

Tárgyas szerkezetek elemzése tenzorfelbontással – áttekintő cikk

Makrai Márton

MTA Nyelvtudományi Intézet
makrai.marton@nytud.hu

Kivonat Áttekintjük a tenzorfelbontás számítógépes nyelvészeti alkalmazásait, különösen az igei argumentumstruktúrára vonatkozókat, és olyan asszociációs mértékekre hívjuk fel a figyelmet, amelyeket eddig nem használtak erre a feladatra.

Kulcsszavak: igei többértelműség, tenzorfelbontás, függőségi elemzés

1. Bevezetés

A *tenzorok* (>2-dimenziós tömbök) a mátrixok általánosításai: ahogy a mátrixok két tengely (sorok és oszlopok) mentén elrendezve tartalmaznak számokat, a tenzoroknak több *tengelyük* (más szóval *módjuk*¹) van. Az együtteselőfordulás-mátrix szingulárisértékfelbontása (*singular value decomposition*, SVD) természetes eszközt kínál arra, hogy általánosításokat modellezzünk két mód között a kölcsönhatásokra vonatkozóan. A két módot alkothatják szavak és a dokumentumok (látens szemantikai elemzés, *latent semantic analysis*, LSA, Landauer és Dumais (1997)), szavak és függőségi kontextusaik (Levy és Goldberg, 2014a), vagy egyszerűen a cél- (avagy fókusz-) és a kontextusszavak (szokásos szóbeágyazások, Mikolov és mtsai (2013b); Levy és Goldberg (2014b); Pennington és mtsai (2014)). Turney és Pantel (2010) szerint négyféleképpen értelmezhetjük az SVD célját: mint valamiféle látens jelentés modellezését, mint zajcsökkentést, mint közvetett (avagy magasabb rendű) együttes előfordulások modellezését (vagyis amikor két szó *hasonló kontextusokban* jelenik meg), vagy mint a ritkaság csökkentését. A nyelvben az intuíciónk szerint vannak többrendű kölcsönhatások: a *lemezját-szó szupermenesdit játszik* kifejezés furcsa (a példa Van de Cruys (2009)-ének módosítása), jöllehet azok a másodrendű kapcsolatok, hogy ⟨játszik, SUBJ, lemezját-szó⟩ és hogy ⟨játszik, OBJ, szupermenesdi⟩ tökéletesek. A mátrixfelbontás tenzorokra való általánosításai (Kolda és Bader, 2009) az ilyen háromirányú kölcsönhatások elemzéséhez nyitnak utat.

A tenzorfelbontás a neurális hálókbán szereplő szóbeágyazáshoz hasonló beágyazásvektorokat biztosít minden módhoz – a mi esetünkben az alany szerepét betöltő főnevekhez, az igékhez és a tárgy szerepét betöltő főnevekhez. Annak a projektnek,

¹ Módookról különösen azokban az alkalmazásokban beszélünk, ahol különböző modalitásból származó adatokat fuzionálnak, ahogy pl. téri és idői koordinátákat az agyi képpalkotásban.

amibe ez a cikk illeszkedik, az a hosszú távú motivációja, hogy szemantikai igeosztályokat nyerjünk az ige-beágyazásvektorok klaszterezésével (felügyeletlen, vagyis annotált adatot nem használó csoportosításával). Ha a klaszterek igeosztályoknak (Levin, 1993) felelnek meg², akkor arra számítunk, hogy a többértelmű igeék, mint a fenti *ját-szik*, kiugrónak (*outlier*) fognak bizonyulni, hiszen a különböző használataik különböző klaszterekbe kívánkoznak.

Az utóbbi évtizedben a vektoros szómodellek (amelyek neurális hálók szóbeágyazásaiként lettek különösen ismertek (Mikolov és mtsai, 2013a)) és a tenzorfelbontási algoritmusok is figyelemre méltó mértékben fejlődtek, és nyelvtechnológiai teszt-halmazokat is használtak élvonalbeli, skálázható, zajtűrő tenzorfelbontó algoritmusok tesztelésénél (Sharan és Valiant, 2017; Bailey és mtsai, 2018; Frandsen és Ge, 2019). A szótöbbértelműség – és különösen az igei szelekció valamint argumentumszerkezet – adatközpontú megértése azonban még nem mondható érettnek. Cikkünk ezt a területet igyekszik bemutatni.

Az 2. szakasz a tenzorszámításról ad egy minimális bevezetőt, és bemutat különféle asszociációs mértékeket, olyanokat is, amelyeket tudomásunk szerint még nem használtak tenzorfelbontásban. A 3. szakasz áttekinti a tenzoros nyelvészeti munkákat, különös tekintettel a bennük alkalmazott minőségi és számszerű kiértékelésre és a kapcsolódó magyar cikkekre.

2. Tenzorfelbontás

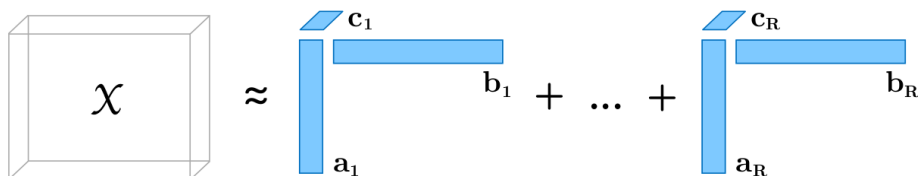
A tenzorszámítással való ismerkedéshez Kolda és Bader (2009) és Rabanser és mtsai (2017) a fő kiindulópontok. Ahogy ezekből is kiderül, nemcsak egyféleképpen lehet általánosítani az SVD alap gondolatát. A következő két szakasz a két legnépszerűbb kiterjesztést, a kanonikus poliadikus felbontást és az általánosabb Tucker-felbontást ismerteti. E két algoritmuscsalád interpretálásának lehetőségeit a jelfeldolgozás és a gépi tanulás kettős szempontjából Sidiropoulos és mtsai (2017) mutatja be.

2.1. Kanonikus poliadikus felbontás

A kanonikus poliadikus felbontás (Canonical Polyadic Decomposition, CPD, más néven CanDecomp, Parallel Factor modell, rangfelbontás vagy Kruskal-felbontás, Carroll és Chang (1970)) a eredeti tenzort 1-rangú tenzorok lineáris kombinációjaként közelíti. Egy 1-rangú tenzor nem más, mint vektorok tenzorszorzata, ugyanúgy ahogy két vektor diádszorzata egy 1-rangú mátrix, lásd az 1-es ábrát.

A váltakozó legkisebb négyzetek algoritmus (Alternating Least Squares, ALS, Carroll és Chang (1970); Harshman (1970)) iteratív módszer a CPD kiszámítására. Egy-egy iterációban egy híján az összes módot rögzítjük, és a fennmaradót illesztjük. Az ALS nem garantálja a konvergenciát, és még ha az meg is történik, nem észlelhető

² Ahogy egy korábbi változat névtelen bírálója megjegyezte, érdekes lehet számos olyan igeosztály szóbeágyazáson alapuló vizsgálata, mint „a thetikus mondatok, egzisztenciális mondatok, aspektusok, határozatlan alanyok. Vajon például megfeleltethetők-e a kapott osztályok valamilyen módon az igei aspektusoknak (pl. igeekötös igeék a magyarban)? ... Lehet-e itt szerepe a határozatlan alanyoknak?”



1. ábra: Kanonikus poliadikus felbontás, ábra Rabanser és mtsai (2017)-től

egykönnyen. Felhívjuk viszont a figyelmet az ALS-nak egy viszonylag új továbbfejlesztésére, az Orth-ALS-ra (Sharan és Valiant, 2017), lásd a 3. szakaszt.

2.2. Tucker-felbontás

Noha CPD elterjedtebb a nyelvészetben, röviden bemutatjuk az általánosabb Tucker-felbontást is. A Tucker-felbontás (más néven magasabb rendű SVD, Tucker (1966)) egy kisebb méretű \mathcal{G} magtenzort ad, amit tengelyenként egy-egy mátrixszal megszorozva az eredeti tenzor közelítését kapjuk, lásd a 2-es ábrát. Ha az eredeti tenzor tengelyei

$$\text{alany} \times \text{ige} \times \text{tárgy},$$

akkor a három mátrix sorai az alanyokat, az igéket illetve a tárgyakat beágyazó vektorok, a \mathcal{G} tenzor elemei pedig az előbbi három közötti kölcsönhatások szintjét határozzák meg.

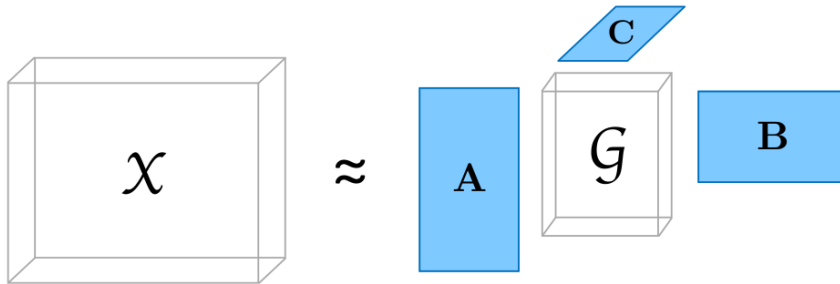
A Tucker-felbontás nem unikus, hiszen \mathcal{G} -t az illesztés romlása nélkül transzformálhatjuk, ha a tényezőmátrixokra ugyanannak a transzformációnak az inverzét alkalmazzuk. Az egyediség további követelmények bevezetésével javítható (Kolda és Bader, 2009; Lahat és mtsai, 2015), mint például ritkaság, kis elemek, \mathcal{G} teljes ortogonalitása (*all-orthogonal*), nem-negativitás vagy függetlenség.

2.3. Az együtt-előfordulások számának súlyozása

	körpusz	tengelyek	asszociációs mérték	rang
Van de Cruys (2009)	holland, .5 B	10 K alany \times 1 K ige \times 10 K direkt tárgy	PPMI	50... 300
Van de Cruys (2011)	holland, .5 B	10 K alany \times 1 K ige \times 10 K direkt tárgy	kétféle PMI	(nincs felbontás)
Van de Cruys és mtsai (2013)	UKWaC, 2 B	10 K alany \times 1 K ige \times 10 K tárgy	PMI	300
Jenatton és mtsai (2012)	2 M Wp-cikk	30 K alany \times 5 K ige \times 30 K direkt tárgy	$\mathbb{P} = 1/(1 + \exp(-s_i \cdot \mathbf{R}_j \otimes \mathbf{o}_k))$	25, 50, 100
Sharan és Valiant (2017)	Wikipedia, 1.5 B	10 K szó \times 10 K szó \times 10 K szó	$\log(f + 1)$, $w_i = s_i \oplus v_i \oplus o_i$ normalizálva	100
Bailey és mtsai (2018)	.3 B a Wp-ből	1000-es gyakorisági cut-off	(\pm eltolt) PPMI, w_i normalizálva	300

1. táblázat. Tenzoros nyelvészeti munkák. A korpuszok méretét többnyire a szavak számában mérve tüntettük fel. A képletekhez némi magyarázatot adunk a szövegben.

A nyelvi gyakoriságok ritka tömböt alkotnak, hiszen a legtöbb szó a legtöbb szóval nem fordul elő együtt empirikusan, és a gyakoriságok sok nagyságrendet ölelnek fel



2. ábra: Tucker-felbontás, ábra Rabanser és mtsai (2017)-től.

(Zipf-törvény avagy hatványeloszlás, Manin (2008); Gittens és mtsai (2017)), ezért ritka tenzorokat érdemes használni a pusztá gyakoriságoknál kifinomultabb társítási mértékekkel (*association measure*) benépesítve (*populate*). Ezekre a mértékekre térünk most rá. Az itt csak hivatkozott nyelvi témájú, tenzorfelbontást alkalmazó munkákat az 1-es táblázat összegzi és a 3. szakasz mutatja be.

A legegyszerűbb választás $\log(f + 1)$, ahol f az együttes előfordulási gyakoriság (Sharan és Valiant, 2017). Jenatton és mtsai (2012) a sokrelációs tanulás kontextusába helyezi az ⟨alany, ige, tárgy⟩ hármassok modellezését, és a log-bilineáris modell (Mnih és Hinton, 2007; Mikolov és mtsai, 2013a) súlyozási függvényét alkalmazza. van de Cruys három cikke és Bailey és mtsai (2018) egy információelméleti mérték, a *(pozitív) pontonkénti kölcsönös információ* (*(positive) pointwise mutual information, (P)PMI*), háromváltozós általánosítását használja (lásd a 2.4 szakaszt). A pozitivitás azt jelenti, hogy annak érdekében, hogy nagyobb pontszámokat tulajdonítsunk a tényleges együtt-előfordulásoknak, mint a nem-látottaknak, a PMI-nél és a következő bekezdésben bemutatott lexikográfiai mértékeknél is kinullázzuk a negatív elemeket.

Egyes lexikográfiai társítási mértékek is hasznosak lehetnek tenzorfelbontásban. A PMI kétféle háromváltozós általánosításáról a következő szakaszban szólnunk. Ezt megelőlegezve bármelyik általánosítást használva magától értődően általánosíthatjuk három változóra a Sketch Engine lexikográfiai szoftverben használt *szembetűnőséget* (*salience*, Kilgariff és mtsai (2004)) is:

$$\log(f(x, y, z)) \cdot PMI(x, y, z).$$

Kísérletezhetünk a Log-Dice (Rychlý, 2008) általánosításával is:

$$\log \frac{3f(x, y, z)}{f(x) + f(y) + f(z)} + c,$$

ahol c -t úgy választjuk, hogy a Log-Dice értékek nem-negatívak legyenek.

2.4. Többváltozós PMI

Mi más lenne a pontonkénti kölcsönös információ (PMI) többváltozós általánosítása, mint

$$\log \frac{p(x, y, z)}{p(x)p(y)p(z)}, \quad (1)$$

– gondolnánk, de valójában ez csak egy a lehetséges általánosítások közül. Van de Cruys (2011) két pontonkénti asszociációs mértéket is bevezet, amelyeknek a várható értéke a kölcsönös információ (Shannon és Weaver, 1949) egy-egy különböző többváltozós általánosítása: az interakciós információ (McGill, 1954) illetve a teljes korreláció (Watanabe, 1960).

Az *interakciós információ* a feltételes kölcsönös információ fogalmán alapul:³

$$\log \frac{p(x, y)p(x, z), p(y, z)}{p(x, y, z)p(x)p(y)p(z)}$$

A *teljes korreláció* a változóiban levő közös információ mennyiségét számszerűsíti. A pontonkénti változat képlete az 1-es egyenletben látható. Az irodalmat követve (Villa-da Moirón, 2005; Van de Cruys, 2009; Van de Cruys és mtsai, 2013; Bailey és mtsai, 2018) ebben a cikkben többváltozós PMI alatt (többváltozós pontonkénti) teljes korrelációt értünk.

Van de Cruys (2011) arról számol be, hogy holland kísérleteikben mindkét módszer ki tudott emelni száliens alany–ige–tárgy hármassokat: prototipikus SVO-kombinációkat, például *szavazás képvisel véleményt* és rögzített kifejezéseket. A *játszik* megfelelőjére szűkítve a vizsgálódást azt találják, hogy az interakciós információ prototipikus SVO kombinációkat talál, pl. *zenekar játszik szimfóniát*, míg az elterjedtebb változat, melyet ők specifikus korrelációnak neveznek, a *szerepet játszik* konstrukciót és ennek száliens alanyait taglalja.

2.5. Ritka tenzorok Python3-ban

Az adattudományban és a nyelvtechnológiában a legnépszerűbb szabad szoftverek jelenleg Python 3-on alapulnak, ezért most az itt elérhető tenzorfelbontó csomagokra térünk rá, különös tekintettel a ritka tenzorokra. A fő Python 3 könyvtárak multilineáris algebrahoz és tenzorfaktorizációkhoz a scikit-tensor-py3 (Nickel és Rol) és a tensorly (Kossai és mtsai, 2016). Mindkét könyvtár támogatja bizonyos mértékig sűrű és ritka tenzorok CPD és Tucker-felbontását.

3. A nyelvi többértelműség tenzoros modelljei

A 1 táblázat összefoglalja a nyelvészeti munkák néhány jellemzőjét. Tudásunk szerint Van de Cruys (2009) vezeti be a nem-negatív tenzorfaktorizációs modellt szelekciós-preferencia-indukcióhoz. Van de Cruys és mtsai (2013) a Kullback–Leibler-divergencia

³ A pontonkénti változat képlete a szitaformulára emlékeztet, csak fel kell cserélni a számlálót és a nevezőt, hogy matematikai értelemben is mértéket kapjunk.

minimalizálására módosítja a tenzorfaktorizációs modellt, ami szerintük az jobban illeszkedik a hosszú farkú eloszláshoz, amelyeneket a nyelvben is találunk. Ebben a cikkben a főnevek rejtett modelljeit (vagyis a szóvektorokat) előre rögzítik akkori hagyományos együttelőfordulás-alapú módszerrel, ami sajnos korlátozza a tenzorok által amúgy felderíthető harmadrendű struktúra kihasználását. Módszerük lényegi részét, a harmadrendű alany-ige-tárgy interakciók indukcióját, pedig a Tucker-felbontás ihlette.

Jenatton és mtsai (2012) a *sokrelációs tanulás (multi-relational learning)* kontextusában tanulnak szemantikus igereprezentációkat, amely paradigma eredetileg olyan entitásokkal (itt főnevek) foglalkozik, amelyek között többféle kapcsolattal (itt ige) állhat fenn, például közösségi hálók, ajánló rendszerek, szemantikus web vagy bioinformatikai adatok. Ebben a paradigmában a kapcsolatok készletét modellezzük: a kapcsolatok maguk is hasonlóak lehetnek egymáshoz különféle szempontokból. Kísérleteik során az entitásoknak egyetlen ábrázolása van az összes relációra vonatkozóan. A nyelvi tenzort (alany, ige, közvetlen tárgy) együtt-előfordulásokkal népesítik be.

Polajnar és mtsai (2014) a zaj-kontrasztív becslés módszerét (amit ők *plausibility training*nek hívnak) alkalmazzák tranzitív ige-tenzorok felbontásához. Zhang és mtsai (2014) azt vizsgálják, hogy miként lehet a kézzel létrehozott szemantikai erőforrásokat neurális szóbeágyazásokkal kombinálni az antonimáknak a szinonimáktól való elválasztására, ami közismerten nehéz az eloszlásalapú eszközök számára. A teaurusz adatait és az eloszlási hasonlóságokat tenzorok egy-egy szeletként (táblájaként) fecskendezik be.

A *függvényes (functional)* megközelítésben a szavaknak különböző rendű tenzorok felelnek meg. Egy szóhoz tartozó tenzor rendje összhangban van a szónak egy kategoriális nyelvtanban való típusával. Például a főnevek atomi típusok, amelyeket egy vektor képvisel, és a melléknevek olyan mátrixok, amelyek függvényként működnek (Baroni és Lenci, 2010). A tranzitív ige harmadrendű tenzor. Ebben a megközelítésben probléma, hogy meglehetősen sok paraméter lehet már alacsony dimenziójú is. (Fried és mtsai, 2015) úgy orvosolja ezt a problémát, hogy a tenzorok alacsony rangú közelítését használják.

Hashimoto és Tsuruoka (2015) is mátrixként ábrázolják a tranzitív igéket. Modelljük impliciten faktorizál egy tenzort abban az értelemben, ahogy a skip-gram is implicit mátrixfelbontás (Levy és Goldberg, 2014b). A tárgyias igék több jelentését megragadják, és egyértelműsítik őket az argumentumaik alapján. A szabad bővítmények hozzájárulását is vizsgálják.

Cotterell és mtsai (2017) a skip-gram modellt általánosítják tenzorfelbontásként, ami lehetővé teszi beágyazások tanítását gazdagabb, magasabb rendű együtt-előfordulásokból, pl. olyan hármásokból, amelyek a kontextusszónak a fókuszszóhoz képesti helyzetére vonatkozó információt is tartalmaznak, vagy morfológiai információt a kapcsolódó szavak közötti paramétermegosztás érdekében. Negyven nyelven kísérleteznek. Ferraro és mtsai (2017) ezt a modellt használva keretszemantikán alapuló tenzorokat tanítanak. A szemantikus proto-szerepeket (*semantic proto-role*, SPR, Dowty (1991)) egyfajta folytonos keretszemantikának (Fillmore és mtsai, 1976) tekintik, ami bizonyos tulajdonságok valószínűségét ragadja meg, a szerepeket pedig e tulajdonságok csoportjaiként jellemzik. Ferraroék ilyen SPR-alapú várható tulajdonságokat rögzítenek szóbeágyazásokban.

Sharan és Valiant (2017)⁴ egy általános szóbeágyazást készít szimmetrikus 3-módú tenzorokból. Cikkük lényege az *Ortogonalis ALS (Orth-ALS)*, a ALS megközelítésnek egy olyan módosítása, ami ugyanolyan hatékony, mint a szokásos ALS, de bizonyíthatóan megtalálja a valódi tényezőket véletlen inicializálással a szokásos inkoherenciafeltételezések mellett, azaz hogy a valódi tényezők kevéssé korrelálnak egymással, ami teljesül az NLP-s alkalmazásokban, ahol a közelítő tenzor rangja általában szignifikánsan szublineáris a tér dimenziójában. Az Orth-ALS időről időre „ortogonalizálja” a tényezők becslését, megakadályozva, hogy több kiszámított tényező ugyanazt a valódi tényezőt „üldözze”. A szóbeágyazásokat úgy hozzák létre, hogy a három kiszámolt faktormátrixot (egyenként 100 látens dimenzió) konkatenálják egy 300-oszlopos mátrixszá, majd normalizálják a sorokat.

Megemlítjük Bailey és mtsai (2018)⁵-at is, akik egy 3-módú szimmetrikus tenzort képeznek, amely azzal a szójelentés-klaszterezési szempontból figyelemre méltó tulajdonsággal bír, hogy egy alkalmas kontextusvektorral való pontonkénti szorzás segítségével jelentésvektorokat kapnak. Végül Frandsen és Ge (2019) *Szintaktikus RAND-WALK* modellje különféle mondattani kapcsolatokat ragad meg egy szóhármassok közötti PMI-t alkalmazó tenzorral.

3.1. Minőségi elemzés a munkákban

<i><athlete, run, race></i>	<i>finish (.29), attend (.27), win (.25)</i>
<i><user, run, command></i>	<i>execute (.42), modify (.40), invoke (.39)</i>
<i><man, damage, car></i>	<i>crash (.43), drive (.35), ride (.35)</i>
<i><car, damage, man></i>	<i>scare (.26), kill (.23), hurt (.23)</i>

2. táblázat. Tranzitív szerkezetben kontextualizált igékhez leghasonlóbb igék Van de Cruys és mtsai (2013)-nál.

A Van de Cruys (2009) által végzett kvalitatív kiértékelés a látens dimenziók elemzésén alapul: az egyes dimenziókat az azokban legnagyobb abszolút értékű koordinátát kapó alanyok, igék és tárgyak szerint értelmezik. Úgy találják, hogy a 100 dimenzió közül 44 keretszemantikát példáz. Egy olyan dimenzióban, amit úgy hívhatunk, hogy *rendőrség letartóztat gyanúsítottat*, a legnagyobb súlyú alanyok, igék és tárgyak olyan szavak, mint például *rendőrség, letartóztat* illetve *gyanúsított*. További példák: *többség támogat javaslatot* vagy *kormány küld csapatot*. További 43 dimenzió szemantikája kevésbé egyértelmű: ezek egyetlen igét képviselnek, esetleg egy ige különféle jelentései keverednek. Tizenhárom rejtett dimenzió konkrét igei szerkezeteket tartalmaz, például *x játszik szerepet*, ahol az alanyi oszlop egyenletesen oszlik meg több tucat szó között, pl. *bosszú, szégyen, intézmény, kultúra* vagy *osztódás*.

⁴ <http://web.stanford.edu/~vsharan/orth-als.html>

⁵ https://github.com/popcorncolonel/tensor_decomp_embedding

Van de Cruys és mtsai (2013) tenzorában a szeletek igéket képviselnek. Ők úgy szemléltetik az adatokat, hogy hármásokhoz a bennük szereplő, kontextualizált igékhez leghasonlóbb igéket mutatják meg, lásd a 2-es táblázatot.

A Frandsen és Ge (2019) kvalitatív kiértékelésében együttes melléknév-főnév illetve ige-tárgy vektorokhoz legközelebbi beágyazású szavakat néznek.

3.2. Számszerű elemzés a munkákban

Van de Cruys (2009) ál-egyértelműsítési feladatban értékeli ki a modelljét, ahol azt kell megítélni, hogy melyik alany (s vagy s') és közvetlen tárgy (o vagy o') valószínűbb egy adott v ige esetében. A tesztkészletet úgy építi fel, hogy $\langle s, v, o \rangle$ -t a korpuszból veszi, míg s' és o' egy-egy véletlenszerűen választott alany illetve közvetlen tárgy a korpuszból, például *fiatal/koalíció iszik sört/részvényt*. Tudomásunk szerint a tesztkészletük nem érhető el.

Grefenstette és Sadrzadeh (2011) tárgyas igék egyértelműsítésére vonatkozó adathalmaza igepárokat tartalmaz egy-egy alannyal és tárggyal. A feladat az igék többértelműségén alapszik (Kartsaklis és Sadrzadeh, 2013; Milajevs és mtsai, 2014; Polajnar és mtsai, 2014). Például az angol *meet* ige többértelmű; egyik jelentésében a *satisfy*-hoz hasonlít, egy másikban a *visit*-hez: *A beach meet standard* kontextusban a *satisfy*-hoz és csak ahhoz, a *representative meet official* környezetben pedig a *visit*-hez. A feladat ezen hasonlóságok predikciója. Van de Cruys és mtsai (2013) ezeken a tárgyas mondatokon értékeli ki számszerűen a rendszerüket.

Kartsaklis és Sadrzadeh (2013) egy másik tesztalmozatot, Mitchell és Lapata (2010) \langle ige, tárgy \rangle szerkezetek hasonlóságára vonatkozó adatát egészíti ki: az eredeti párokat alanyokkal látja el úgy, hogy a hasonlóság mértékét igyekeznek őrizni, hogy az emberi hasonlóságítéletek érvényesek maradjanak. Kartsaklis és Sadrzadeh (2014) olyan változatát adja közre az adathalmaznak,⁶ ahol már a hármásokat értékeltetik ki Amazon Türrkel. Polajnar és mtsai (2014), Fried és mtsai (2015) és Hashimoto és Tsuruoka (2015) Grefenstette-ék adathalmazán és ez(ek)en értékelnek ki.

Jenatton és mtsai (2012) két feladatban értékeli ki modelljeiket: adott alanyhoz és közvetlen tárgyhoz jósolják be a megfelelő igét, illetve lexikai hasonlósági osztályozást végeznek. Ők is közzéteszik a tesztadatot, de mi azt találtuk, hogy hármasaik kissé zajosabbak pl. Grefenstette-ékéinél.

Noha nem triviális, hogy az antonímia jobban támaszkodik-e a hárommódú együttes előfordulásokra, mint például a szinonímia, Zhang és mtsai (2014) GRE antonim kérdésekben (Mohammad és mtsai, 2008) értékeli ki a munkájukat. A tesztalmozatról Zhangék azt írják, hogy az adatkészlet „szemmel láthatóan köztulajdonban” van, és kérésre elérhető.

A szövektorok tesztelésének egyik legnépszerűbb módszere a szóhasonlósági rangsorolási feladat (Cotterell és mtsai (2017) is így értékelnek ki), kiváltképp a SimLex-999 (Hill és mtsai, 2014). Noha a szópárok hasonlósága nem közvetlenül a háromirányú interakciókat célozza meg, úgy gondoljuk, hogy a SimLex-999 igei megfelelője, a Sim-Verb (Gerz és mtsai, 2016), sőt az alanyok és a tárgyak tekintetében maga a SimLex-999

⁶ <http://www.cs.ox.ac.uk/activities/compdistmeaning/>

is, hasznos józanság-ellenőrzést (*sanity check*) nyújt a tenzorfelbontó modellek számára is.

Sharan és Valiant (2017) nem teszleti kifejezetten a >2 -rendű kapcsolatok modellezését, hanem szokásos szóanalógiában („a *kutyához* úgy viszonyul a *kan*, mint a *macskához* a(z) *x*”) és szemantikai szóhasonlósági feladatokban értékelik ki az ortogonalizált tenzorként nyert beágyazásokat. Azzal, hogy az Orth-ALS-t használják a szokásos ALS helyett, jelentős javulást kapnak, ám a mátrix-SVD módszer továbbra is felülmúlja a tenzoralapú módszereket. Miután felvetik azt a pesszimista magyarázatot, miszerint a természetes nyelv esetleg nem tartalmaz eléggé gazdag magasabb rendű függőségeket a szűk kontextusban megjelenő szavak között a 2-módú szerkezeten túl, másodikként azt a lehetséges magyarázatot adják a gyenge teljesítményre, hogy csak a két vizsgált feladathoz nem szükséges ez a fajta magasabb rendű statisztika. Végül Frandsen és Ge (2019) a már említett melléknév–főnév szókapcsolatok hasonlóságára vonatkozó feladatban (Mitchell és Lapata, 2010) értékelik ki a munkájukat.

3.3. Magyar munkák

Bár a cikkünk angol tárgynyelvű, a magyar konferencia közönsége számára érdekes lehet a kapcsolódó magyar munkák bemutatása – a teljesség leghalványabb igénye nélkül. A magyar nyelvű szójelentés-osztályozás (*word sense disambiguation*) a gépi tanulás szempontjából legalább Miháltz (2005)-ig és Vincze és mtsai (2008)-ig nyúlik vissza. Az igéknek sok kutató szentelte a figyelmét nyelvészektől a szűkebb értelemben vett nyelvtechnológusokig (Dressler és Ladányi, 2000; Kuti és mtsai, 2010; Miháltz és Sass, 2013).

A magyar igei konstrukciók fő adatbázisai a megjelenés sorrendjében a Mazsola (Sass, 2015, 2018), a Tádé (Kornai és mtsai, 2016) és a Manócska (Kalivoda és mtsai, 2018; Kalivoda, 2019). A magyar szóbeágyazós munkák közül Makrai (2015); Siklósi (2016); Berend (2018); Kardos és mtsai (2019) és Döbrössy és mtsai (2019) munkáit emeljük ki, mint cseppeket a tengerből.

4. Következtetés és a jövőbeni kutatás

Összességében elmondhatjuk, hogy a tenzorfelbontás a fősdorra (Hewitt és Manning, 2019) merőleges irányt kínál a nyelvi szerkezet adatvezérelt megértésében. Hosszú távon, amint azt az 1. szakaszban már említettük, szemantikai igeosztályokat szeretnénk felügyeletlenül tanulni. Ha az igei beágyazás-vektorok Levin (1993)-féle igeosztályoknak megfelelő klaszterekbe rendeződnek, akkor a többértelmű igéket a klaszterekből kimaradó vektorok formájában azonosíthatjuk be. Ez a kutatási vonal a többnyelvű paradigmába is kiterjeszthető (Vulić és mtsai, 2017; Majewska és mtsai, 2018; Sun és mtsai, 2010).

Köszönetnyilvánítás

Hálás vagyok Tülay Adalinnak, aki a 2018-as DeepLearn nyári egyetem lelkesítő előadójaként felhívta a figyelmemet a tenzorfelbontás általi interpretációban rejlő lehetőségekre, valamint Berend Gábornak, Borbély Gábornak, Indig Balázsnak, Kalivoda

Ágnesnek, Kornai Andrásnak, Sass Bálintnak, Simon Eszternek, Szécsényi Tibornak és korábbi változatok névtelen bírálóinak (MSZNY 2019, ACL 2019, Repl4NLP 2019, Maleczki 65) hasznos megjegyzéseikért. Kutatásomat részben a 2018-1.2.1-NKP-2018-00008 *A mesterséges intelligencia matematikai alapjai* és az NKFIH 120145-ös *Szó-szerkezet felismerése mélytanulással* projekt támogatta.

Hivatkozások

- Bailey, E., Meyer, C., Aeron, S.: Learning semantic word representations via tensor factorization (2018), <https://openreview.net/forum?id=BlkIr-WRb>, arXiv:1705.08968 [cs.AI]
- Baroni, M., Lenci, A.: Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4), 673–721 (2010)
- Berend, G.: Towards cross-lingual utilization of sparse word representations. In: Vincze, V. (szerk.) XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018). pp. 272–280. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2018)
- Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition. *Psychometrika* 35, 283–319 (1970)
- Cotterell, R., Poliak, A., Durme, B.V., Eisner, J.: Explaining and generalizing skip-gram through exponential family pca. In: ACL (2017)
- Van de Cruys, T.: A non-negative tensor factorization model for selectional preference induction. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics. pp. 83–90. Association for Computational Linguistics, Athens, Greece (Mar 2009), <https://www.aclweb.org/anthology/W09-0211>
- Van de Cruys, T.: Two multivariate generalizations of pointwise mutual information. In: Proceedings of the Workshop on Distributional Semantics and Compositionality. pp. 16–20. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), <https://www.aclweb.org/anthology/W11-1303>
- Van de Cruys, T., Poibeau, T., Korhonen, A.: A tensor-based factorization model of semantic compositionality. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1142–1151. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), <https://www.aclweb.org/anthology/N13-1134>
- Döbrössy, B., Makrai, M., Tarján, B., Szaszák, G.: Investigating sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian. In: Proceedings of the 4th Workshop on Representation Learning for NLP (Repl4NLP-2019). pp. 187–193. Association for Computational Linguistics, Florence, Italy (Aug 2019), <https://www.aclweb.org/anthology/W19-4321>
- Dowty, D.: Thematic proto-roles and argument selection. *Language* 67(3), 547–619 (1991)
- Dressler, W.U., Ladányi, M.: Productivity in word formation (wf): a morphological approach. *Acta Linguistica Hungarica* 47(1-4), 103–145 (2000)
- Ferraro, F., Poliak, A., Cotterell, R., Durme, B.V.: Frame-based continuous lexical semantics through exponential family tensor factorization and semantic proto-roles. In: Joint Conference on Lexical and Computational Semantics (*SEM) (2017)

- Fillmore, C.J., és mtsai: Frame semantics and the nature of language. In: *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*. pp. 20–32 (1976)
- Frandsen, A., Ge, R.: Understanding composition of word embeddings via tensor decomposition. In: *7th International Conference on Learning Representations, ICLR 2019* (2019), <https://openreview.net/forum?id=H1eqjiCctX>, arXiv preprint arXiv:1902.00613
- Fried, D., Polajnar, T., Clark, S.: Low-rank tensors for verbs in compositional distributional semantics. In: *ACL* (2015)
- Gerz, D., Vulić, I., Hill, F., Reichart, R., Korhonen, A.: SimVerb-3500: A large-scale evaluation set of verb similarity. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pp. 2173–2182. Association for Computational Linguistics, Austin, Texas (Nov 2016), <https://www.aclweb.org/anthology/D16-1235>
- Gittens, A., Achlioptas, D., Mahoney, M.W.: Skip-gram – zipf + uniform = vector additivity. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 69–76. Association for Computational Linguistics (2017), <http://aclweb.org/anthology/P17-1007>
- Grefenstette, E., Sadrzadeh, M.: Experimenting with transitive verbs in a DisCoCat. In: *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. pp. 62–66. Association for Computational Linguistics, Edinburgh, UK (Jul 2011), <https://www.aclweb.org/anthology/W11-2507>
- Harshman, R.A.: Foundations of the parafac procedure: Models and conditions for an “explanatory” multi-modal factor analysis. *UCLA Working Papers in Phonetics* 16, 1–84 (1970), <http://publish.uwo.ca/~harshman/wpppfac0.pdf>
- Hashimoto, K., Tsuruoka, Y.: Learning embeddings for transitive verb disambiguation by implicit tensor factorization. In: *3rd Workshop on Continuous Vector Space Models and their Compositionality* (2015)
- Hewitt, J., Manning, C.D.: A structural probe for finding syntax in word representations. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp. 4129–4138 (2019)
- Hill, F., Reichart, R., Korhonen, A.: Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics* 2(10), 285–296 (2014)
- Jenatton, R., Roux, N.L., Bordes, A., Obozinski, G.: A latent factor model for highly multi-relational data. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*. pp. 3167–3175. NIPS’12, Curran Associates Inc., USA (2012), <http://dl.acm.org/citation.cfm?id=2999325.2999488>
- Kalivoda, A.: Véges erőforrás végtelen sok igekötős ígére [A finite resource for infinitely many Hungarian particle verbs]. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) *XV. Magyar Számítógépes Nyelvészeti Konferencia*. pp. 331–344 (January 2019)
- Kalivoda, Á., Vadász, N., Indig, B.: MANÓCSKA: A Unified Verb Frame Database for Hungarian. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (szerk.) *Proceedings of*

- the 21st International Conference on Text, Speech and Dialogue—TSD 2018, Brno, Czech Republic. *Lecture Notes in Artificial Intelligence*, vol. 11107, pp. 135–143. Springer-Verlag (Sep 2018)
- Kardos, P., Berend, G., Farkas, R.: Kísérletek tudásbázis- és mondatkörnyezet-alapú be-
 ágyazásokkal magyar nyelvre. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV.
 Magyar Számítógépes Nyelvészeti Konferencia. pp. 153–162. Szegedi Tudomány-
 egyetem, Informatikai Intézet (2019)
- Kartsaklis, D., Sadrzadeh, M.: Prior disambiguation of word tensors for constructing
 sentence vectors. In: EMNLP (2013)
- Kartsaklis, D., Sadrzadeh, M.: A study of entanglement in a categorical framework of
 natural language. In: The 11th workshop on Quantum Physics and Logic (6 2014),
 arXiv:1412.8102
- Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: Sketch engine. In: Williams, G., Vessi-
 er, S. (szerk.) Proceedings of Euralex. pp. 105–116. Lorient, Université de Bretagne-
 Sud, Faculté des lettres et des sciences humaines (July 2004)
- Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM review* 51(3),
 455–500 (2009)
- Kornai, A., Nemeskey, D.M., Recski, G.: Detecting optional arguments of verbs. In:
 Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B.,
 Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of
 the Tenth International Conference on Language Resources and Evaluation (LREC
 2016). pp. 2815–2818. European Language Resources Association (ELRA), Paris,
 France (may 2016)
- Kossaiifi, J., Panagakis, Y., Anandkumar, A., Pantic, M.: Tensorly: Tensor learning in
 python. *Journal of Machine Learning Research (JMLR)* 20, 1–6 (2016), arXiv pre-
 print arXiv:1610.09555
- Kuti, J., Héja, E., Sass, B.: Sense disambiguation – „ambiguous sensation”? evaluating
 sense inventories for verbal wsd in hungarian. In: Calzolari, N., Choukri, K., Mae-
 gaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., Tapias, D. (szerk.) Procee-
 dings of LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools
 for Central and (South-)Eastern European Languages. pp. 23–30. European Lan-
 guage Resources Association (ELRA) (2010)
- Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, chal-
 lenges, and prospects. *Proceedings of the IEEE* 103(9), 1449–1477 (2015)
- Landauer, T.K., Dumais, S.T.: A solution to plato’s problem: The latent semantic analy-
 sis theory of acquisition, induction, and representation of knowledge. *Psychological
 review* 104(2), 211 (1997)
- Levin, B.: *English Verb Classes and Alternations: A Preliminary Investigation*. Univer-
 sity of Chicago Press (1993)
- Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the
 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2:
 Short Papers). pp. 302–308. Association for Computational Linguistics, Baltimore,
 Maryland (June 2014a), <http://www.aclweb.org/anthology/P14-2050>
- Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In:
 Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (szerk.) *Ad-
 vances in Neural Information Processing Systems* 27. pp. 2177–2185 (2014b)

- Majewska, O., Vulić, I., McCarthy, D., Huang, Y., Murakami, A., Laippala, V., Korhonen, A.: Investigating the cross-lingual translatability of VerbNet-style classification. *Language Resources and Evaluation* 52(3), 771–799 (2018)
- Makrai, M.: Comparison of distributed language models on medium-resourced languages. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015). pp. 22–33. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2015)
- Manin, D.Y.: Zipf’s law and avoidance of excessive synonymy. *Cognitive Science* 32, 1075–1098 (2008)
- McGill, W.: Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory* 4(4), 93–111 (1954)
- Miháltz, M.: Towards a hybrid approach to word-sense disambiguation in machine translation. In: Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., Nikolov, N. (szerk.) RANLP-2005 Workshop: Modern Approaches in Translation Technologies (September 2005)
- Miháltz, M., Sass, B.: What do we drink? automatically extending hungarian wordnet with selectional preference relations. In: *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*. pp. 105–109 (2013)
- Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (szerk.) 1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings (May 2013a), <http://arxiv.org/abs/1301.3781>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (szerk.) *Advances in Neural Information Processing Systems* 26, pp. 3111–3119. Curran Associates, Inc. (2013b)
- Milajevs, D., Kartsaklis, D., Sadrzadeh, M., Purver, M.: Evaluating neural word representations in tensor-based compositional settings. In: *EMNLP*. pp. 708–719 (2014)
- Mitchell, J., Lapata, M.: Composition in distributional models of semantics. *Cognitive Science* 34, 1388–1429 (2010)
- Mnih, A., Hinton, G.: Three new graphical models for statistical language modelling. In: *Proceedings of the 24th international conference on Machine learning*. pp. 641–648. ACM (2007)
- Mohammad, S., Dorr, B., Hirst, G.: Computing word-pair antonymy. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp. 982–991. Association for Computational Linguistics (2008)
- Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543. Association for Computational Linguistics (2014), <http://www.aclweb.org/anthology/D14-1162>
- Polajnar, T., Rimell, L., Clark, S.: Using sentence plausibility to learn the semantics of transitive verbs. In: *NIPS Learning Semantics Workshop (2014)*, in arXiv, some minor errata fixed.
- Rabanser, S., Shchur, O., Günnemann, S.: Introduction to tensor decompositions and their applications in machine learning (2017), <http://arxiv.org/abs/1711.10781v1>, arXiv:1711.10781 [stat.ML]

- Rychlý, P.: A lexicographer-friendly association score. In: Proceedings of Recent Advances in Slavonic Natural Language Processing. pp. 6–9 (2008)
- Sass, B.: A lattice based algebraic model for verb centered constructions. In: TSD. pp. 231–238. Springer (2018)
- Sass, B.: 28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet [28 million syntactically analyzed sentences and 500 000 verb constructions in Hungarian]. In: Attila, T., Viktor, V., Veronika, V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015). pp. 303–308. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2015)
- Shannon, C.E., Weaver, W.W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)
- Sharan, V., Valiant, G.: Orthogonalized ALS: A theoretically principled tensor decomposition algorithm for practical use. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017. pp. 3095–3104 (August 2017), <http://proceedings.mlr.press/v70/sharan17a.html>
- Sidiropoulos, N.D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E.E., Faloutsos, C.: Tensor decomposition for signal processing and machine learning. IEEE Transactions on signal processing 65(13), 3551–3582 (Jul 2017), <https://doi.org/10.1109/TSP.2017.2690524>
- Siklósi, B.: Using embedding models for lexical categorization in morphologically rich languages. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 115–126. Springer (2016)
- Sun, L., Korhonen, A., Poibeau, T., Messiant, C.: Investigating the cross-linguistic potential of VerbNet: style classification. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 1056–1064. Association for Computational Linguistics (2010)
- Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika 31(3), 279–311 (1966)
- Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37, 141–188 (2010)
- Villada Moirón, M.B.: Data-driven identification of fixed expressions and their modifiability. Ph.D.-értekezés, University of Groningen (2005)
- Vincze, V., Szarvas, G., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, C., Csirik, J.: Hungarian word-sense disambiguated corpus. In: Calzolari, N., ChoukriK, Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D. (szerk.) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). pp. 3344–3349. European Language Resources Association (ELRA) (2008)
- Vulić, I., Mrkšić, N., Korhonen, A.: Cross-lingual induction and transfer of verb classes based on word vector space specialisation. arXiv preprint arXiv:1707.06945 pp. 2546–2558 (Sep 2017), <https://www.aclweb.org/anthology/D17-1270>
- Watanabe, S.: Information theoretical analysis of multivariate correlation. IBM Journal of research and development 4(1), 66–82 (1960)
- Zhang, J., Salwen, J., Glass, M., Gliozzo, A.: Word semantic representations using Bayesian probabilistic tensor factorization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1522–

1531. Association for Computational Linguistics, Doha, Qatar (Oct 2014), <https://www.aclweb.org/anthology/D14-1161>