionálisan alkalmazott beállítását is teszteltük, számos új variánst javasoltunk mi is. Továbbá, számos új konfigurációt teszteltünk, amik közül némelyek az általánosan használt, standard konfigurációknál messze jobb eredményt érnek el, és az eddig ismert legjobb konfigurációknál is jobb eredményeket érnek el.

Első körben az elemzésünket angolra végeztük el, és az eredményeket azon elemeztük ki részletesen (Dobó and Csirik, 2019a). Ezt követően megismételtük ugyanezt az elemzést, néhány paraméter esetén bővített beállítási opciókkal, angolra, spanyolra és magyarra is, és a különböző nyelvekre levont konklúziókat összevetettük egymással (Dobó and Csirik, 2019b).

Konklúziók

Az értekezésben az eloszlás alapú szemantikai modellek tulajdonságvektorainak készítése és összehasonlítása során használt paramétereknek egy nagyon részletes és szisztematikus elemzését prezentáltuk angolra, spanyolra és magyarra, amivel egy komoly kutatási hiányt töltöttünk be. Gyakorisági-vektor-alapú modellek esetén 10, míg prediktív és tudás-gráf-alapú modellek esetén 4 fontos paramétert azonosítottunk, és ezek mindegyikéhez számos beállítást teszteltünk. Az elemzésünk során teszteltünk új paramétereket és új paraméter-beállításokat, továbbá minden paramétert párhuzamosan vizsgáltunk, ezáltal ezek esetleges egymásra hatását is figyelembe véve. Tudomásunk szerint mi voltunk az elsők, akik e paraméterek ilyen részletes elemzését elvégezték, és szintén mi voltunk az elsők, akik a különböző nyelvek esetén levont konklúziókat részletesen összehasonlították.

A két lépéses heurisztikus módszerünk segítségével mindhárom nyelvre megkerestük a legjobb konfigurációt, ami során olyan konfigurációkat találtunk, amik egy része új paraméter-beállításokat is tartalmaz, amik lényegesen jobb eredményt érnek el az általánosan használt beállításoknál. Habár egy heurisztikus módszert használtunk a kereséshez a lehetséges konfigurációk óriási száma miatt, igazolni tudtuk e módszerünk helyességét és az általa adott eredmények megbízhatóságát és helyességét. Továbbá igazolni tudtuk azt is, hogy egy adott bemeneti adattípuson jól működő konfiguráció másik azonos típusú bemenet használata esetén is jól működik.

A kezdeti sejtésünknek megfelelően volt jó néhány olyan paraméter, ami mindhárom nyelv esetén nagyon hasonlóan működött. Találtunk olyan paramétereket is, amik spanyolra és magyarra hasonlóan működnek, de angolra másképp, ami szintén várható volt. Mindemellett meglepődve tapasztaltuk azt, hogy volt olyan paraméter is, ami angolra és magyarra hasonlóan működött, míg spanyolra máshogyan, illetve nem találtunk olyan paramétert, amit a két indoeurópai nyelvre azonosan működött volna, de magyarra másképp. Habár azt tapasztaltuk, hogy a legjobb eredményt a különböző nyelvek esetén különböző konfigurációval lehet elérni, a nyelvek közötti tesztjeink megmutatták azt, hogy ezek mindegyike meglehetősen jól működik mindhárom nyelv esetén. Ez alapján úgy gondoljuk, hogy sikerült olyan konfigurációkat találni, melyek meglehetősen nyelv-függetlenek, és robosztus és megbízható eredményeket adnak.

Annak érdekében, hogy az eredményeinket össze tudjuk hasonlítani az eddig ismert legjobb módszerek eredményeivel, olyan teszteket is futtattunk, amiben azonos bemeneti adatokat használtunk a jelenleg ismert legjobb konfigurációkhoz és a mi konfigurációinkhoz is. Nyers frekvenciák bemenetként való használata esetén, amikor mind a 10 paramétert tudtunk vizsgálni, a legjobb konfigurációink tartalmaztak új paraméter-beállításokat és a eddigi legjobb konfiguráci-

óknál egyértelműen jobb eredményeket értek el, általában számottevően magasabb pontszámokkal. Szemantikus vektorok bemenetként való használata esetén, amikor a 10-ből csak 4 paramétert tudtunk vizsgálni, a legjobb konfigurációink legalább olyan jól teljesítettek, mint a eddigi legjobb konfigurációk, és néhány esetben kis fölénnyel is rendelkeztek. Igazából a legjobb modellünk abszolút legjobb eredményt ért el, minden eddigi modellnél jobban teljesítve a legfontosabb tesztadathalmazokon. Ezek alapján úgy gondoljuk, hogy az elemzésünk sikeres volt, és sikerült olyan új paraméter-beállításokat és új konfigurációkat bemutatni, amik az eddig ismertnél jobb eredményeket tudnak elérni, és ezáltal túlszárnyalják az eddig ismert legjobb konfigurációkat.

Ahogy a tesztek során látszódott, a bemenetként használt korpusz, illetve az alkalmazott információ-kinyerési módszer nagyban befolyásolja az eredményeket. Ebből kifolyólag úgy gondoljuk, egy a mostanihoz hasonló elemzés elvégzése a DSM-ek információkinyerési fázisán kulcsfontosságú iránya lehetne a jövőbeni kutatásoknak. Továbbá, véleményünk szerint fontos lenne az általunk újonnan javasolt konfigurációkat az általunk használt szöveges korpuszoknál nagyságrendekkel nagyobbakon tesztelni. Ennél még jobb lenne, ha ezeken az óriási korpuszokon a teljes elemzést meg lehetne ismételni. Ezen felül, habár az eredményeink meglehetősen robosztusnak és megbízhatónak tűnnek spanyolra és magyarra is, érdekes lenne az elemzésünket megismételni nagyobb és megbízhatóbb spanyol és magyar tesztadatbázisokon is, amint ilyen adathalmazok elérhetővé válnak.

Úgy gondoljuk, hogy e tanulmánnyal nagyban hozzájárultunk a DSM-ek működésének és tulajdonságainak a megértéséhez. Habár az eredményeinkből teljesen megbízható konklúziókat csak DSM-ekre tekintettel tudunk levonni, úgy gondoljuk, hogy hasonló konklúziók érvényesek lennének más, szintén vektor-tér modelleken alapuló rendszerek esetén is. Ezért úgy gondoljuk, hogy az eredményeink hasznosak lehetnek (némi fenntartással) a DSM-ek tárgykörén kívül is, más, vektor-tér modelleket alkalmazó számítógépes nyelvészeti vagy tetszőleges egyéb probléma esetén is.

Az értekezés során elért legfontosabb eredmények a következő fejezetekben kerülnek összefoglalásra. Minden fejezet végén felsorolásra kerülnek azok a publikációk, amelyek az adott eredményekhez kapcsolódnak.

Tézis 1: Eloszlás alapú szemantikai modellek tulajdonságvektorainak készítése és összehasonlítása során használt paraméterek elemzése

A kutatásunk során kitaláltunk egy heurisztikus módszert az eloszlás alapú szemantikai modellek tulajdonságvektorainak készítése és összehasonlítása során használt paraméterek elemzése. Ehhez először megvalósítottunk egy szózsák-alapú információ-kinyerési módszert, aminek segítségével a nyers input adatot készítettük a heurisztikus módszerünkhöz. Ezt követően megvizsgáltuk a lehetséges paramétereket és azok potenciális beállításait, és kiválasztottuk ezek közül fontosakat az elemzésünkhöz. Amellett, hogy az általánosan használt, standard beállításokkal is kísérleteztünk minden paraméter esetén, számos új változatot is terveztünk, különösképpen a vektorhasonlósági mértékek és a súlyozási sémák esetén. Ebben az értekezésben egy nagyon részletes áttekintését adtuk az alkalmazott heurisztikus módszernek, illetve a használt paramétereknek és azok tesztelt beállításainak, többek között a részletes képletét is prezentálva számos mértéknek. Ezáltal a munkánk többek között hasznos referenciaként is szolgálhat olyan jövőbeni kutatásokhoz, amik eloszlás alapú szemantikai modellek paramétereit elemzik, vagy egyszerűen csak azokat használnák valamilyen specifikus beállítással.

Főbb eredmények:

- Egy olyan heurisztikus módszer kitalálása, ami lehetővé tette a DSM-ek tulajdonságvektorainak készítése és összehasonlítása során lehetséges számtalan konfiguráció elemzését;
- Rengeteg új vektorhasonlósági mérték kreálása;
- Javaslat nagyszámú új súlyozási sémára;
- Számos új beállítás kitalálása a további paraméterekre is;
- Kapcsolódó publikációk: (Dobó and Csirik, 2012, 2013; Dobó, 2018; Dobó and Csirik, 2019a,b)
  András Dobó and János Csirik. Magyar és angol szavak szemantikai hasonlóságának automatikus kiszámítása. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 213–224, 2012

  András Dobó and János Csirik. Computing semantic similarity using large static corpora. In *SOFSEM 2013: Theory and Practice of Computer Science. LNCS 7741*, pages 491–502, 2013

András Dobó. Multi-D Kneser-Ney Smoothing Preserving the Original Marginal Distributions. *Research in Computing Science*, 147(6):11–25, 2018

András Dobó and János Csirik. A comprehensive study of the parameters in the creation and comparison of feature vectors in distributional semantic models. *Journal of Quantitative Linguistics*, 2019a

András Dobó and János Csirik. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT (in press)*, 2019b

Tézis 2: Angol szavak szemantikai hasonlósága

A heurisztikus elemzésünket elsőként az angol nyelvre próbáltuk ki, melynek a részletes eredményeiit az értekezés tartalmazza. Annak érdekében, hogy a módszerünk helyességéről meggyőződjünk, több kísérletet is végeztünk különféle gyakori-sági-vektor-alapú, prediktív és tudás-gráf-alapú adatokkal inputként. Ezek eredményeként meg tudtuk határozni a legjobb konfigurációt mindhárom típusú input adatok esetén. Továbbá, részletesen összehasonlítottuk a legjobb beállítási konfigurációinkat a hagyományos illetve az ismert legjobb beállítási konfigurációkkal.

Főbb eredmények:

- Legjobb konfigurációk meghatározása angol gyakorisági-vektor-alapú, prediktív és tudás-gráf-alapú modellekhez;
- A heurisztikus módszerünk helyességének és megbízhatóságának, illetve a segítségével elért eredmények robosztusságának és megbízhatóságának demonstrálása;
- Főbb konklúziók:
  - A különféle paraméterek egymástól függnek, így ahelyett, hogy külön-külön kerülnek elemzésre, azokat együttesen kell elemezni, azok egymásra hatását is figyelembe véve;
  - Másfajta konfigurációk használata szükséges gyakorisági-vektor-alapú, prediktív és tudás-gráf-alapú modellekhez;
  - Egy olyan konfiguráció, ami jól működik egy adott bemeneti adattípussal, másik azonos típusú bemenet használata esetén is jól működik;
  - Nyers frekvenciák bemenetként való használata esetén, amikor mind a 10 paramétert tudtunk vizsgálni, a legjobb konfigurációink tartalmaztak új paraméter-beállításokat és a eddigi legjobb konfigurációknál jobb eredményeket értek el;
  - Szemantikus vektorok bemenetként való használata esetén, amikor a 10-ből csak 4 paramétert tudtunk vizsgálni, a legjobb konfigurációink legalább olyan jól teljesítettek, mint a eddigi legjobb konfigurációk, és néhány esetben kis fölénnyel is rendelkeztek;
  - A legjobb modellünk, BestSv2UsingSv, ami tudás-gráf-ra épülő szemantikai vektorokon alapszik, abszolút legjobb eredményt ért el, minden eddigi modellnél jobban teljesítve a legfontosabb tesztadathalmazokon.
- Kapcsolódó publikációk: (Dobó and Csirik, 2012, 2013, 2019a,b)
  András Dobó and János Csirik. Magyar és angol szavak szemantikai hasonlóságának auto-

matikus kiszámítása. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 213–224, 2012

András Dobó and János Csirik. Computing semantic similarity using large static corpora. In *SOFSEM 2013: Theory and Practice of Computer Science. LNCS 7741*, pages 491–502, 2013

András Dobó and János Csirik. A comprehensive study of the parameters in the creation and comparison of feature vectors in distributional semantic models. *Journal of Quantitative Linguistics*, 2019a

András Dobó and János Csirik. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT (in press)*, 2019b

Tézis 3: Az angol, spanyol és magyar nyelv esetén levont konklúziók összehasonlítása

Az angol nyelvre vonatkozó elemzésünk után ugyanazt ez átfogó elemzést elvégeztük spanyolra és magyarra is. Annak érdekében, hogy az eredményeinket ki tudjuk értékelni magyarra, létre kellett hozzunk magyar nyelvű szemantikus modellek kiértékelésére alkalmas tesztadathalmazokat, mivel ilyenek nem álltak rendelkezésre. Miután megvolt minden szükséges tesztadathalmaz, a heurisztikus módszerünk segítségével meghatároztuk a legjobb konfigurációt spanyol és magyar gyakorisági-vektor-alapú modellekre is. Továbbá, összehasonlítottuk az elemzésünk első és második fázisának eredményeit a különböző nyelvekre, és ezek alapján konklúziókat vontunk le.

Főbb eredmények:

- Új általános tesztadathalmazok készítése magyar szemantikai modellek kiértékeléséhez, amik az értekezés Függelékében kerültek publikálásra;
- A legjobb konfiguráció meghatározása spanyol és magyar gyakorisági-vektor-alapú modellekre;
- Főbb konklúziók:
  - A tesztek és verifikációs lépések alapján a heurisztikus módszerünk sikeres volt magyarra és spanyolra is, és robosztus és megbízható eredményeket ad;
  - A különböző nyelvek esetén különféle konfigurációk adják a legjobb eredményt, de ezek mindegyike meglehetősen nyelv-független és jól teljesít a többi nyelv esetén is;
  - Mint ahogyan azt vártuk, voltak olyan paraméterek amik hasonlóan eredményt adtak spanyol és magyar esetén, de különbözőt az angol nyelv esetében;
  - Várakozásainkkal ellentétben nem találtunk olyan paramétert, ami angolra és spanyolra hasonlóan működött volna, de magyarra másképp;
  - Meglepetésünkre találtunk egy olyan paramétert is, ami hasonló eredményeket produkált angolra és magyarra, de ezektől különbözőt spanyolra.
- Kapcsolódó publikációk: (Dobó and Csirik, 2012, 2013, 2019a,b)
  András Dobó and János Csirik. Magyar és angol szavak szemantikai hasonlóságának automatikus kiszámítása. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 213–224, 2012

András Dobó and János Csirik. Computing semantic similarity using large static corpora. In *SOFSEM 2013: Theory and Practice of Computer Science. LNCS 7741*, pages 491–502, 2013

András Dobó and János Csirik. A comprehensive study of the parameters in the creation and comparison of feature vectors in distributional semantic models. *Journal of Quantitative Linguistics*, 2019a

András Dobó and János Csirik. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT (in press)*, 2019b

Hivatkozások

Marco Baroni and Alessandro Lenci. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd ACL*, pages 238–247, 2014.

Elia Bruni, Nam Khan Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 48:1–47, 2013.

Alexander Budanitsky and Graeme Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources*, page 2, 2001.

András Dobó. Multi-D Kneser-Ney Smoothing Preserving the Original Marginal Distributions. *Research in Computing Science*, 147(6):11–25, 2018.

András Dobó and János Csirik. Magyar és angol szavak szemantikai hasonlóságának automatikus kiszámítása. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 213–224, 2012.

András Dobó and János Csirik. Computing semantic similarity using large static corpora. In *SOFSEM 2013: Theory and Practice of Computer Science. LNCS 7741*, pages 491–502, 2013.

András Dobó and János Csirik. A comprehensive study of the parameters in the creation and comparison of feature vectors in distributional semantic models. *Journal of Quantitative Linguistics*, 2019a.

András Dobó and János Csirik. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT (in press)*, 2019b.

András Dobó and Stephen G Pulman. Interpreting noun compounds using paraphrases. *Procesamiento del Lenguaje Natural*, 46:59–66, 2011.

Zellig S. Harris. Distributional Structure. *WORD*, 10(2-3):146–162, 1954.

Felix Hill, Roi Reichart, and Anna Korhonen. Multi-Modal Models for Concrete and Abstract Concept Meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296, 2014.

Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides G M Petrakis, and Evangelos Milios. Information retrieval by semantic similarity. *International journal on semantic Web and information systems*, 2(3):55–73, 2006.

Aminul Islam and Diana Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2):1–25, 2008.

Douwe Kiela and Stephen Clark. A systematic study of semantic vector space model parameters. In *2nd CVSC at EACL*, pages 21–30, 2014.

Gabriella Lapesa and Stefan Evert. A Large Scale Evaluation of Distributional Semantic Models: Parameters , Interactions and Model Selection. *Transactions of the Association for Computational Linguistics*, 2:531–545, 2014.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the ACL*, 3:211–225, 2015.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. LexVec, 2018. URL `https://github.com/alexandres/lexvec/blob/master/README.md`. [Accessed 01.04.2019].

Robert Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *31st AAAI*, pages 4444–4451, 2017.