SEGÍTHETNEK-E A SZÓBEÁGYAZÁSI MODELLEK A TÁRSADALOMTUDÓSOKNAK?

Novák Attila[1], Siklósi Borbála[1], Prószéky Gábor[1,2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
[2] MTA Nyelvtudományi Intézet

Bevezetés

- disztribúcióalapú szójelentés-reprezentációs modellek
  - szavak jelentése
  - grammatikai és szemantikai dimenziók
  - szociolektális dimenziók

→ gazdag tudásforrást jelenthetnek azoknak, akik számára a szövegek alapvető nyersanyagként szolgálnak

- néhány példa társadalomtudósok számára...

Disztribúciós modellek

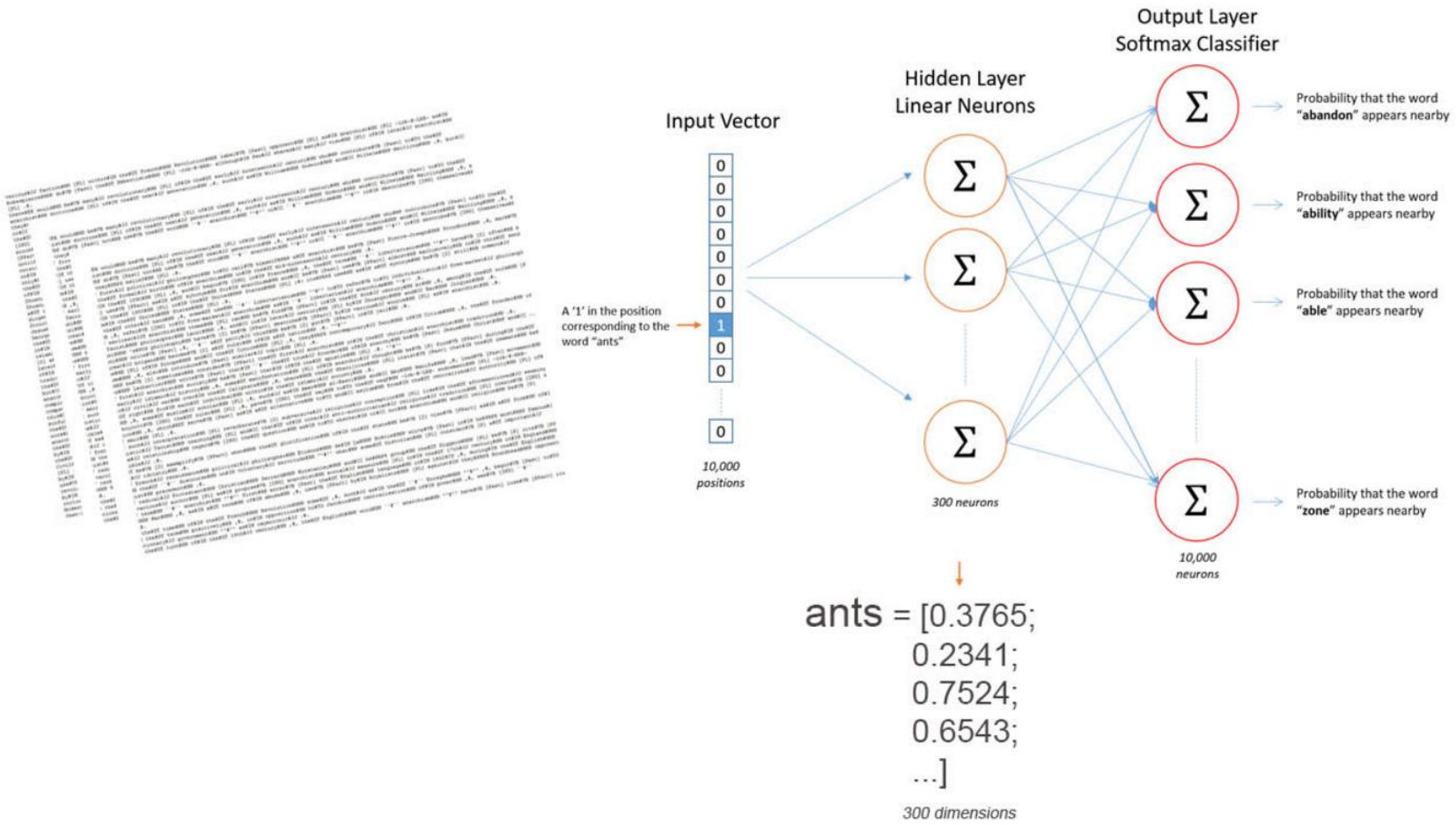
- strukturalista nyelvészek: a nyelvi tudás elsődleges forrása a szavak és morfémák disztribúciója

- disztribúciós szemantika: a szavak jelentése szorosan összefügg azzal, hogy milyen kontextusokban használjuk őket

Disztribúciós modellek

- számítógépes megvalósítás
  - hagyományos: az adott szó előre meghatározott méretű környezetében előforduló szavak egy nagy korpuszból számított előfordulási statisztikája
    - nagyon sok dimenziós, ritka reprezentáció
  - modern: mesterséges neurális hálózatok
    - hatékony gépi háttér
    - folytonos vektortérbeli tömör reprezentáció

Disztribúciós modellek



Disztribúciós modellek

- nagy méretű szöveges korpusz (előfeldolgozás)
- néhányszáz dimenziós vektorok
- az egyes dimenzióknak nincs saját jelentésük
- a lexikai elemek egy valós vektortér pontjai
  - az egymáshoz szemantikailag és/vagy morfológiailag hasonló szavak egymáshoz közel, a jelentésben eltérő elemek egymástól távol esnek
  - vektoralgebrai műveletek

A korpusz előkészítése

- több mint 1 milliárd szavas webkorpusz
- automatikusan egyértelműsített morfológiai elemzés
- szótövek és címkék

```
A török megszállás nem feltétlenül jelentette a népesség pusztulását.

a [Det] török [Adj] megszállás [N] nem [Neg] feltétlenül [Adv] jelent
[V.Past.3Sg.Def] a [Det] népesség [N] pusztulás [N.Poss3Sg.Acc] . [.]
```

- grammatikai és szemantikai hasonlóság
- stiláris, szociolektális dimenziók


pógár
semmittevő
köpködő
napszámos
lumpenproli
ingyenélő
szoci
zember
komcsi
burzsuj
proli
komenista
tróger
panelproli
álrokkant
bolsi
fülkeforradalmár
csicska
kommer
kommancs
panellakó
posztkomcsi
nyugdíjas
cigó
segélyes
fidesznyik
jobbér
macskajancsi
segélyes
kisnyugdíjas
libsi
hülyegyerek
melós
gyurcsányista
szadeszos
maszopos
nyugdijas
nyögdíjas
nyugdíjjas
fityeszes
bértollnok
szájtépő
közszolga
pártaktivista
szakszervezetis

Nyelvi rétegződés

- a modellből lekérdezhető a benne szereplő szavakhoz legközelebb elhelyezkedő további szavak listája (távolság szerint)
- egy-egy szóból kiindulva feltérképezhető egy adott régió szókincse
- olyan típusú kategóriák is elkülönülnek, amilyen típusú megkülönböztetés semmilyen létező szótárban nem szerepel

kempel	ficc	macula	balíz
wowozik	fic	sárgafolt	balízcsoport
farmol	fici	degeneratio	vezérlőjel
fearless	fanfic	atrophia	főjelző
healel	törid	glaukóma	transzponder
VF-ezik	ficu	látóidegfő	vágányút
hackel	drarry	szürkehályog	vezérlőegység
maxol	fanfiction	makula	EVC
castol	sztory	ideghártya	jelsorozat
turret	snarry	látóhártya	menetengedély
leöl	SSHG	zöldhályog	kijelzés
sentry	oneshot	centralis	DMI
questel	feji	látóideg	vezérlőközpont
betámad	függővég	glaucoma	riasztóközpont
lewarezol	manga	naevus	komparátor
limpel	dorama	erythema	nyugtázás


Szebasztián
Márk
Dániel
Dominik
Rajmund
Kristóf
Miklós
Patrik
Levente
Bence
Renátó
András
Gergő
Imre
Zoltán
Csaba
Ákos
Lajos
Bendegúz
Balázs
Krisztián
László
Ede
Krisztofer
István
Béla
Aladár
György
Ödön
Jenő
Ádám
Alajos
Mihály
Márió
Ernő
Emil
Nándor
Márton
Rezső

Doménadaptáció és -szelekció

- nagy mennyiségű tanítóanyag szükséges jó minőségű modellek létrehozásához
- egy adott réteg vagy szaknyelv vizsgálatához a nagyobb általános korpuszból létrehozott modellből kiindulva a rendszert a szaknyelvi korpuszon tovább tanítva a köznyelvben dominánsan az adott rétegnyelvtől eltérő jelentésben használt szavak reprezentációja a rétegnyelvben domináns jelentéshez közelít

Matematikai transzformációk a vektortéren

- ellentétes polaritású elemek szétválasztása
- homonim alakok kezelése

Többnyelvűség

- a különböző nyelveken készített szóbeágyazási modellek topológiája általában hasonló
- néhány ezer fordítási szópár megadásával viszonylag pontos leképezés definiálható két különböző nyelvhez készült modell között
- „rokon” lexikális mezői közötti leképezés
- lehetővé teszi az egyik oldalról kiindulva a másik oldal felfedezését

busó	pörc	cigó
reveler	bacon	thug
reveller	dough	strikebreaker
parade	sauce	racist
re-enactor	sliced	troublemaker
clown	gravy	Palestinians
townspeople	soup	rioter
carnival	curd	hoodlum
festival-goer	steak	Tutsis
townsfolk	stew	Jew
villager	pastry	Arab
onlooker	tortilla	bigot
festivity	lard	whites
mummer	butter	fascist
maypole	flatbread	drunk
procession	mayonnaise	bookie

Összefoglalás

- nagyméretű korpuszokból neurális hálózatok segítségével épített szóbeágyazási modellek
- a szövegekre alapozott kutatásokat végző társadalomtudósok számára
- nyelvi szinten tetten érhető tudás megragadható