Magyar szóbeágyazási modellek kézi kiértékelése

Novák Attila, Novák Borbála

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

Kivonat A szóbeágyazási modellek a lexikai szemantikai tudásreprezentáció hatékony eszközének bizonyultak, és természetes módon illeszkednek az utóbbi években elterjedt neurálisháló-alapú modellekbe, ezért alkalmazásuk nagyon népszerűvé vált. Ezeket a modelleket leggyakrabban valamilyen gépi osztályozási vagy annotációs feladatot végző rendszer részeként értékelik ki. Ebben a cikkben ezzel szemben magyar nyelvű szóbeágyazási modellek közvetlen kézi kiértékelésének eredményét mutatjuk be. Az eredmények azt mutatják, hogy a morfológiailag elemzett előfeldolgozott korpuszból készített modellek jobb teljesítményt nyújtanak a kézi kiértékelésünk alapját képező szóhasonlósági teszteken, mint a nyers szövegekből készített modellek.

1. Bevezetés és kapcsolódó kutatások

A szavak megfelelő gépi ábrázolása alapvető feladat a természetesnyelv-feldolgozó rendszerekben. Az utóbbi években elterjedt neurális szóbeágyazási modellek hatékony szóreprezentációs formalizmusnak bizonyultak [1,2,3]. Ezen modellek kiértékelésére számos módszert javasoltak. A modellek kiértékelésével foglalkozó tanulmányok egy része a külső kiértékelés mellett foglal állást: a beágyazási modellek által a szavakhoz rendelt vektorokat valamilyen osztályozási vagy annotációs feladathoz használják jegyekként, és azt mérik, hogy a modellválasztás hogyan befolyásolja a rendszer adott feladatban nyújtott teljesítményét. Ez a módszer nem közvetlenül méri a beágyazási modell minőségét, hanem csak annak hatását jellemzi az azt felhasználó rendszer teljesítményére, ez utóbbi azonban jelentős mértékben függ a feladat jellegétől.

A beágyazási modellek kiértékelésével foglalkozó tanulmányok nagy része [4,3] úgy végez kiértékelést, hogy azt mérik, hogy milyen korrelációt mutat az adott modell által egy adott tesztanyagban szereplő szópárokhoz rendelt szóhasonlósági mérték azzal, amennyire az adott (rendszerint korábban elvégzett) kísérletben résztvevő emberek hasonlónak ítélték azokat. Azonban mint Schnabel és társai kimutatták [5], az emberek, illetve a gépi modellek által felállított különböző tartományokba eső pontszámokon alapuló különböző rangsorok nem feltétlenül jól összehasonlíthatóak, és az így kapott pontszámok aggregálásával kapott eredmények félrevezetőek lehetnek.

Problémát jelent például, hogy csak a szópárok egy részhalmazára vonatkozó rangsorolás áll rendelkezésre a kiértékelés alapjául szolgáló anyagokban. Azonban ahhoz, hogy a rangsor-korreláció tényleg megbízhatóan használható metrika lehessen, az olyan szavak közötti hasonlóság emberek szerint kvantifikált mértékére is szükség lenne, amelyeknek „semmi közük egymáshoz". A kiértékeléshez használt szópárlistákkal kapcsolatban azt is problémaként vetik fel, hogy azok nem kiegyensúlyozott mértékben reprezentálják a szókincset. Ezért azt javasolják, hogy szóbeágyazási modellek kiértékelésére olyan hívószólistát használjunk, amely gyakoriság, szófaj és a jelentés absztraktsága szempontjából is kiegyensúlyozott.

A legtöbb tanulmány elsősorban a szóbeágyazási modellek angolra, illetve más általában viszonylag egyszerű morfológiájú nyelvre való alkalmazására koncentrál. Emberek által a szóhasonlóság mértéke szempontjából kiértékelt szógyűjtemények is csak néhány nyelvre állnak rendelkezésre. A viszonylag korlátos szókincs, illetve a viszonylag kötött szórend ezeknek a nyelveknek az esetében lehetővé teszi, hogy a felhasznált egyszerű neurális hálózatok a szójelentést jól megragadó vektortérmodelleket hozzanak létre.

A kiértékelést általában úgy végzik, hogy egy adott korpuszból különböző algoritmussal, illetve különböző paraméterbeállításokkal készített modelleket hasonlítanak össze, és az algoritmusválasztás, illetve a paraméterbeállítások hatását mérik. A morfológiailag komplex nyelvek esetében azonban fontos kérdés az is, hogy a modell mennyire képes az adathiánnyal kapcsolatos problémákat kezelni.

Ebert és kollégái [6] lemmatizált, illetve tövesített korpuszokból készítettek szóbeágyazási modelleket morfológiai szempontból különböző mértékben komplex nyelvekre (többek között magyarra is), és arra az eredményre jutottak, hogy a szóhasonlósági feladatokban az általuk vizsgáltak közül a lemmatizált korpuszból épített modellek teljesítenek legjobban. Kiértékelésüket a WordNetre alapozták. Azt vizsgálták, hogy a modell által az adott szóhoz generált hasonlósági listán hányadik pozícióban szerepel az első olyan szó, amely az adott szó WordNet-beli reprezentációjából két lépésben elérhető valamely a WordNetben szereplő reláció mentén. Ezen pozíció reciprokának átlagát (MRR: Mean Reciprocal Rank) használták a modelleket kiértékelő metrikaként.

Sajnos a magyar WordNet (HuWN) [7] alacsony és kiegyenlítetlen lexikai fedése miatt nem igazán alkalmas arra, hogy megbízható referenciaanyagként szolgáljon a szóhasonlósági modellek kiértékeléséhez. A HuWN 19400 főnévi synsetjének 83%-a tulajdonnév (elsősorban településnév), ugyanakkor a viszonylag gyakori köznevek fedése nem nagyon jó. Egy másik probléma, hogy az angol (2.0) WordNet struktúráját követi, ezért a csomópontok 6.7%-a olyan szellemcsomópont, amihez egyáltalán nem kapcsolódik magyar lexikai elem. Ebert és munkatársai [6] feltehetőleg többek között ezek miatt a problémák miatt kaptak jóval alacsonyabb MRR pontszámokat magyarra, mint más nyelvekre. A 3 részben leírt kísérleteinkben mindenesetre a magyar WordNetet is felvettük a tesztelt erőforrások közé.

Az itt bemutatott kutatásban a célunk az volt, hogy felmérjük, hogy a korpusz előfeldolgozása milyen hatással van a létrejövő modell minőségére. A kor-

puszt, a vektortérmodell létrehozására alkalmazott algoritmust, és annak paramétereit rögzítettük. Ezután részletes közvetlen kézi kiértékelést végeztünk a különböző módon annotált magyar nyelvű korpuszból készült szóbeágyazási modelleken.

2. Magyar szóbeágyazási modellek

Négy modellt építettünk a `word2vec` eszköz CBOW modelljét használva. A korpusz mindegyik esetben ugyanaz az 1,2 milliárd szavas webről gyűjtött korpusz volt. A modellek paraméterei mindegyik esetben a következők voltak: 10 szavas kontextusablak, 300 dimenzió, minimális korpuszgyakoriság: 5.

Első modellünk a tokenizált, de egyébként nyers korpuszból készült (`SURF`). Ebben a modellben felszíni szóalakok szerepelnek. Ez a modell – legalábbis a viszonylag gyakori szavak esetében – a jelentés mellett jól megragadja a morfológiai hasonlóságokat is (amennyire azok a szavak szintaktikai viselkedését is befolyásolják). A morfológiailag analóg *jó – rossz* és *jobb – rosszabb* szópárok hasonlósága jóval nagyobb ebben a modellben, mint a lemma és valamelyik toldalékolt alak pl. *jó – jobb*, illetve *rossz – rosszabb* hasonlósága. Az 1. táblázat első két oszlopa felszíni szóalakok nyers korpuszból épített modellbeli legközelebbi szomszédait mutatja. A *kenyerek* kérdőszóhoz többnyire többes számú péksütemények hasonlítanak legjobban. Ez a modell azonban érzékeny az adathiányproblémára. A *kenyér* szónak 197 különböző ragozott alakja szerepel a korpuszban, a 'kenyér' jelentésreprezentációja ezek között oszlik meg. A `SURF` modell gyakran nem képes arra, hogy a ritka szóalakok jelentését megbízható módon megragadja. Az 1. táblázat második oszlopában szereplő *Vakkalit* szóalakhoz (a Vakkali név tárgyesete) a modell teljesen oda nem illő szavakat rendel hozzá.

kenyerek(2270) SURF	Vakkalit(5) SURF	kenyér(147000) ANA	Vakkali(23) ANA
kiflik(349)	tevedesnek(5)	hús(136814)	Ánanda(321)
zsemlék(283)	áfa-jának(7)	kalács(10658)	Avalokitésvara(39)
lepények(202)	mot-nak(5)	rizs(31678)	Dordzse(270)
pogácsák(539)	Villanysze(5)	zsemle(6690)	Babaji(82)
pékáruk(771)	oktávtól(5)	pogácsa(11066)	Bodhidharma(210)
péksütemények(997)	Isten-imádat(5)	sajt(46660)	Gautama(574)
sonkák(613)	Nagycsajszi(5)	kifli(9715)	Mahakásjapa(25)
tészták(2466)	-fontosnak(7)	krumpli(37271)	Maitreya(426)
kalácsok(277)	tárgykörből(5)	búzakenyér(306)	Bódhidharma(115)

1. táblázat. Egy gyakori és egy ritka szó legközelebbi szomszédai a `SURF` és az `ANA` modellből. A zárójeles számok a korpuszbeli előfordulások számát mutatják.

A korpusz morfológiailag annotált változatából **készítettünk egy lemmatizált modellt is (`LEM`)**, hogy a morfoszintaktikai és szintaktikai hasonlóságok

megragadása helyett inkább a szemantikai analógiák megragadására helyezzük a hangsúlyt, illetve hogy enyhítsük a fentebb említett ritka szavakkal kapcsolatos adathiányproblémát. A lemmákat is tartalmazó annotációt a magyar Humor mofológiai elemzőt [8,9] használó PurePos taggerrel [10] készítettük. Az annotációból csak a lemmákat hagytuk meg, és ezek sorozatán tanítottuk be a szóbeágyazási modellt.

A [11] cikkben leírt módon **létrehoztunk egy másik elemzett korpuszon alapuló modellt is (`ANA`).** Ebben a modellben nemcsak a lemmák szerepelnek, hanem minden eredeti tokent két token reprezentál: a lemmát a szó morfoszintaktikai címkéje külön tokenként követi.

Az alábbi példa a *Szeretlek, kedvesem.* mondat így előfeldolgozott változatát mutatja:

```
szeret [IGE][Ie1] , [PUNCT] kedves [FN][PSe1][NOM]
```

Mivel ebben a reprezentációban a címkéket a hozzájuk tartozó szó mellett tartottuk, a `LEM` modellel ellentétben a címkék által hordozott morfoszintaktikai információ továbbra is szerepet játszott a vektorreprezentáció meghatározásában. Ugyanakkor az adathiányt jól kezeli ez a reprezentáció, mert a szó különböző ragozott alakjait egyetlen lemma képviseli. Az 1. táblázat második két oszlopa az `ANA` modell által generált legközelebbi szomszédokat mutat be. Látható, hogy ez a modell a `SURF` modellnél jobban meg tudja ragadni a ritka lexikai elemek jelentését, mert míg a lemmatizálás enyhíti az adathiányproblémát, a környezetben jelenlévő morfoszintaktikai annotáció megőrzi a grammatikai információ túlnyomó részét (v.ö. a táblázat 2. és 4. oszlopát). A *Vakkali* szóhoz az `ANA` modellben legközelebbi szavak világosan jelzik, hogy ez a modell képes volt annak a ténynek a kihámozására a korpuszból, hogy egy buddhista személyiségről van szó.

Az `ANA` modell egy apró módosításával egy még **kifinomultabb elemzett modellt hoztunk létre (`POS`)**. A fő szófajcímkét a lemmán hagytuk, és csak a morfoszintaktikai címke többi részét választottuk le külön tokenként.

Az alábbi példa a *Szeretlek, kedvesem.* mondat így előfeldolgozott változatát mutatja:

```
szeret[IGE] [Ie1] ,[PUNCT] kedves[FN] [PSe1][NOM]
```

Ez a modell szófajonként külön reprezentációt hoz létre a különböző szófajú homonim szavakhoz (pl. *vár, nyúl, csűr, reggeli, ír* stb.), ezzel részben kezeli azt a problémát, hogy a szóbeágyazási modelleket létrehozó algoritmusok többsége nem tud értelmes módon mit kezdeni a homonímia/poliszémia jelenségével. A CBOW/skipgram modellek által létrehozott vektorterekben ugyan valamilyen értelemben sokszor visszanyerhető a több jelentést egybeejtő vektorreprezentációból a több jelentés, de hogy ez mennyire sikerülhet, az nagy mértékben múlik azon, hogy hány jelentése van az adott lexikai elemnek, és hogy azoknak a korpuszbeli jelenléte mennyire kiegyensúlyozott.

3. Kísérletek

Mivel – ellentétben az angollal – olyan magyar nyelvű erőforrások nem léteznek, amelyekben anyanyelvi beszélők jellemeztek volna szópárokat a jelentéshasonlóságuk mértéke szempontjából, és amelyeket a szóbeágyazási modellek kiértékeléséhez közvetlenül lehetne használni, ezért más módszert kellett alkalmaznunk. Egy olyan webes felületet hoztunk létre, melyen keresztül összehasonlító értékeléseket gyűjtöttünk a különböző modellekből kinyert és a felhasználó számára bemutatott példák alapján [5]. A kiértékelésben résztvevőket arra kértük, hogy rangsorolják a modelleket az alapján, hogy egy adott hívószóhoz kinyert kapcsolódó szavak listája mennyire releváns. Az `ANA`, `POS` és `LEM` modellek mellett a `SURF` modell egy módosított változatát (l. alább), egy szabadon hozzáférhető, nyers magyar nyelvű szövegeken tanított skip-gram modellt [12] és egy a magyar WordNeten alapuló modellt is bevettünk a kiértékelésbe.

A nyers korpuszból épített modellek (pl. `SURF`) a szavak toldalékolt alakjait is tartalmazzák, ezek megjelennek a lekérdezésekre kapott válaszokban is. Annak érdekében, hogy ezek a szólisták (amikben tehát egy szóhoz a legközelebbi szavak nagy része ugyanannak a lemmának a különböző alakjai) összehasonlíthatóak legyenek az előfeldolgozott korpuszból létrehozott, csak lemmákat tartalmazó modellből kinyert szólistákkal, a nyers modellek kimenetén utólag alkalmaztunk tövesítést. Az adott hívószóhoz az n legközelebbi szót pedig úgy adtuk meg, hogy az ismétlődő lemmák eliminálása után kapjunk n hosszú listát (`SURFL`). A [6] cikk szerzői hasonló transzformációt alkalmaztak a nyers és lemmatizált modellek összehasonlításához.

Az általunk létrehozott modelleket egy 1,2 milliárd szavas magyar webkorpuszból építettük. A referenciaként használt szabadon letölthető modell egy jóval nagyobb, 4,6 milliárd szavas nyers korpuszból készült [12]. Míg a mi modelljeink tanítása során a CBOW architektúrát alkalmaztuk, addig a nagyobb korpuszból épült modell létrehozása a skip-gram architektúrával történt, és 200 dimenziós vektorokat tartalmaz. Ennek a modellnek szintén az utólag tövesített eredményét vettük be a kiértékelésbe (`SGL`).

A magyar WordNet-et, mint kézzel ellenőrzött szemantikai adatbázist vettük be a kiértékelésbe. [6] alapján minden tesztszóhoz kinyertük azokat a szavakat, amik valamilyen WordNet reláció mentén három lépésen belüli távolságban voltak az adott szótól. Ezt a listát a távolság alapján rendeztük sorba úgy, hogy a vizsgált szó szinonimáit 0 távolságra lévő szavaknak tekintettük.

3.1. A lekérdezések

Schnabel és munkatársai egy olyan 100 szóból álló angol szólistát állítottak össze angol nyelvű szóbeágyazási modellek kiértékeléséhez, ami a lekérdezések kiindulásaként használt szavakat korpuszbeli szógyakoriság, a jelentés absztraktsága, illetve szófaj szempontjából kiegyensúlyozott arányban tartalmazza [5].[1] Ezt a

[1] Az erőforrás elérhető online a `http://www.cs.cornell.edu/~schnabts/eval/` oldalon.

szólistát magyarra fordítottuk úgy, hogy minden szónak a megfelelő szófajú és ugyanabba a jelentéskategóriába tartozó megfelelőjét választottuk, majd ellenőriztük, hogy a magyar nyelvű korpuszban hasonló gyakorisággal szerepel-e. Ha a fordítás nem ugyanabba a tartományba esett, akkor egy másik, hasonló jelentéscsoportba tartozó, ugyanolyan szófajú, de megfelelő gyakoriságú szót választottunk a korpuszból. Ezáltal egy, az eredeti szólistához hasonlóan kiegyensúlyozott magyar nyelvű listát kaptunk. Ezt a listát kiegészítettük még 7 további olyan homonim szóval, amiknek a különböző jelentései különböző szófajúak (pl. *vár*, *reggeli*).

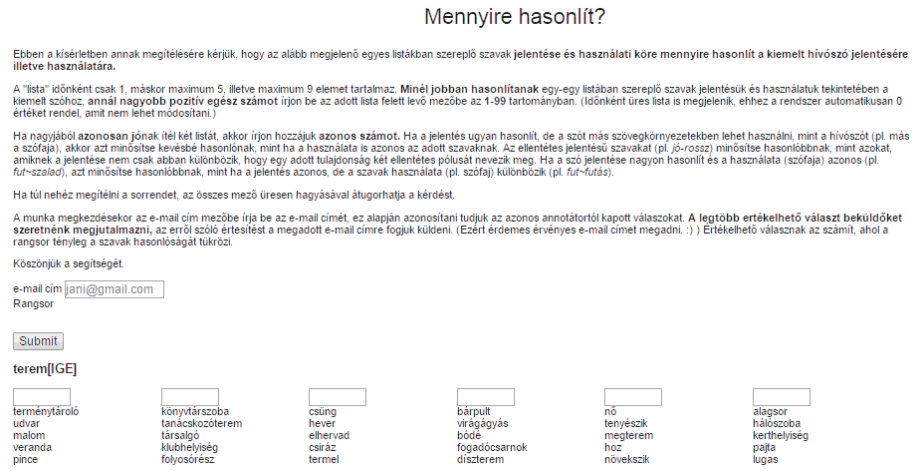
3.2. A rangsorolási feladat

Magyar anyanyelvű annotátorokat kértünk meg arra, hogy a létrehozott webes felületet használva rangsorolják a hat modell által generált válaszokat (`ANA`, `SURFL`, `POS`, `LEM`, `SGL`, `WN`). A kiértékelésben résztvevő felhasználóknak minden körben az előre összeállított listából egy hívószót és az egyes modellekből 1, 5, vagy 9 kapcsolódó szót mutattunk meg. A modellek neve természetesen rejtve volt a felhasználók elől, és a megjelenítés sorrendje is véletlenszerűen változott. Az 1, 5, vagy 9 elemű megjelenített szólisták az egyes modellek szerint a vizsgált szó k legközelebbi szomszédját tartalmazó listák 1, 5, illetve 9 egymás utáni elemét tartalmazó részei voltak, vagy a távolság szerint rendezett lista első elemétől, vagy a 30. elemétől kezdve (egy konkrét kérdés esetén mindig ugyanabból a tartományból). Ez alól egy esetben tettünk kivételt: mivel a WordNet-ből létrehozott modellben a kapcsolódó szavak száma a 107 vizsgált szóból 97 esetben 30-nál kevesebb volt, ezért az olyan kérdéseknél, amikor a többi modellből a 30. pozíciótól számítva jelenítettük meg a kapcsolódó szavakat, a `WN` modellt kihagytuk a rangsorolási feladatból. A magyar WordNet fedése egyébként is elég alacsony volt, a kísérlet során vizsgált 107 tesztszó közül csupán 70 szerepelt benne (65%). Ha egy hívószó esetén valamelyik modell nem adott választ, akkor üres listát jelenítettünk meg a felhasználó számára.

Az annotátoroknak a megjelenített listákat rangsorolni kellett az alapján, hogy mennyire érezték a hívószó és az egyes listák elemei között a kapcsolódást. A rangsoroláskor 1 és 99 közötti pontszámot adhattak (minél magasabb, annál jobb a lista minősége), az üres listákhoz automatikusan 0 pontot rendeltünk és megengedtük az egyenlő pontszámokat is.

Mivel a hasonlóság mértékének meghatározása szubjektív feladat, ezért ha egy annotátor nem tudott egy kérdésben rangsort felállítani (az is előfordulhatott, hogy a ritka szavak esetén a hívószó jelentését sem ismerte), akkor válaszadás nélkül továbbléphetett a következő kérdésre. A válaszok összesítésekor az egyes listákhoz rendelt pontszámok értékét figyelmen kívül hagytuk, csupán a rangsort vettük figyelembe. A kiértékeléshez használt felület és az annotátorok számára adott instrukciók az 1. ábrán láthatók.

A válaszok kiértékeléséhez a nyerési arány metrikát használtuk. Ez azt fejezi ki, hogy egy adott modell az esetek mekkora részében kapott magasabb pontszámot egy másik modellnél. Ez tehát a modellek páronkénti összehasonlítását jelenti minden egyes tesztkérdés alapján a felhasználók által adott rang-



1. ábra. A kiértékeléshez használt felület

sor szerint. Minden „győzelem” egy pontot ér, a végső pontszám pedig a annak az aránya, hogy az összes összehasonlításból hányszor került ki győztesként az adott modell. Ezt a módszert alkalmazzák gépi fordító rendszerek emberi kiértékelése során is, mivel ez bizonyult a legmegbízhatóbb módszernek arra, hogy egymástól egészen különböző rendszerek minősége összemérhető legyen [13]. Az összesítő kiértékelést minden paraméter szerint külön-külön is elvégeztük, így a modellek minőségét a hívószavak gyakorisága, szófaja, szemantikai kategóriája, és a válaszlisták kezdőpozíciója (1 vagy 30) szerint is vizsgálni tudtuk.

Bár a felkért annotátorok magyar anyanyelvű felhasználók voltak, a valódi válaszadás megkezdése előtt egy tesztfázison kellett átesniük. A 107 elemű listán felül kiválasztottunk 5 olyan szót, amire a modellek válaszai egyértelműen rangsorolhatók, így egy gold standard-nek tekinthető választ várhattunk ezekre. Ezt az 5 szót minden felhasználónak az első 10 kérdés között megjelenítettük. Ha egy annotátor ebből az 5 kérdésből legalább 4-re nem tudta a helyes sorrendet felállítani, akkor az ő eredményeit nem vettük figyelembe a kiértékelésnél.

A kísérlet során végül 15 annotátortól kaptunk használható válaszokat, akik mind legalább 14, legfeljebb 102 különböző kérdésre adtak választ. Minden egyes tesztszóra legalább 3-3 rangsorolás érkezett, mind az 1., mind a 30. pozíciótól számított listák alapján.

4. Eredmények

A kiértékelés eredménye az 2. ábrán látható. Az egyik legegyértelműbb tény, ami az ábrán jól látszik, hogy a referenciaként használt 200 dimenziós vektorokból álló skip-gram modell (SGL) teljesített legrosszabbul szinte mindegyik kiértékelési szempont esetén ($p < 0.05$). Ez alól kivétel csak a melléknevekre és a nagyon

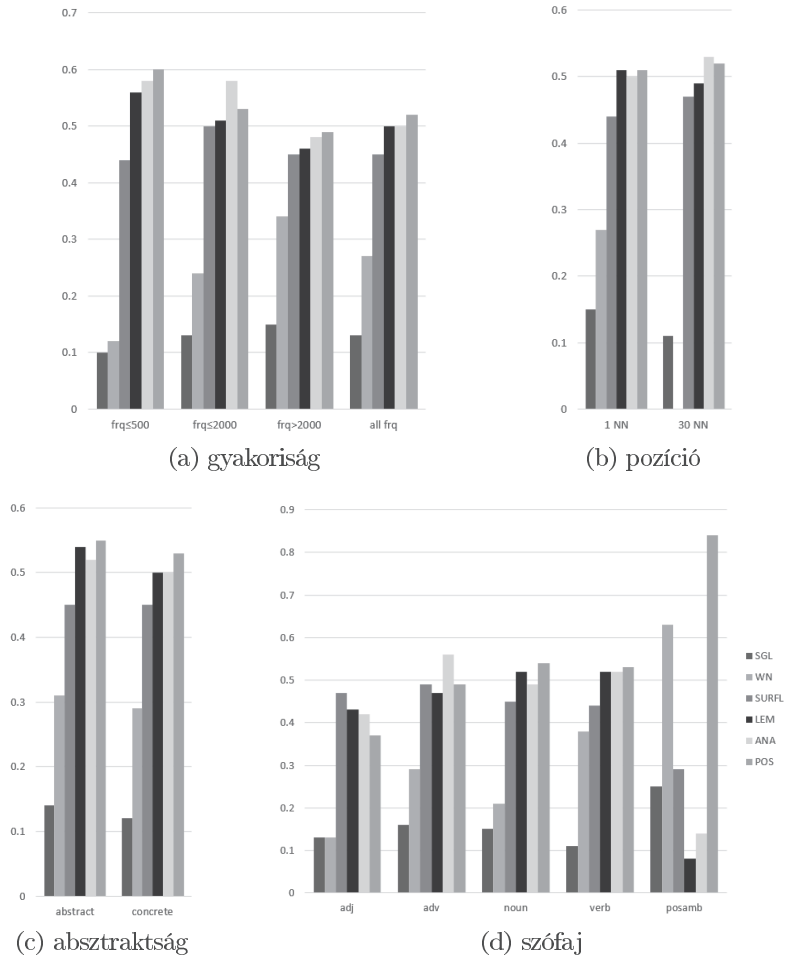
ritka szavakra (gyakoriság ≤ 500) szűkített eredmény, ahol a `WN` modell csaknem ugyanolyan rossz eredményt produkált (az alacsony fedés miatt), illetve a szófajon átívelő többértelműség esete (lásd a 2d. ábra elemzését később). Az `SGL` modellből a gyakori szavakhoz kapott válaszlisták is sok olyan szót tartalmaztak, amik nem, vagy nagyon gyengén kapcsolódnak a hívószóhoz (pl. a *kenyerek* szó esetén a *búzalisztből, sütéshez, karfiolból* szavak), a ritkább szavakhoz pedig sokszor teljesen értelmetlen eredmények szerepeltek a listában.[2]

A modellek összesített teljesítménye a,z 2a ábra *all frq* oszlopában látható. Mindegyik általunk készített modell jobban teljesített, mint a magyar WordNet. Ez megerősítette azt a feltevésünket, hogy bár az angol szakirodalomban népszerű a WordNet-hez viszonyított kiértékelés, a magyar WordNet azonban nem alkalmas arra, hogy a beágyazási modellek kiértékelésének alapjául szolgáljon: a fedése ehhez nem elég jó.

Ahogy a 2a. ábrán látható, az előfeldolgozott és lemmatizált korpuszból épített szóbeágyazási modellek jobban teljesítettek a ritka szavak esetén, míg a gyakori szavak esetén kevésbé érvényesült az előfeldolgozás hatása.

A listán belüli kezdőpozíciót, mint paramétert megfigyelve (2b. ábra) az látható, hogy az `SGL` modell teljesítménye tovább romlik, ha a kapcsolódó szavak listájában hátrébb szereplő elemeket vizsgáljuk. Megfigyelhető továbbá, hogy az `ANA` és a `POS` modellek minősége a 30. pozíciónál jobb, mint a `LEM` modell minősége. Míg a lemmatizált modell jól kezeli az adathiány problémáját (lásd a ritka szavakra mutatott teljesítménynövekedést a nem lemmatizált modellekhez képest), ebben a modellben a szemantikai hasonlóság elnyomja a grammatikai szempontból való hasonlóság szempontjait. A `LEM` modellből kapott válaszlisták gyakran változatos szófajú elemeket tartalmaznak. Az `ANA` modell ezzel szemben megőrzi a grammatikai (szintaktikai) relációkat is, mivel az előfeldolgozás során megtartottuk a morfoszintaktikai tulajdonságokat kódoló címkéket a szavak környezetében. Ezért a létrejött szóbeágyazási vektorok tanítása során ez az információ is érvényesülhetett. A `POS` modell ezen felül a szófaj alapján megkülönböztethető homonim szóalakok között is különbséget tud tenni. Ezek a modellek tehát egy jó minőségű köztes megoldást biztosítanak a nyers korpuszból tanított szóalapú modellek (ahol a morfoszintaktikai viszonyok ugyan jól megjelennek, de a ritkább szavak esetén az adathiány rontja a modell minőségét) és a szótő alapú modellek között (ahol szinte kizárólag szemantikai szempontok érvényesülnek). Fontos megjegyezni, hogy az általunk létrehozott modellek tanítása azonos paraméterekkel történt, ez mégsem ugyanazt jelenti, hiszen az `ANA` és a `POS` modellek esetén a környezetet figyelembe vevő ablakban valójában fele annyi szó szerepel, hiszen minden szót két tokenként reprezentáltunk a korpuszban. A fele akkora környezet ellenére a modellek minősége nem rosszabb, mint a nagyobb környezetet figyelembe vevő `LEM` és `SURF` modelleké.

[2] Mivel ezt a modellt nem mi készítettük, nem egyértelmű, hogy pontosan milyen tényezők játszhattak szerepet abban, hogy ennyivel gyengébb teljesítményt nyújt a többinél. Nem valószínű, hogy az lenne az oka, hogy a CBOW helyett a skipgram architektúrát használták a modell készítésekor.



(a) gyakoriság (b) pozíció

(c) absztraktság (d) szófaj

2. ábra. A modellek teljesítménye (nyerési arány) különböző szempontok szerint vizsgálva

Ami a jelentés absztrakt vagy konkrét voltát illeti, a modellek teljesítményében ez a szempont nem nagyon játszik szerepet (2c. ábra), hasonlóan teljesítenek mindkét kategóriában. A szófajok szerinti bontásban azonban vannak különbségek (2d. ábra). Egyértelműen látszik, hogy az elemzett korpuszból tanított modellek jobban teljesítenek igék és főnevek esetén. Ez nem meglepő, hiszen ezek a leggyakrabban toldalékolódó szófaji kategóriák. A melléknevek esetén azonban a nyers szövegen tanított modell teljesített jobban.

A homonim alakokra a `POS` modell – mint az várható volt – a mezőnyből messze kiemelkedő eredményt nyújtott (2d. ábra *posamb* oszlopa). Ebben a kategóriában a `WN` modell lett a második legjobb, ami annak köszönhető, hogy a

WordNet-ben (a POS modellhez hasonlóan) szerepel szófaji információ, tehát az azonos alakú, különböző szófajú szavak jól megkülönböztethetők. A WordNet alacsony fedése miatt azonban a POS modell jobban teljesített. Érdekes továbbá, hogy ez volt az egyetlen olyan szempont, amely szerint a LEM és ANA modellek teljesítménye gyengébb volt, mint a nyers korpuszból tanított modelleké (SGL és SURFL).

5. Konklúzió

Cikkünkben egy magyar nyelvű szóbeágyazási modellek kiértékelésére irányuló kísérlet eredményeit foglaltuk össze. A kiértékelésben 6 modellt hasonlítottunk össze, melyek közül négynek a létrehozását is részletesen bemutattuk, a különböző előfeldolgozási lépésekkel együtt. A további két modell közül az egyik egy mások által más anyagon és más architektúrán betanított, általunk csak a kiértékelésben felhasznált szóbeágyazási modell, a másik pedig a magyar WordNeten alapuló modell volt. A modellek kiértékelését egy webes felületen keresztül humán annotátorok által adott rangsorok alapján végeztük.

Az eredmények alapján általánosságban elmondható, hogy a morfológiai annotációt tartalmazó korpuszból tanított modellek jobban teljesítettek, mint a nyers korpuszból tanított modellek, enyhítve az agglutinációból fakadó adatritkaság által okozott problémákat. A bemutatott ANA és POS modellek az egyszerűen lemmatizált modelleknél is jobb teljesítményt nyújtottak, annak ellenére, hogy a tanítás során kisebb kontextust vettek figyelembe a szóvektorok létrehozásakor, mert ezeknek a modelleknek a betanításakor a morfoszintaktikai információ jelen volt az adott lexikai elem környezetében, így a grammatikai hasonlóságok megragadására alkalmasabbak a pusztán lemmatizált modellnél.

Köszönetnyilvánítás

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

Hivatkozások

1. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
2. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. (2013) 746–751
3. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, Association for Computational Linguistics (2014) 238–247

4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. (2013) 3111–3119
5. Schnabel, T., Labutov, I., Mimno, D.M., Joachims, T.: Evaluation methods for unsupervised word embeddings. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y., eds.: EMNLP, The Association for Computational Linguistics (2015) 298–307
6. Ebert, S., Müller, T., Schütze, H.: LAMB: A good shepherd of morphologically rich languages. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA (2016)
7. Miháltz, M., Hatvani, C., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian WordNet project. In: Proceedings of The Fourth Global WordNet Conference. (2008) 311–321
8. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
9. Novák, A.: A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 1068–1073 ACL Anthology Identifier: L14-1207.
10. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, Incoma Ltd. Shoumen, Bulgaria (2013) 539–545
11. Siklósi, B.: Using embedding models for lexical categorization in morphologically rich languages. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, Springer International Publishing, Cham. (2016)
12. Szántó, Z., Vincze, V., Farkas, R.: Magyar nyelvű szó- és karakterszintű szóbeágyazások. In Tanács, A., Varga, V., Vincze, V., eds.: XIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2017) 323–328
13. Bojar, O., Ercegovčević, M., Popel, M., Zaidan, O.F.: A grain of salt for the wmt manual evaluation. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. WMT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1–11