Dimenzija spola v računalniškem jezikoslovju

Senja Pollak, Inštitut “Jožef Stefan”

Prezrte dimenzije spola v znanstvenih raziskavah
9. mar 2020
ZRC SAZU,

Računalniško jezikoslovje

Področje *računalniškega jezikoslovja* oz. *procesiranje naravnega jezika* se ukvarja z računalniškimi metodami za:

- razumevanje, analizo in tvorjenje jezika.

Primeri nalog:

- označevanje besedilnih korpusov
- luščenje informacij
- strojno prevajanje
- klasifikacija in iskanje dokumentov
- avtomatska izdelava povzetkov
- pripisovanje avtorstva, profiliranje avtorjev besedil
- analiza sentimenta
- ...

Zelo produktivno znanstveno in aplikativno področje

Jezikovni korpusi

- čas “velepodatkov”
- velike količine besedil v elektronski obliki, vsako leto se skoraj podvojijo
- velik napredek v računalniskem jezikoslovju
- **Jezikovni korpusi** (knjige, časopisi, znanstveni članki, socialni mediji) omogočajo kvantitativne analize povezave med jezikom in spolom, vsebujejo in odsevajo tudi družbene neenakosti, stereotipe in pristranosti, tudi v povezavi s spolom.

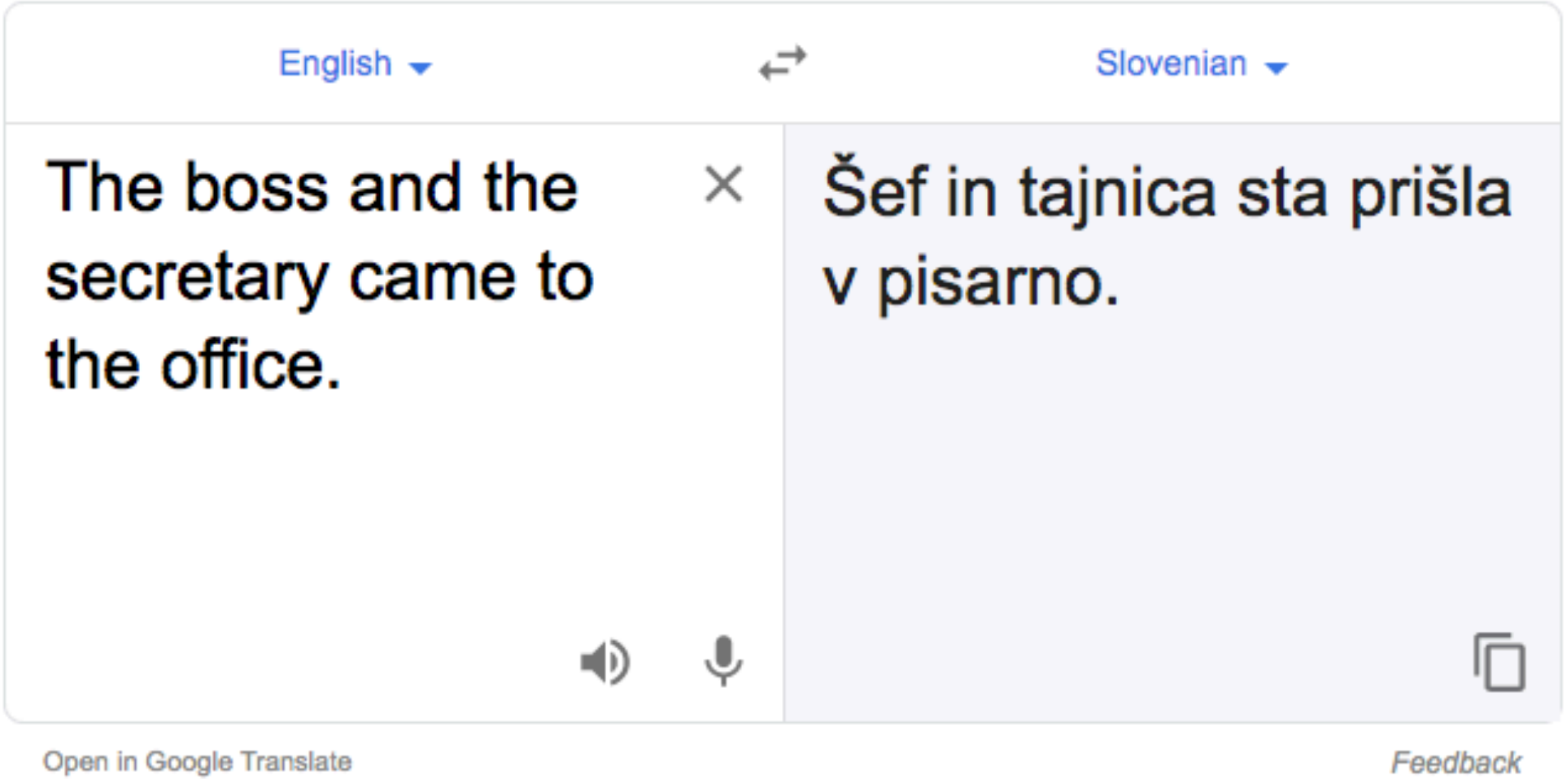
Računalniški modeli in jezikovni korpusi:

- analiza razlik v povezavi s spolom
- reproduciranje pristranosti v računalniških modelih

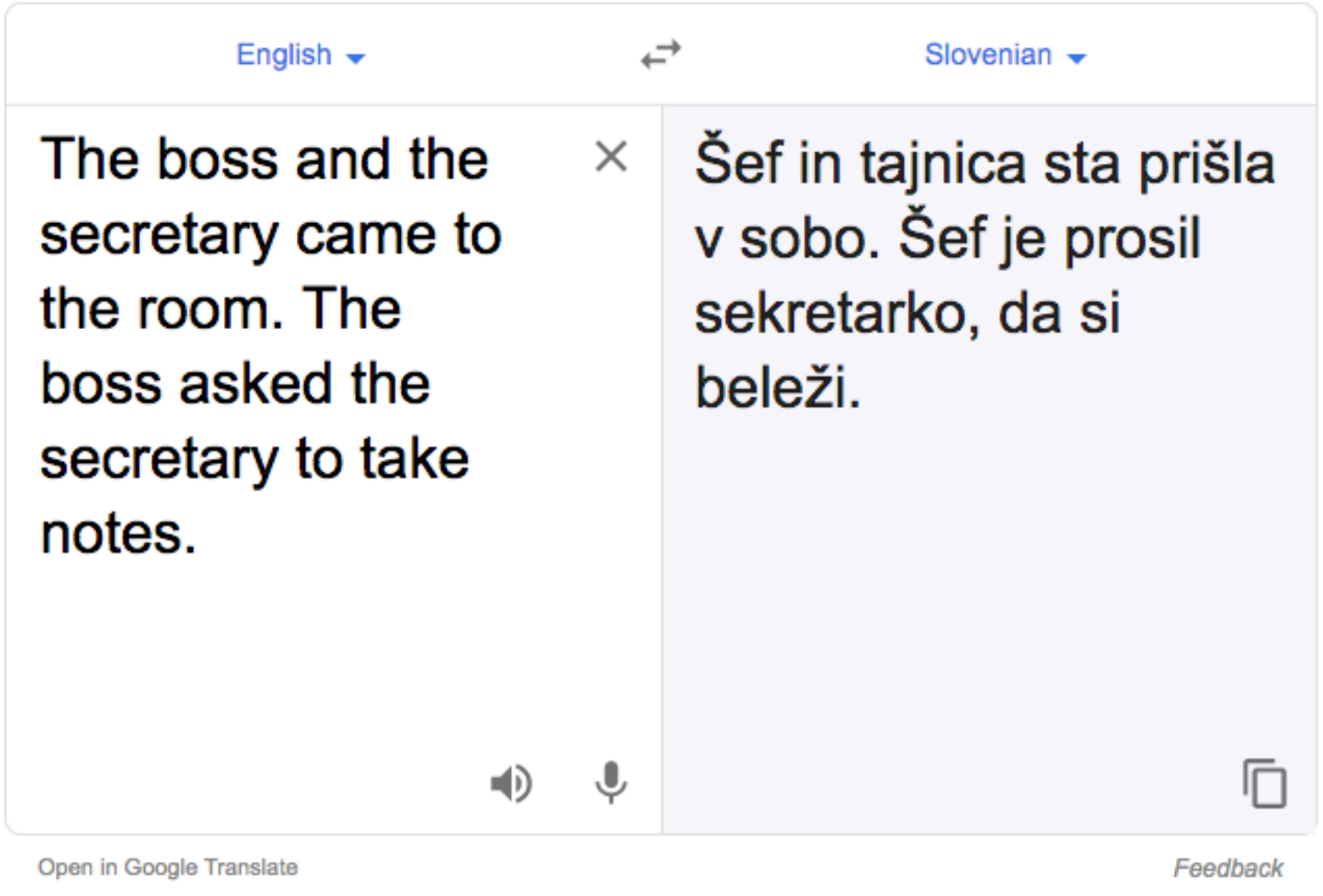
Profiliranje avtorjev besedil

- Računalniško razpoznavanje spola: PAN shared tasks, profiliranje za slovenščino (npr. Verhoeven et al. 2017: tviti, Škrjanec et al. 2018: blogi)
- Algoritmi dosežejo točnost okoli 90%
- Sociolingvistični izsledki:
  - slovnična raven (npr. ženska oblika deležnika na -l, predvsem v samonanašalnih kontekstih (sem rekla))
  - tematska raven (šport in alkoholne pijače so značilne teme za moške)
  - slogovna raven (npr. moški uporabljajo več vulgarizmov, številk in simbolov, žensk e pa več zaimkov in znakov za izražanje čustev, emotikonov).
- Zasnova raziskovalnih vprašanj, interpretacija, etc.

Google translate



Google translate



Besedne vektorske vložitve

- Besedne vektorske vložitve – vektorske reprezentacije besed v velikem večdimensionalnem prostoru, naučene z algoritmi (word2vec, GloVe, FastText...)

"You shall know a word by the company it keeps"

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In Studies in Linguistic Analysis, p. 11. Blackwell, Oxford.

	aardvark	computer	data	pinch	result	sugar	…
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	

sugar, a sliced lemon, a tablespoonful of **apricot** jam, a pinch each of,
their enjoyment. Cautiously she sampled her first **pineapple** and another fruit whose taste she likened
well suited to programming on the digital **computer**. In finding the optimal R-stage policy from
for the purpose of gathering data and **information** necessary for the study authorized in the

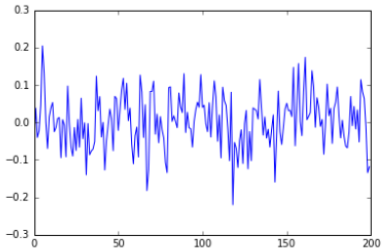
Besedne vektorske vložitve

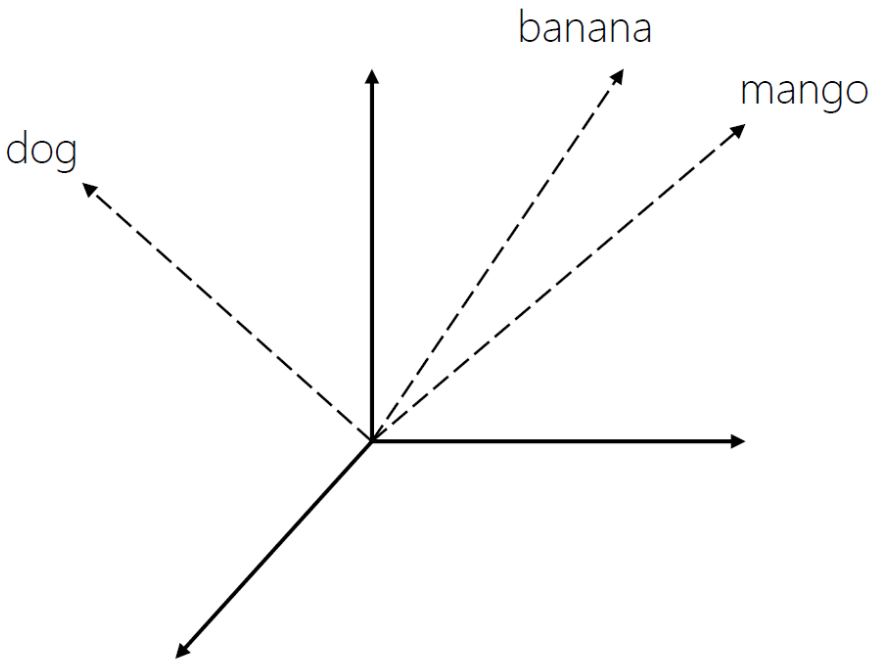
Dense. Dim = 200 (for example)

```
In [67]: print(vec['banana'])
         plt.plot(vec['banana'])
```

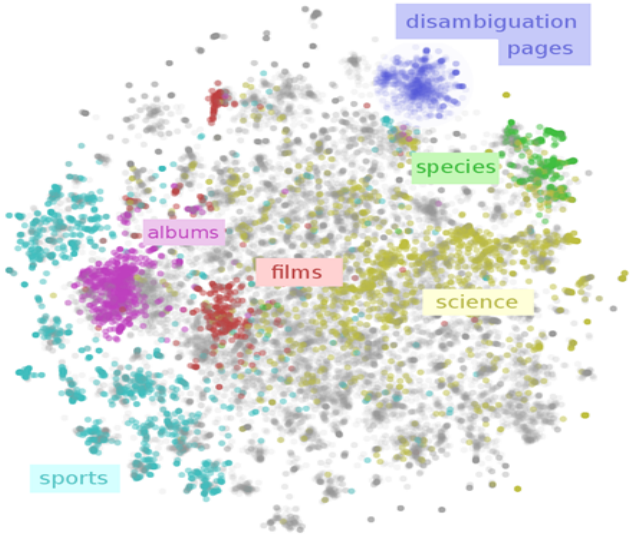
```
[-0.065091, 0.037847, -0.040299, -0.022862, 0.046481, 0.204306, 0.132157, 0.000275, -0.069716, 0.014626, 0.038425, 0.053029, -
0.024947, -0.013991, 0.010317, 0.012735, -0.094237, 0.007101, -0.007268, -0.091869, 0.097138, -0.002357, -0.065102, -0.089856,
 -0.013727, -0.074923, 0.007938, -0.066188, 0.064525, -0.0436, -0.001177, -0.140017, -0.003096, -0.086315, -0.0763, -0.071214,
 -0.051458, 0.123467, 0.031151, 0.068839, -0.039029, 4e-06, -0.127185, -0.049415, -0.007708, 0.035502, 0.009538, -0.075545, 0.0
69583, 0.062794, -0.021556, 0.031155, 0.087352, 0.117663, 0.034883, 0.104613, 0.004534, 0.037999, -0.058016, -0.110679, -0.0353
5, -0.012488, -0.0924, 0.126315, 0.080949, -0.040334, 0.047046, -0.182169, -0.1268, 0.082376, 0.082963, 0.110073, -0.031732, 0.
022219, -0.054332, 0.015394, -0.019853, -0.04169, -0.106969, -0.134253, 0.093094, 0.094716, 0.002643, 0.017417, 0.00309, -0.014
145, 0.078464, 0.041464, 0.026328, 0.12988, -0.02715, 0.027002, -0.014312, -0.017305, -0.066002, 0.002747, 0.033995, 0.053829,
 0.040628, 0.127369, 0.040216, 0.045803, -0.003395, -0.024843, 0.052411, -0.039267, 0.043378, 0.110868, 0.067947, -0.050505, 0.
019753, -0.094825, 0.094058, 0.057547, 0.045447, -0.016258, -0.102323, 0.080506, -0.219969, -0.053595, -0.069609, -0.120579, -
0.048799, -0.019837, -0.109987, -0.002571, 0.031825, -0.124037, -0.024646, -0.102276, 0.038512, 0.035166, 0.031713, 0.008979,
 0.114415, 0.0421, -0.034152, 0.014497, -0.04199, -0.018534, -0.065822, -0.020059, 0.019861, -0.159393, -0.03374, 0.083666, -0.
025234, -0.058921, -0.014924, 0.035292, 0.050979, 0.031609, 0.0322, 0.015638, 0.146793, -0.062475, 0.042192, 0.157084, 0.00237
1, -0.035507, 0.08275, 0.173776, 0.007175, 0.016044, 0.025942, 0.137863, 0.094541, -0.013125, 0.065621, 0.040823, -0.010574, 0.
007796, -0.085031, -0.003617, 0.102267, 0.018047, 0.037613, -0.056187, 0.036693, 0.053867, 0.094616, 0.015941, -0.041536, 0.005
796, -0.03694, -0.063241, -0.067796, -0.026023, 0.069142, -0.008786, 0.042428, -0.017718, 0.03318, -0.052277, 0.114012, 0.08154
2, 0.063282, -0.012149, -0.134274, -0.118431]
```

```
Out[67]: [<matplotlib.lines.Line2D at 0x12a60774e48>]
```





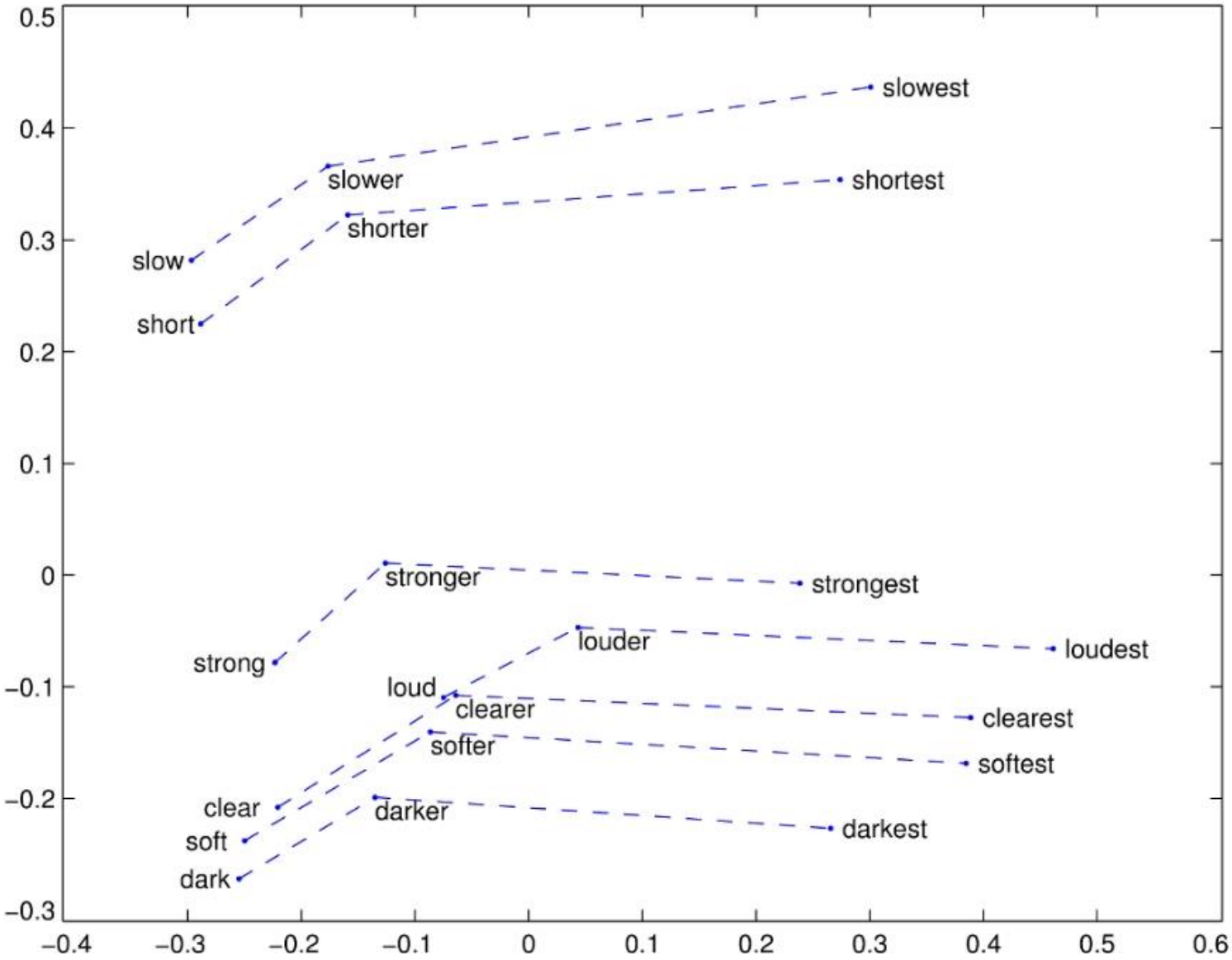
Besedne vektorske vložitve

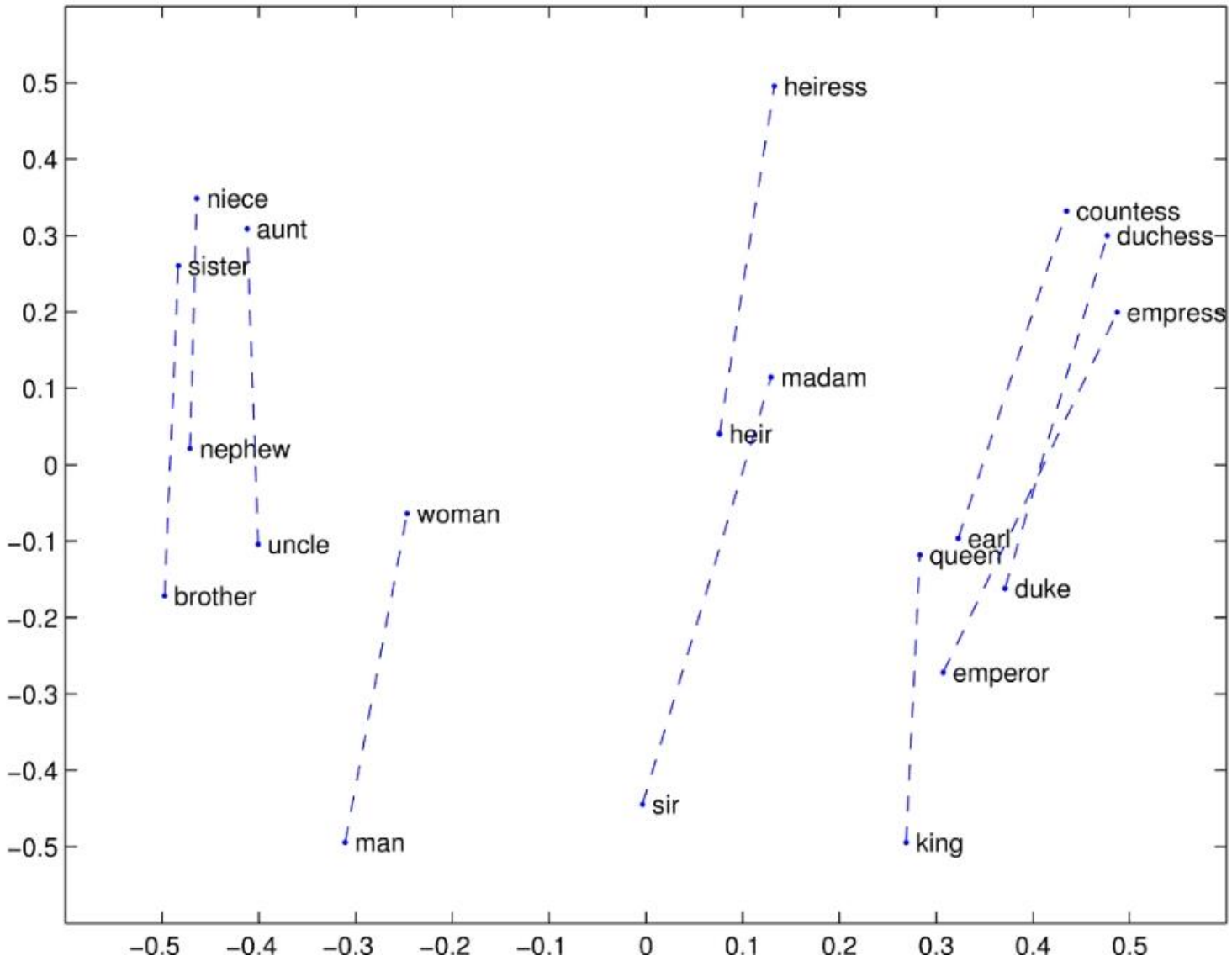


Delno ohranjajo semantične relacije

- Analogije

Pariz – Francija ≈ Berlin– Nemčija


slowest
slower
shortest
shorter
slow
short
stronger
strongest
louder
loudest
strong
loud
clearer
clearest
softer
softest
clear
darker
darkest
soft
dark
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6


0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
niece
aunt
sister
nephew
uncle
brother
woman
man
heiress
madam
heir
sir
countess
duchess
empress
earl
queen
duke
emperor
king

Raziskave na področju spolne pristranosti v besednih vložitvah

1. **Pristranost v besednih vektorskih vložitvah**
2. **Analiza družbe z besednimi vektorskimi vložitvami**
3. **Vpliv pristranosti na orodja**

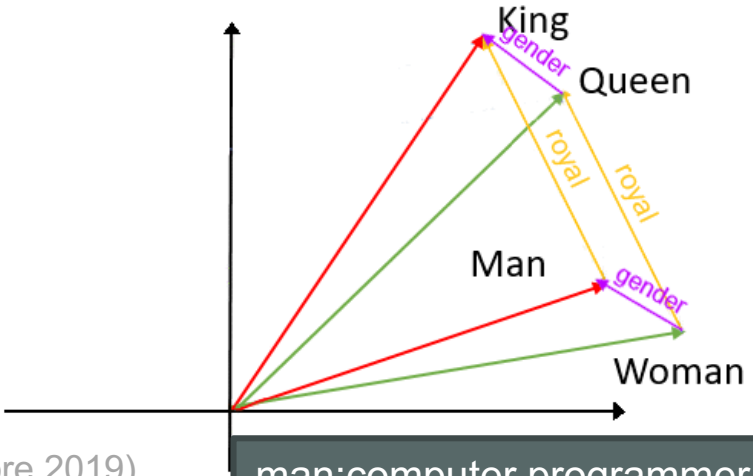
1) Pristranost v vložitvah

Analogije in pristranost

Bolukbasi et al. 2016: Man is to Computer programmer as woman in to Homemaker

Učenje na korpusih (Google News) - odsevanje in prenašanje pristranosti in stereotipov

$$X = woman + king - man \approx queen$$ man:king :: woman:x



(Buonocore 2019)

man:computer programmer :: woman: X , X= **homemaker**

father : doctor :: mother : X, X=nurse

Vložitve, analogije in pristranost

Bolukbasi et al. 2016: Man is to Computer programmer as woman in to Homemaker

Učenje na korpusih - odsevanje in prenašanje pristranosti in stereotipov

- Spolno nevtralne besede (npr. *nurse, boss*)
- Spolno zaznamovane besede (*brother, sister*)

- Spolno nevtralne besede – vseeno zaznamovane - problem pristranosti, kar vpliva na
- uporabo algoritmov v vrsti nalog, podobni rezultati z MTurk
- **brother : man :: sister : woman** - OK
- **doctor : man :: nurse : woman** - pristranost

- Metode *odstranjevanja pristranosti* : odstranjevanje neželenih povezav, a ohranjanje povezav med spolom in besedami, ki jim je spol inherenten

- “Razpristranjevanje” pomaga k temu, da ne prenašamo stereotipov skozi algoritme

Vložitve, IAT in pristranost

Caliskan et al. 2017. : Semantics derived automatically from language corpora contain human-like biases

Učenje na korpusih - odsevanje in prenašanje pristranosti (GloVe vložitve)

- WEAT : na besednih vložitvah ponovijo spekter znanih pristranosti, ki so bile že pokazane s pomočjo *testa implicitnih asociacij (IAT, Greenwald et al. 2004)* na ljudeh
- Prijeten-Neprijeten
  - družbeno neproblematične asociacije *rože vs. insekti, glasbeni inštrumenti vs. orožje*
  - stereotipne škodljive asociacije *evro-ameriška vs. afro-ameriška imena*
- Moški-Ženska
  - družbeno neproblematične asociacije: povezanost ženskih in moških imen z Ž in M
  - stereotipne asociacije: ženske besede (woman, girl) bolj povezane z *umetnostjo* kot z *mate matiko,* bolj povezane tudi z *umetnostjo* kot *znanostjo*
  - GloVe tudi korelirajo z odstotkom žensk na zaposlitev
- besedne vložitve pokažejo tako stereotipe kot tudi realno družbeno stanje (e.g. poklici)
- pristranost se lahko prenaša z algoritmi umetne inteligence

Cf. CV in trg dela, Bertrand & Mullainathan, 2004

Previdnost pri preučevanju pristranosti

Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor (Nissim et al. 2019)

- Rezultati so pogosto pretirani in poudarjajo pristranosti
- A : B :: C : D - že v zasnovi eksperimentov pogosto ni dopuščeno ponavljanje
  - man : actor :: woman : X, X=actress
  - man : doctor :: woman : X, X=?
- Vpliv parametrov

	man : doctor woman : X	he : doctor she : X	man : computer_programmer woman : X
3COSADD unconstrained	gynecologist doctor	nurse doctor	homemaker computer_programmer
3COSMUL unconstrained	gynecologist doctor	nurse doctor	homemaker computer_programmer
BOLUKBASI unconstrained	midwife gynecologist	nurse nurse	schoolteacher

- Pomembnost transparentnosti eksperimentov

Slovenščina - besedne vložitve

Supej et al., 2019

- Slovenščina – izraženost spola v imenu poklica v večini primerov

Eksperimenti:

- Analogije s poklici preko slovenskih besednih vložitev
- *word2vec* algoritem
- Korpusu približno 600.000 besed (akademski teksti, novice, knjige itd.) (Plahuta 2019)

- 48 parov poklicev iz treh skupin (ULRS 28/1997):
  1) Zakonodajalci/zakonodajalke, visoki uradniki/visoke uradnice, menedžerji/menedžerke
  2) Strokovnjaki/strokovnjakinje
  3) Uradniki/uradnice

1,2 večje razlike v plačah
3 najmanjše

- moški:menedžer :: ženska:*X*
- ženska:menedžerka :: moški:*X*
- **Hipoteza: *X* je ženska/moška oblika poklica**

10 besed z najvišjo kosinusno podobnostjo

Slovenščina - besedne vložitve

Supej et al., 2019

Osnova - Ž	Analogije - M		Osnova - M	Analogije - Ž
direktorica	0.78 direktor		direktor	0.75 direktorica
	0.68 generalni_direktor			0.59 šefica
	0.65 pomočnik_direktoria			0.58 generalna_direktorica
	0.64 namestnik_direktoria			0.58 šefinja
	0.64 šef			0.58 predstavnica
	0.63 vodja			0.57 menedžerka
	0.59 izvršni_direktor			0.56 izvršna_direktorica
	0.58 ravnatelj			0.56 tajnica
	0.58 soustanoviteli			0.56 uprava
	0.58 član_upravnega_odbora			0.56 upravnica

Slovenščina - besedne vložitve

Supej et al., 2019

Rezultati

- Večina - pravilni rezultati analogije kot prvi zadetek (tj. X je pričakovan poklic in ima najvišjo kosinusno podobnost)
  - v 71% primerov pri iskanju moških in 87% pri iskanju ženskih analogij
- Skoraj vedno (96% oz. 98% primerov) kot eden izmed 10 zadetkov
  - moški:duhovnik:: ženska:**nuna**
  - ženska:tajnica:: moški:**šef**

- Nekateri rezultati med prvimi 10 zadetki nepovezani s poklicem v analogiji (npr. *hišnik*, *mehanik*, *taksist* za moške in *služkinja*, *gospodinja*, *medicinska sestra*, *kuharica* za ženske)

- Med 20-imi najbolj pogostimi besedami, ki so se pojavile, je bilo na „moški strani“ dosti poklicev z visokim statusom (npr. *odvetnik*, *šef*, *direktor*, *profesor*), skupno 50 (in le 26 na „ženski strani“, *tajnica* najpogostejši poklic)

- Nekaj stereotipnih ali žaljivih analogij med 10 zadetki (npr. plesalec – *striptizeta*; pismonoša – *ciganka*)

2) Analiza reprezentacije spolov

Diahrone analize

Reprezentacija spolov – diahrone analize

Garg et al. 2018: Word embeddings quantify 100 years of gender and ethnic stereotypes

- besedne vložitve natrenirane na korpusih, ki vključujejo 100 let različne literature
- besedne vložitve "ujamejo" spreminjajoče se stereotipe do spolov in različnih manjšin
- spremembe v vložitvah pokažejo na konkretne dogodke (npr. feminizem v ZDA, razlika 60. in 70. leta)

Primer pridevnikov

- povezava z inteligenco, premišljenostjo in logiko: pozitiven trend
- fizični izgled: ni stat. sign. sprememb

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

3) Vpliv na orodja UI

Pristranost in orodja rač. jez.

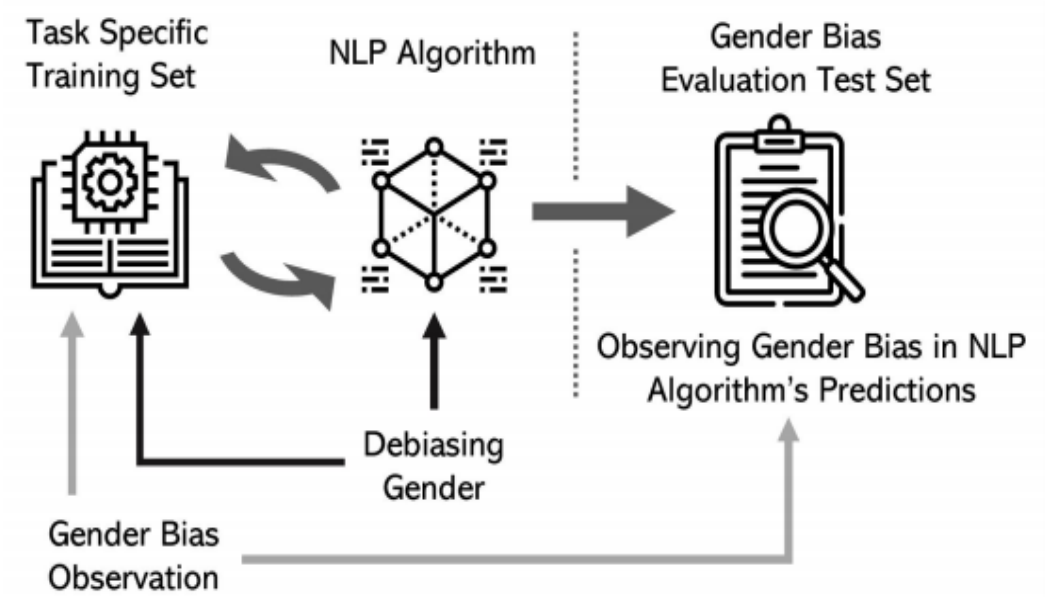
Sun et al. 2019. : Mitigating Gender Bias in Natural Language Processing: Literature Review

- Pristranost razporeditve: algoritmi delujejo bolje na podatkih spola, ki je v podatkih prevladujoč (pomembnost učnih množic)
- Pristranost reprezentacije: pristrane povezave med spolom in koncepti

Task	Example of Representation Bias in the Context of Gender
Machine Translation	Translating "He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor." (Douglas, 2017)
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018).
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017).
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018).
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016).

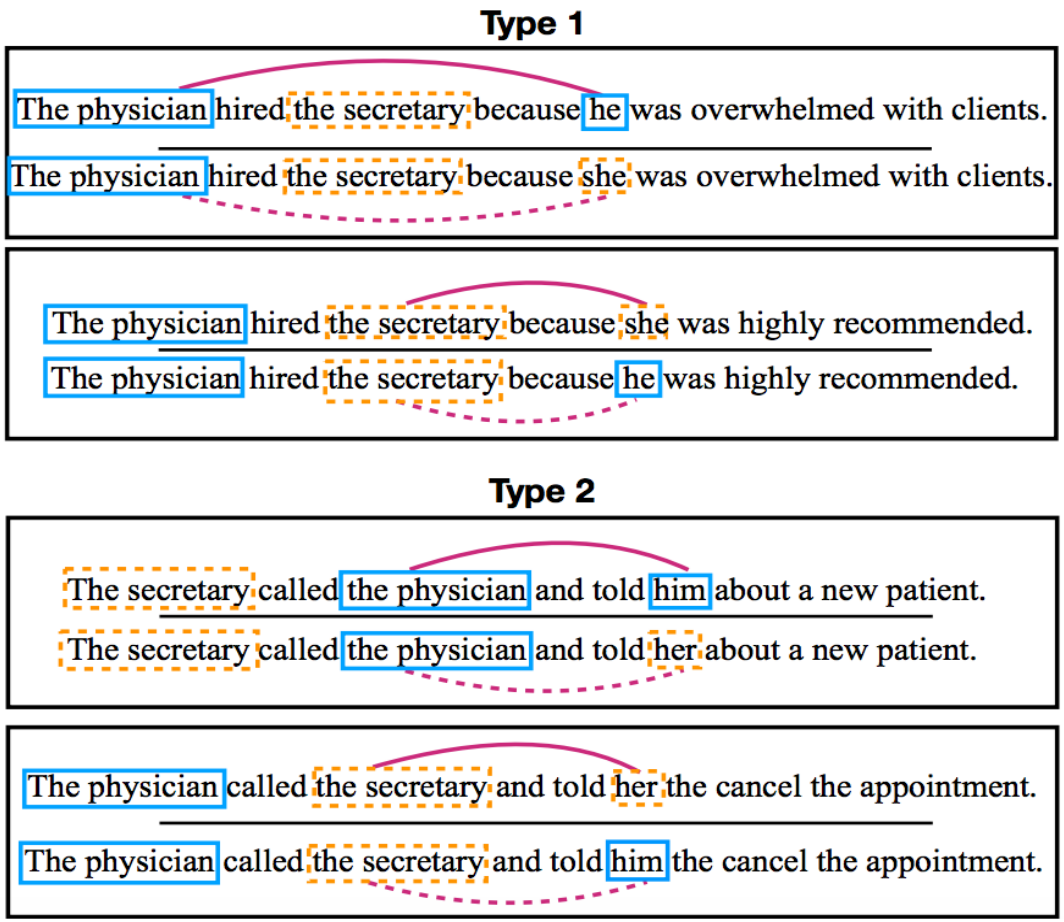
Pristranost in orodja rač. jez.

Sun et al. 2019. : Mitigating Gender Bias in Natural Language Processing: Literature Review



Pristranost in orodja rač. jez.: testne množice

Zhao et al. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods



Pristranost in orodja rač. jez.

Sun et al. 2019. : Mitigating Gender Bias in Natural Language Processing: Literature Review

Veliko raziskav za gradnjo ne- oz. manj pristranih modelov

- obstajajo različni načini odstranjevanja pristranosti in študije področja (Zhao et al., Gonen and Goldberg, Prost et al.)
- pogosto nemožno doseči nepristranost;
- tehnike so bile večinoma preverjene le na posameznih nalogah, pri nekaterih lahko poslabšajo točnost
- potrebno še veliko celovitih raziskav

Zaključki

- Področje računalniškega jezikoslovja se je začelo ukvarjati s **pristranostjo**
- Vektorskih besednih vložitvah
- **Analizo** reprezentacije spola v jezikovnih korpusih
- Prenašanje **pristranosti v algoritmih**
- Pomembnost analize učnih množic in modelov
- Orodja za ”razpristranjevanje” modelov
- Transparentno raziskovanje, odprta znanost
- Ne gre le za spol, ampak tudi za druge aspekte
- Financiranje za transparente raziskave in odprto in strpno družbo
- Diseminacija in razumevanje orodij UI v javnosti

EMBEDDIA: Cross-Lingual Embeddings for Less-Represented Languages in European NewsMedia

www.embeddia.eu

@embeddiaproject

Zahvala: A. Supej, M. Purver, M. Robnik Šikonja