What It Takes to Control ~~Societal Bias in~~ Natural Language Processing

Kai-Wei Chang
UCLA

References: http://kwchang.net

Always working?!

http://viralscape.com/travel-expectations-vs-reality/

Performance on Benchmarks

Performance in reality

NLP Models are Brittle

Generating Natural Language Adversarial Examples [ASEHSC(EMNLP 18)]

Original Text Prediction: **Entailment** (Confidence = 86%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **runner** wants to head for the finish line.*
Adversarial Text Prediction: **Contradiction** (Confidence = 43%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **racer** wants to head for the finish line.*

Retrofitting Contextualized Word Embeddings with Paraphrases [SCZC (EMNLP 19)]

Paraphrased contexts	L2	Cosine
How can I make **bigger** my arms? How do I make my arms **bigger**?	6.42	0.27
Some people believe earth is **flat**. Why? Why do people still believe in **flat** earth?	7.59	0.46
It is a very **small** window. I have a **large** suitcase.	5.44	0.26

Training NLP models Require Large Data

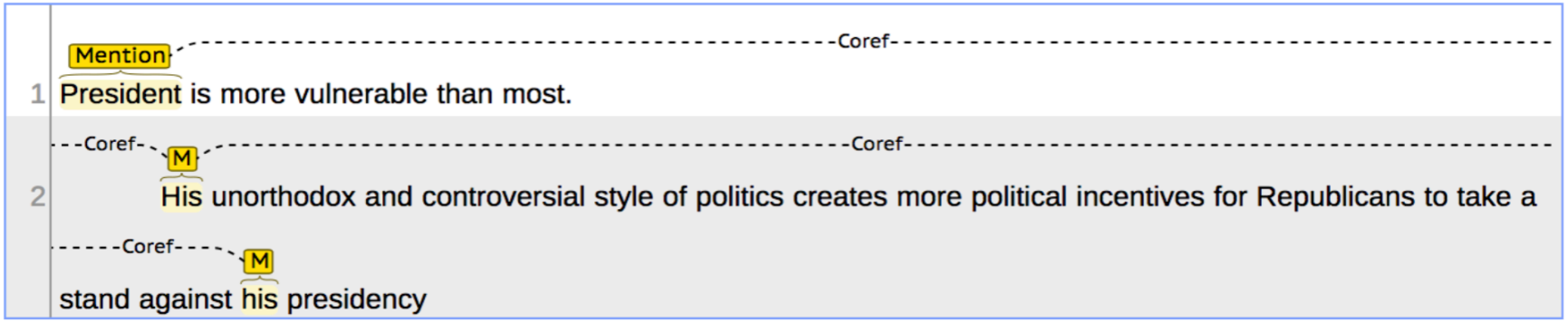
How about low-resource languages?

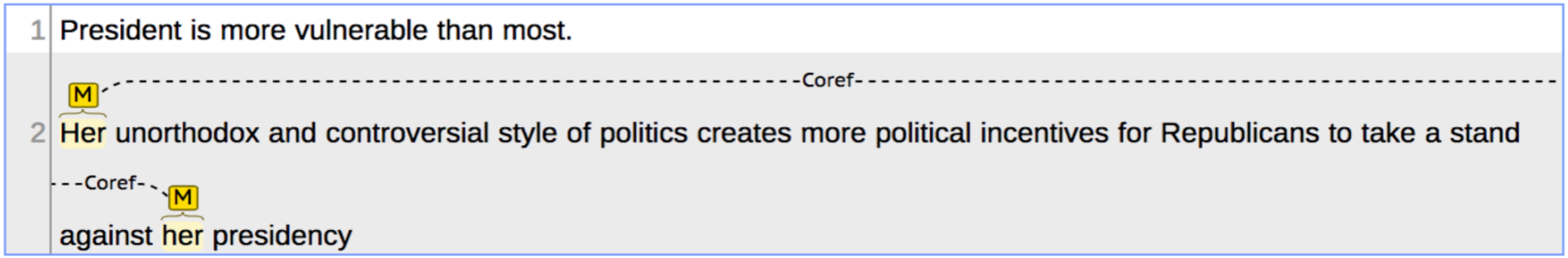
How about domains where annotations are expansive?

NLP Model is biased

[1] Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods [ZWYOC NAACL 2018]

- Coreference resolution is biased[1,2]
  - Model fails for female when given same context



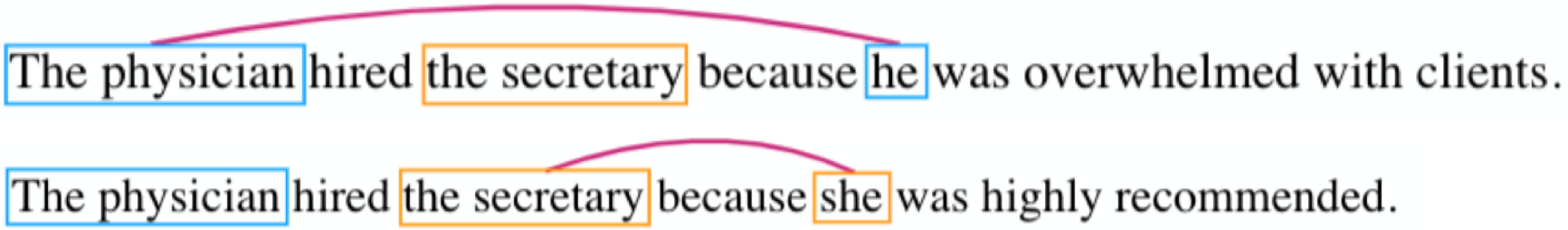


[2] Rudinger et al. Gender Bias in Coreference Resolution. NAACL 2018

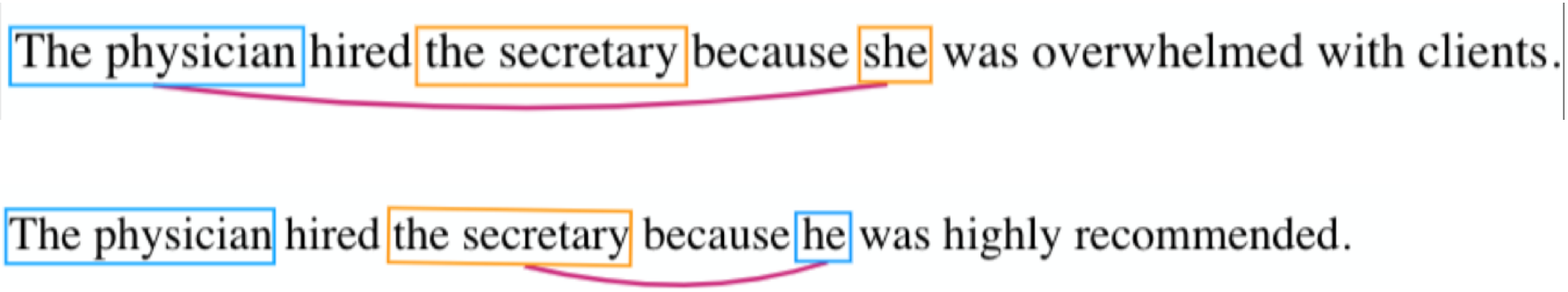
Wino-bias data

[ZWYOC NAACL 2018]

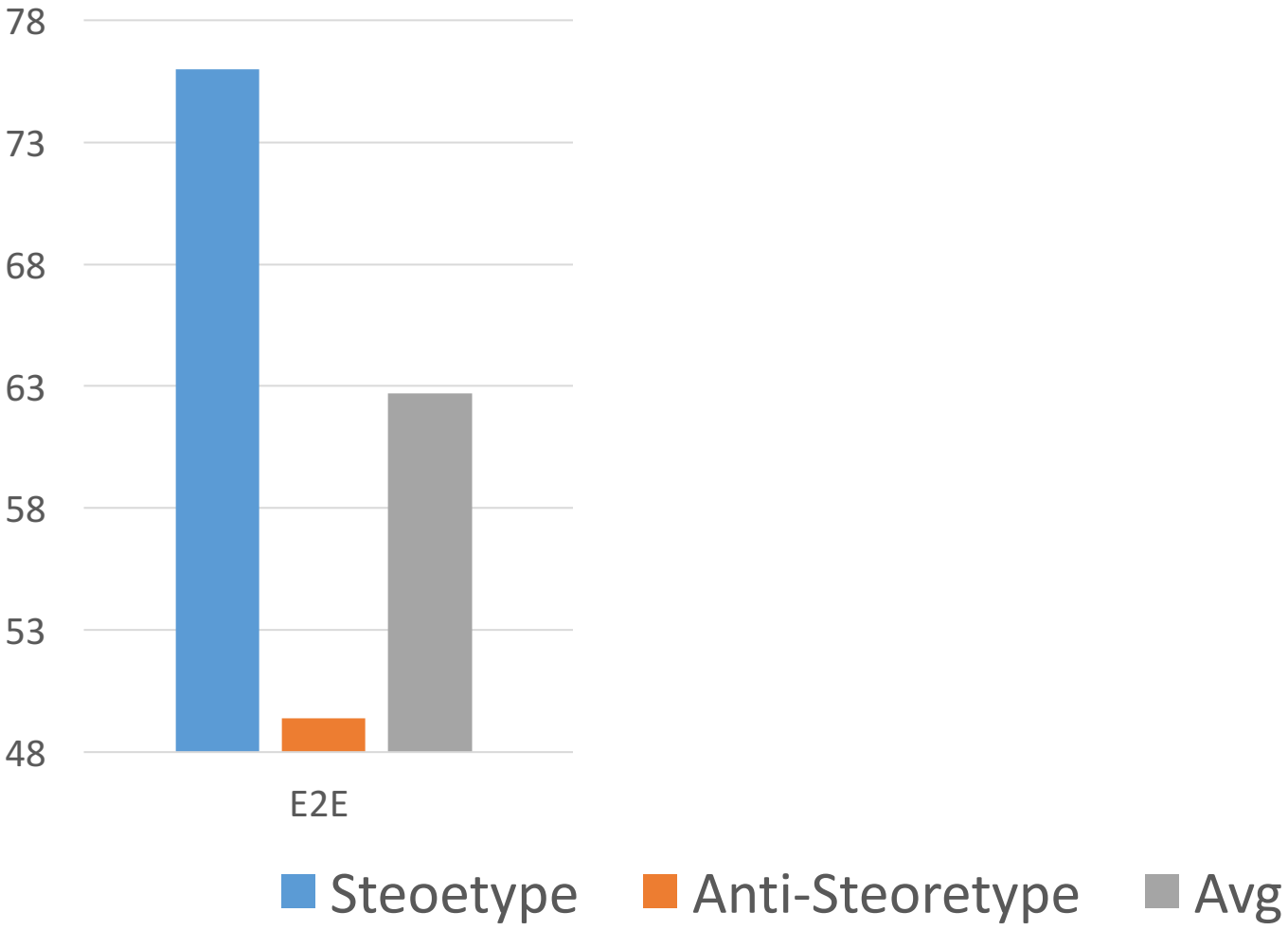
- Stereotypical dataset



- Anti-stereotypical dataset



Gender bias in Coref System



NLP Model is biased

The Woman Worked as a Babysitter: On Biases in Language Generation [SCNP EMNLP 2019]

- Language generation is biased

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Outline

- Gender Bias in NLP
  - Representational harm
  - Performance gap in downstream applications

arXiv.org > cs > arXiv:1906.08976

Computer Science > Computation and Language

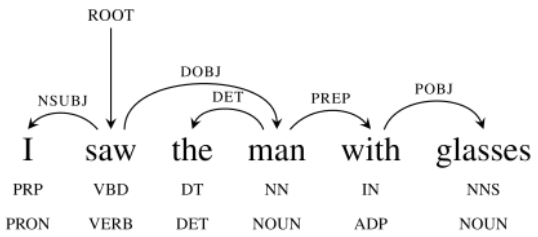
Mitigating Gender Bias in Natural Language Processing: Literature Review

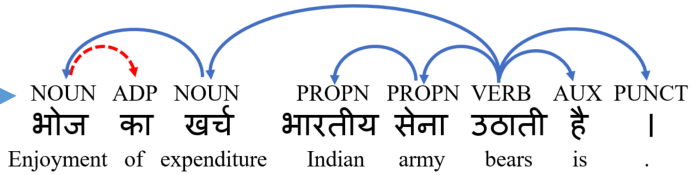
Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, William Yang Wang

(Submitted on 21 Jun 2019)

[ACL 2019]

- Cross-lingual Dependency Parsing



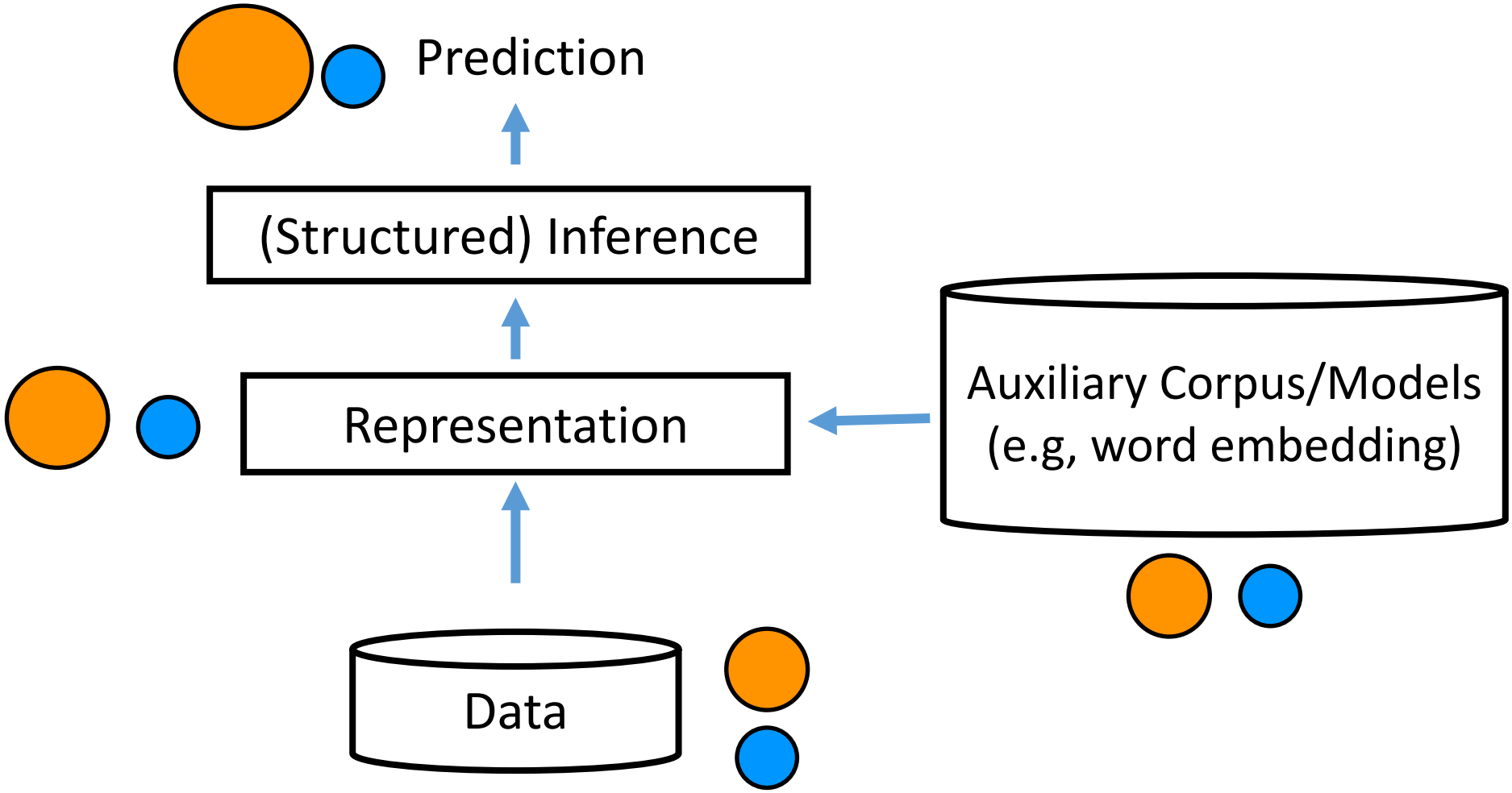


I will show you…

- How to ***unlearn*** unwanted bias in training data
- How to inject knowledge that are not present in training data
- Some ILP formulations

$$\max_{\{y^i\}\in\{Y^i\}} \sum_i f_\theta(y^i, i), \quad \text{s.t.} \quad A\sum_i y^i - b \leq 0$$

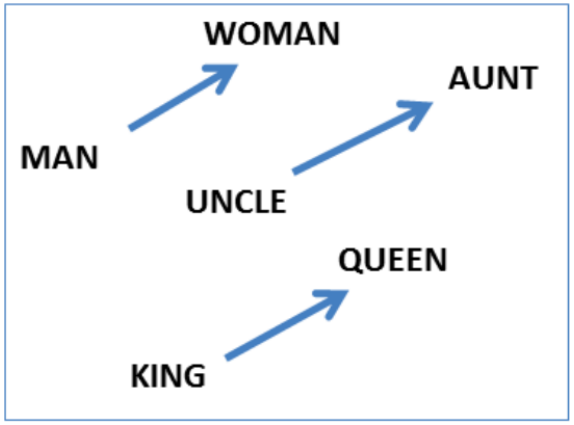
A carton of ML (NLP) pipeline



Representational Harm in NLP: Word Embeddings can be Sexist

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings [BCZSK NeurIPS16]

- $v_{man} - v_{woman} + v_{uncle} \sim v_{aunt}$



he: _____	she: _____
brother	sister
beer	
physician	
professor	

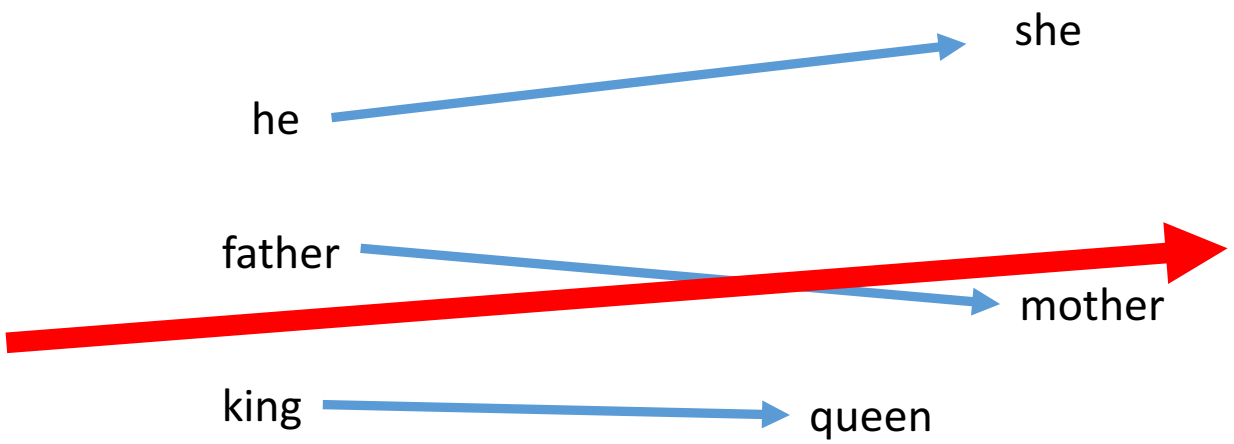
We use Google w2v embedding trained from the news

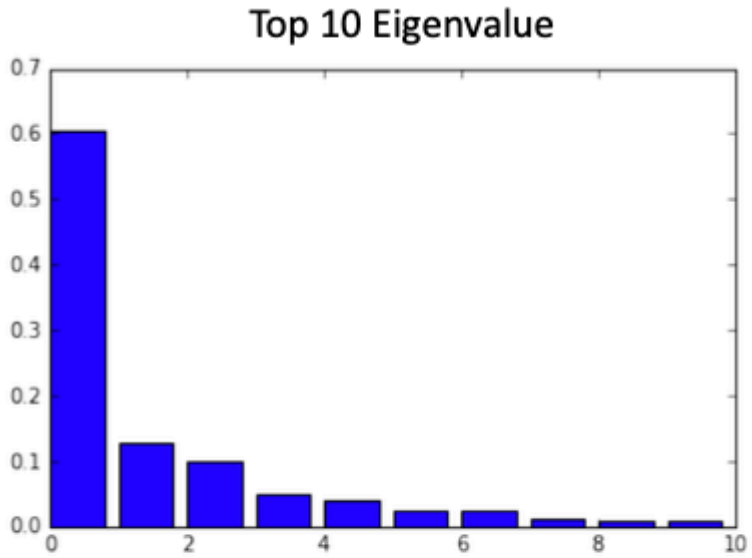
Concurrent work: replicated IAT findings using word embeddings

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]
+ See all authors and affiliations

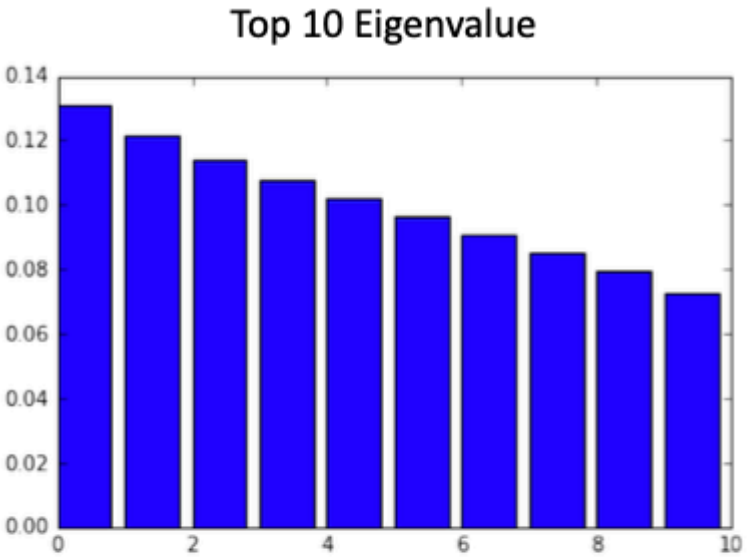
Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230





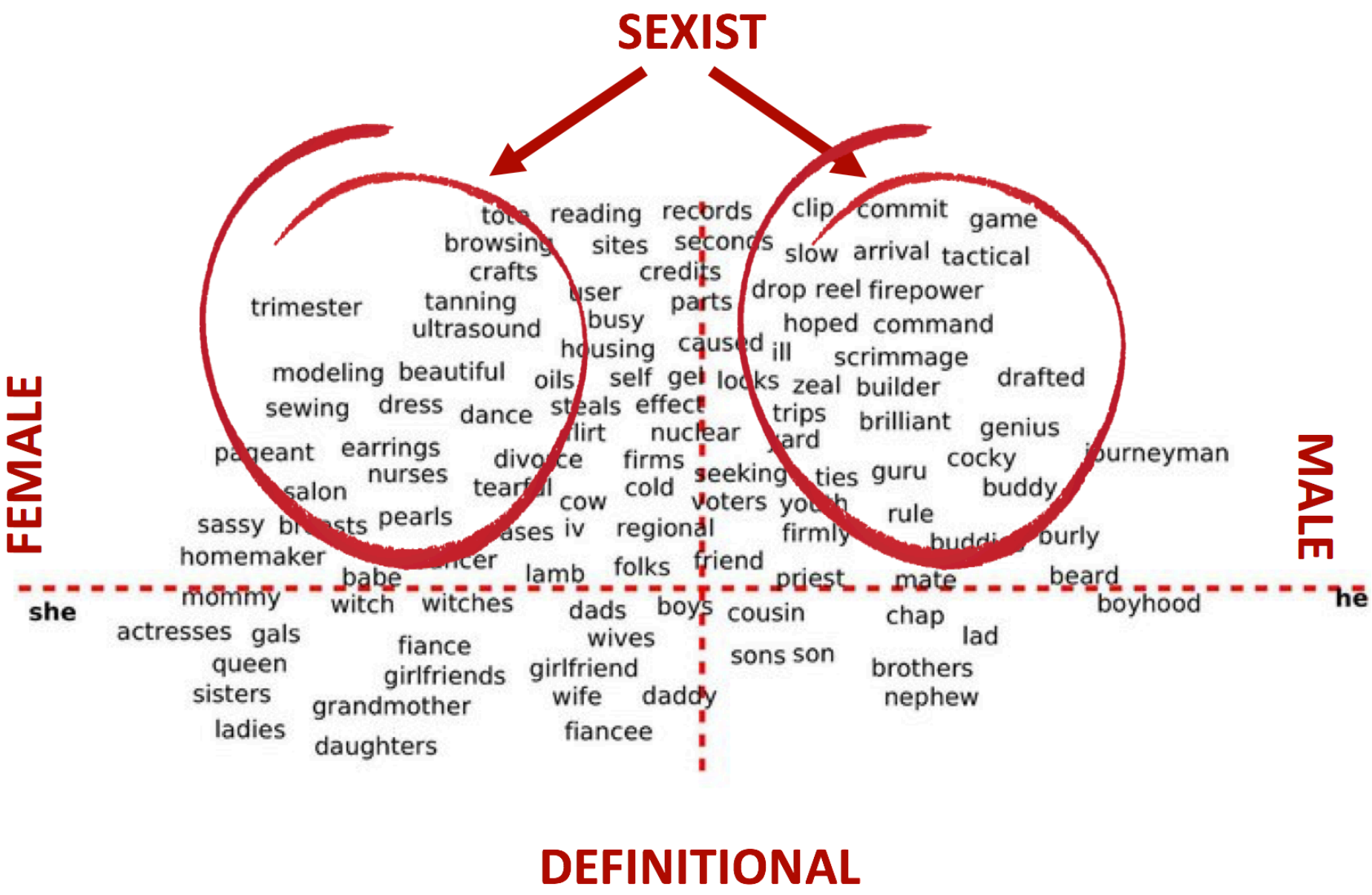
PCA (“he”- “she”, “father”-”mother”,…)

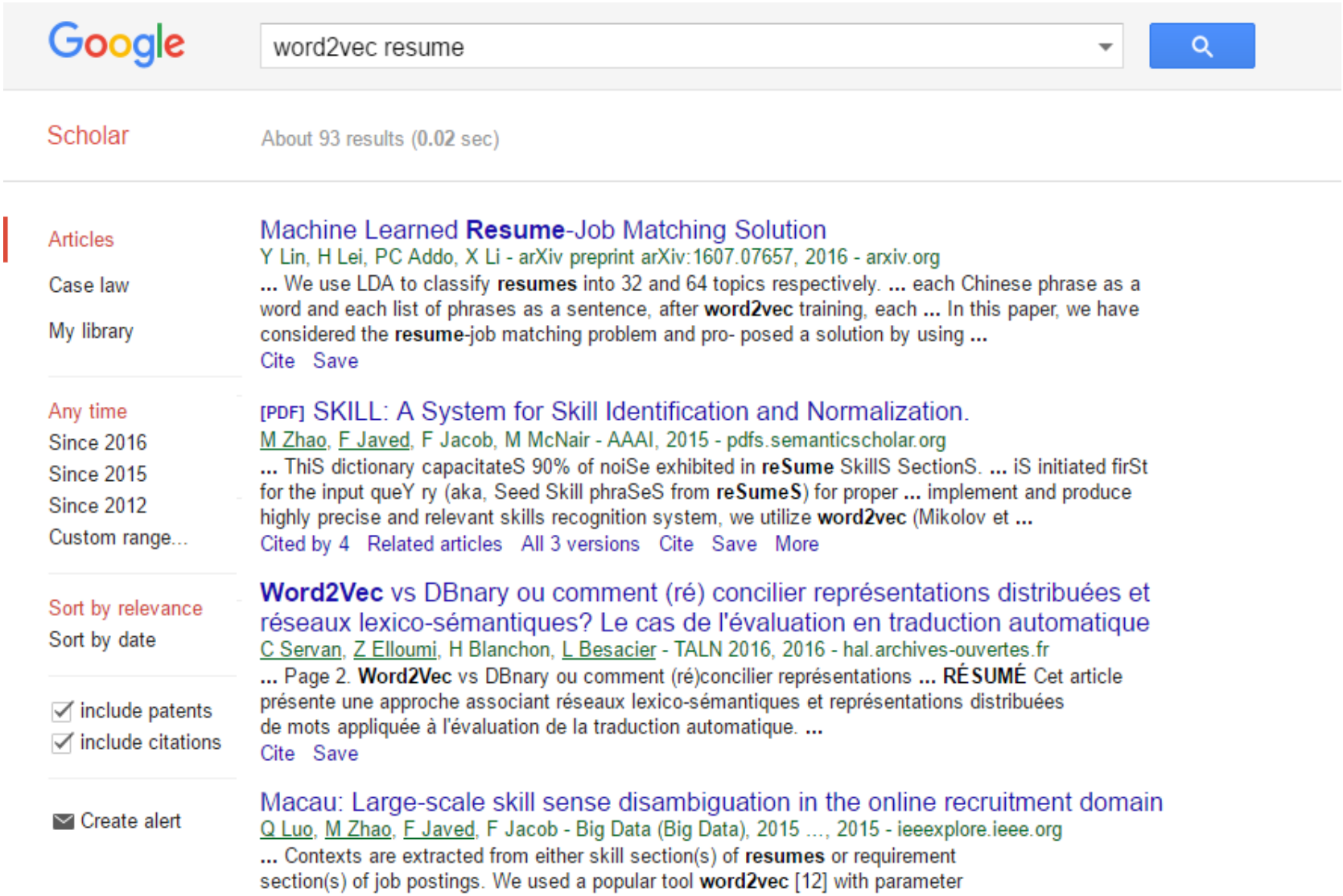
Gender Pair



PCA (“dog”- “cat”, “house”-”building”,…)

Random Pair

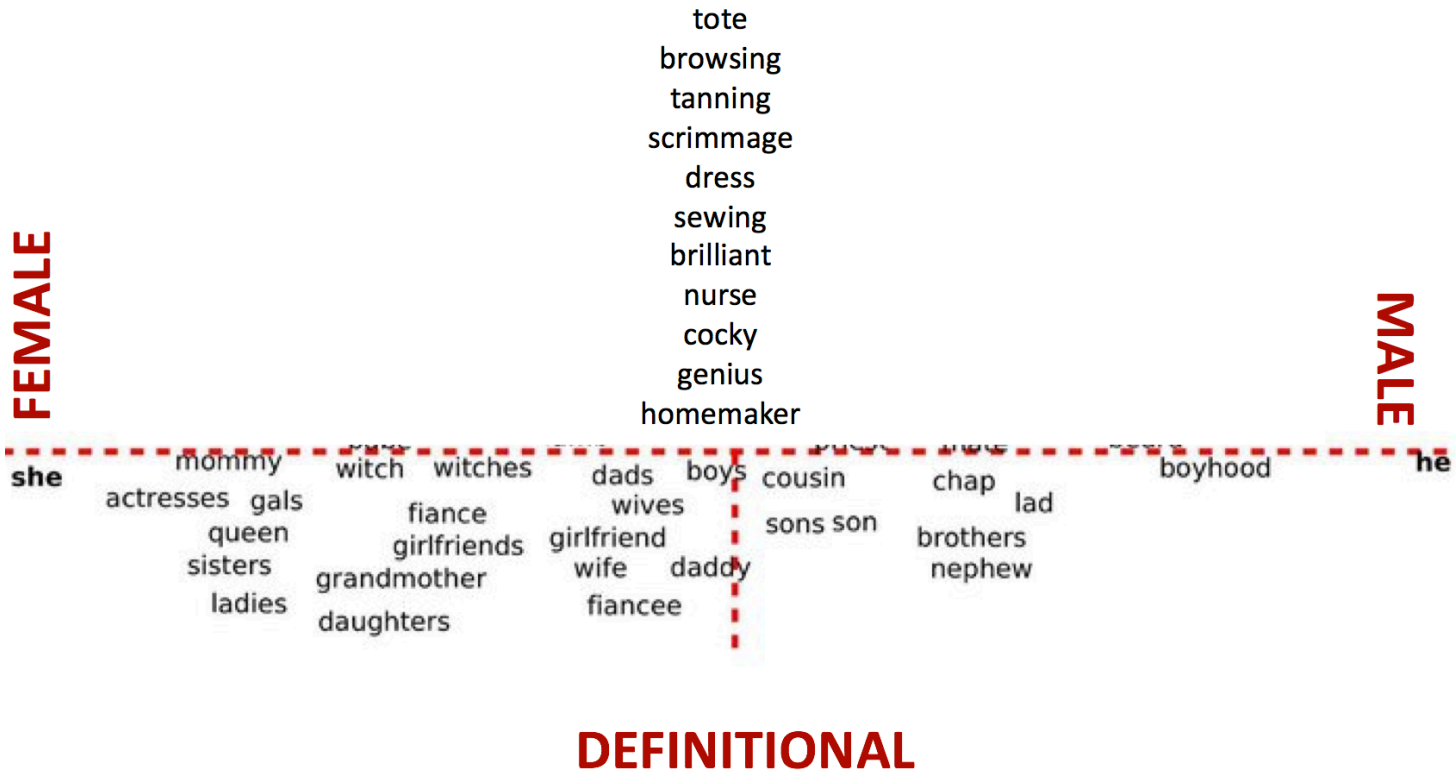

SEXIST
FEMALE
MALE
she
he
DEFINITIONAL
reading
records
clip
commit
game
browsing
sites
seconds
slow
arrival
tactical
crafts
credits
trimester
tanning
user
parts
drop
reel
firepower
ultrasound
busy
hoped
command
housing
caused
ill
scrimmage
modeling
beautiful
oils
self
gel
locks
zeal
builder
drafted
sewing
dress
dance
steals
effect
trips
brilliant
genius
flirt
nuclear
earrings
firms
cocky
journeyman
nurses
seeking
ties
guru
salon
cold
buddy
cow
voters
pearls
rule
regional
firmly
homemaker
burly
lamb
folks
friend
priest
mate
beard
babe
mommy
witch
witches
dads
boys
cousin
chap
boyhood
actresses
gals
wives
lad
fiance
sons
son
queen
girlfriends
girlfriend
brothers
sisters
wife
daddy
nephew
grandmother
ladies
fiancee
daughters



May cause allocative harms in downstream applications

~~SEXIST~~

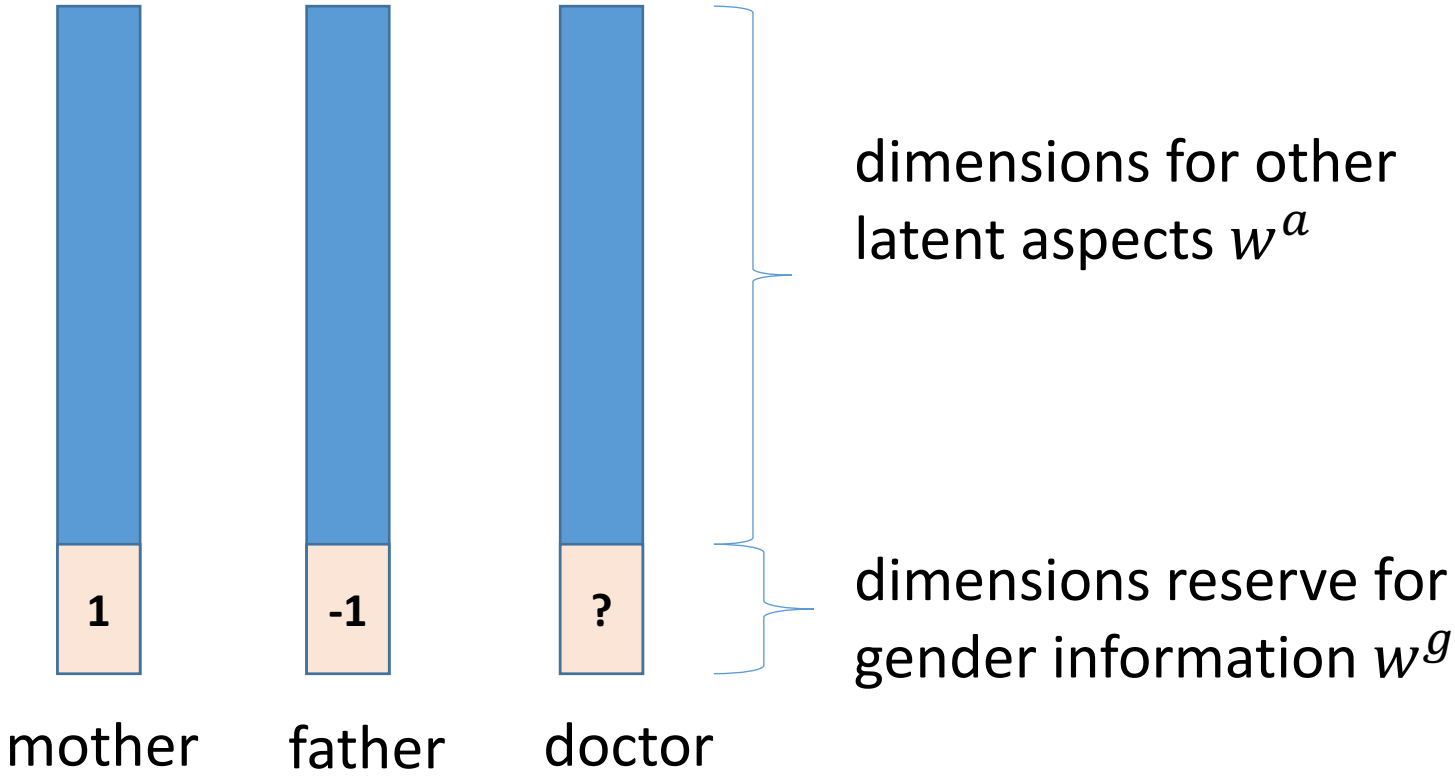
[BCZSK; NeurIPS 16]



This can be done by projecting gender direction out from gender neutral words using linear operations

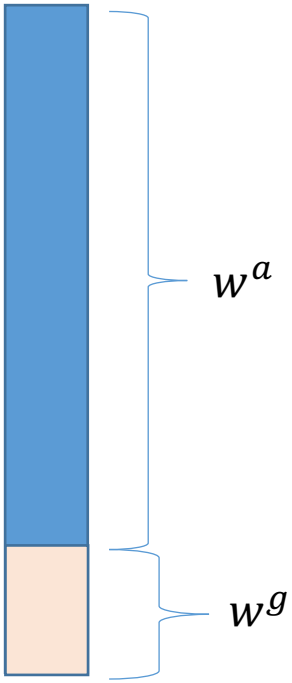
Make Gender Information Transparent in Word Embedding

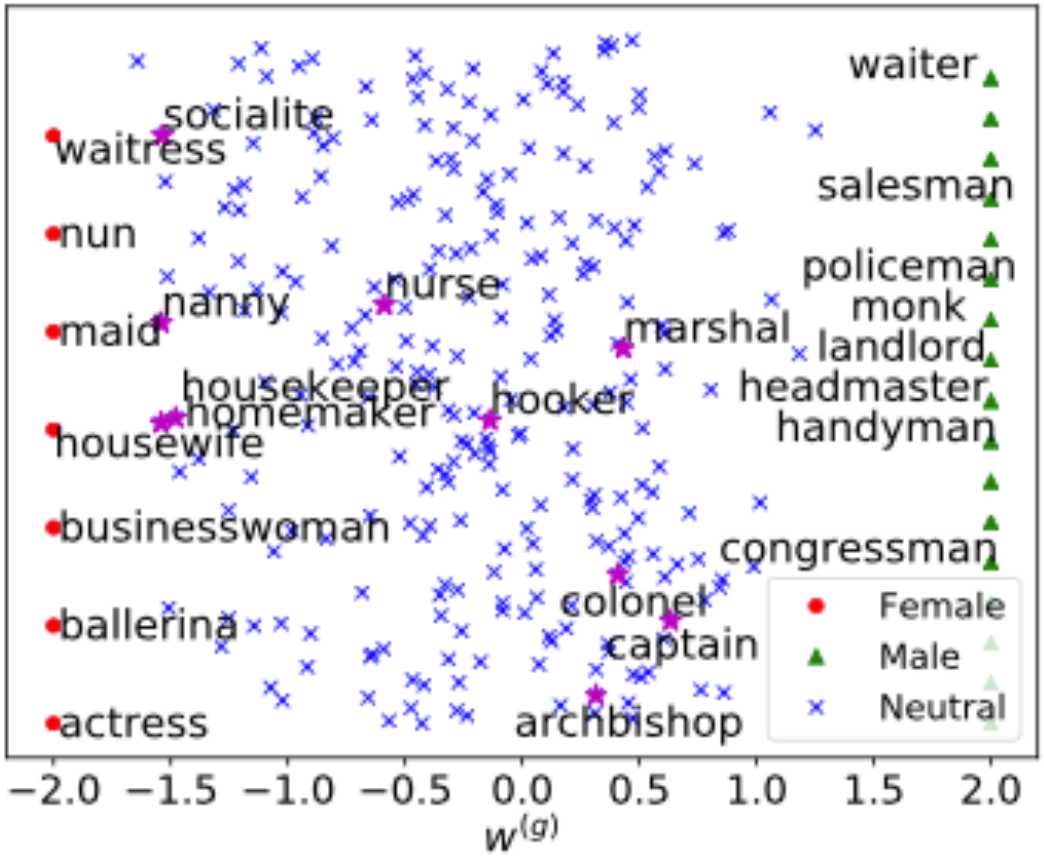
Learning Gender-Neutral Word Embeddings [ZZLWC; EMNLP18]



Make Gender Information Transparent in Word Embedding

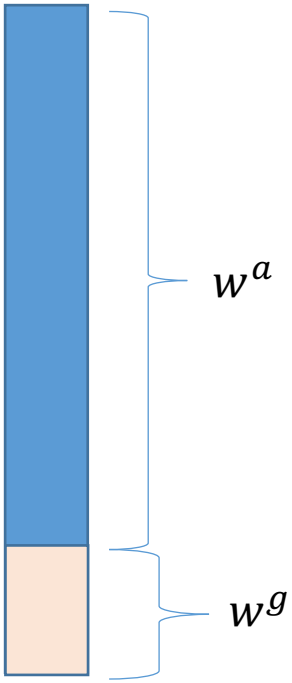
Learning Gender-Neutral Word Embeddings [ZZLWC; EMNLP18]

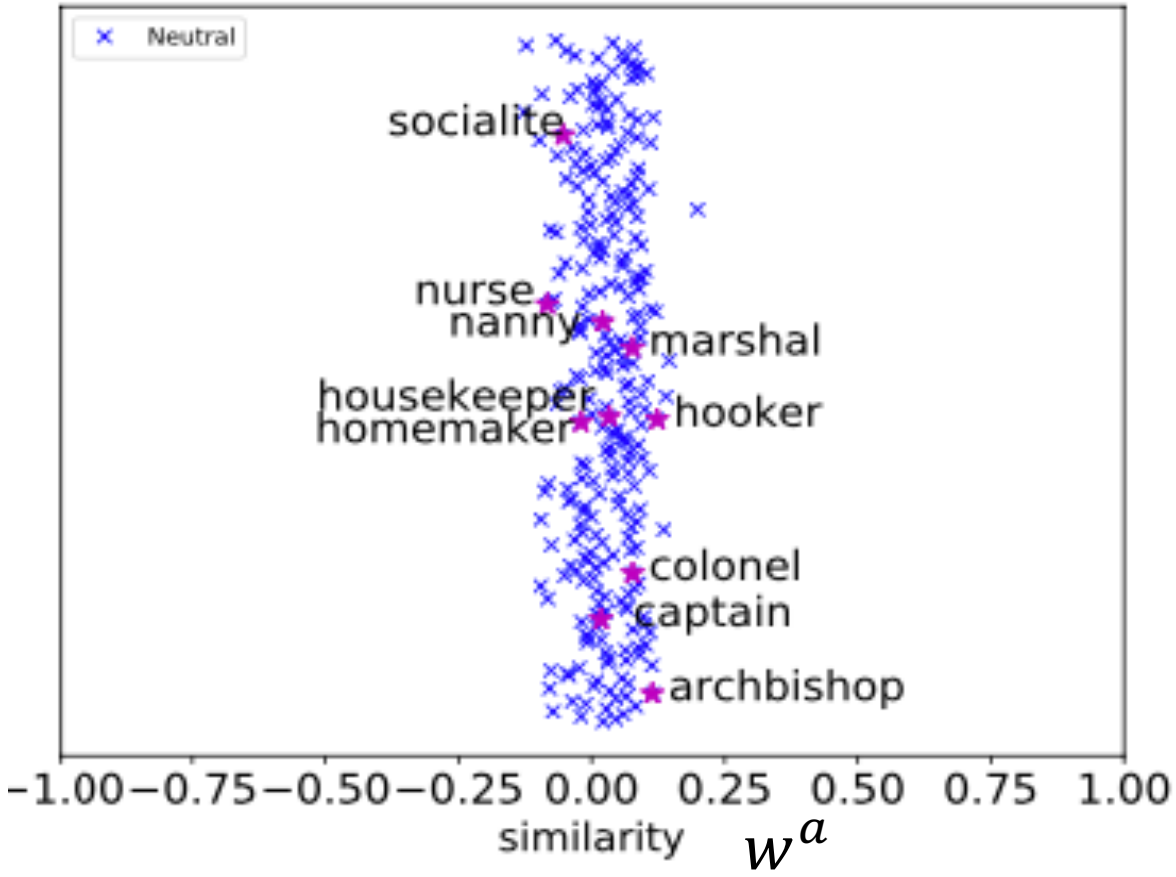




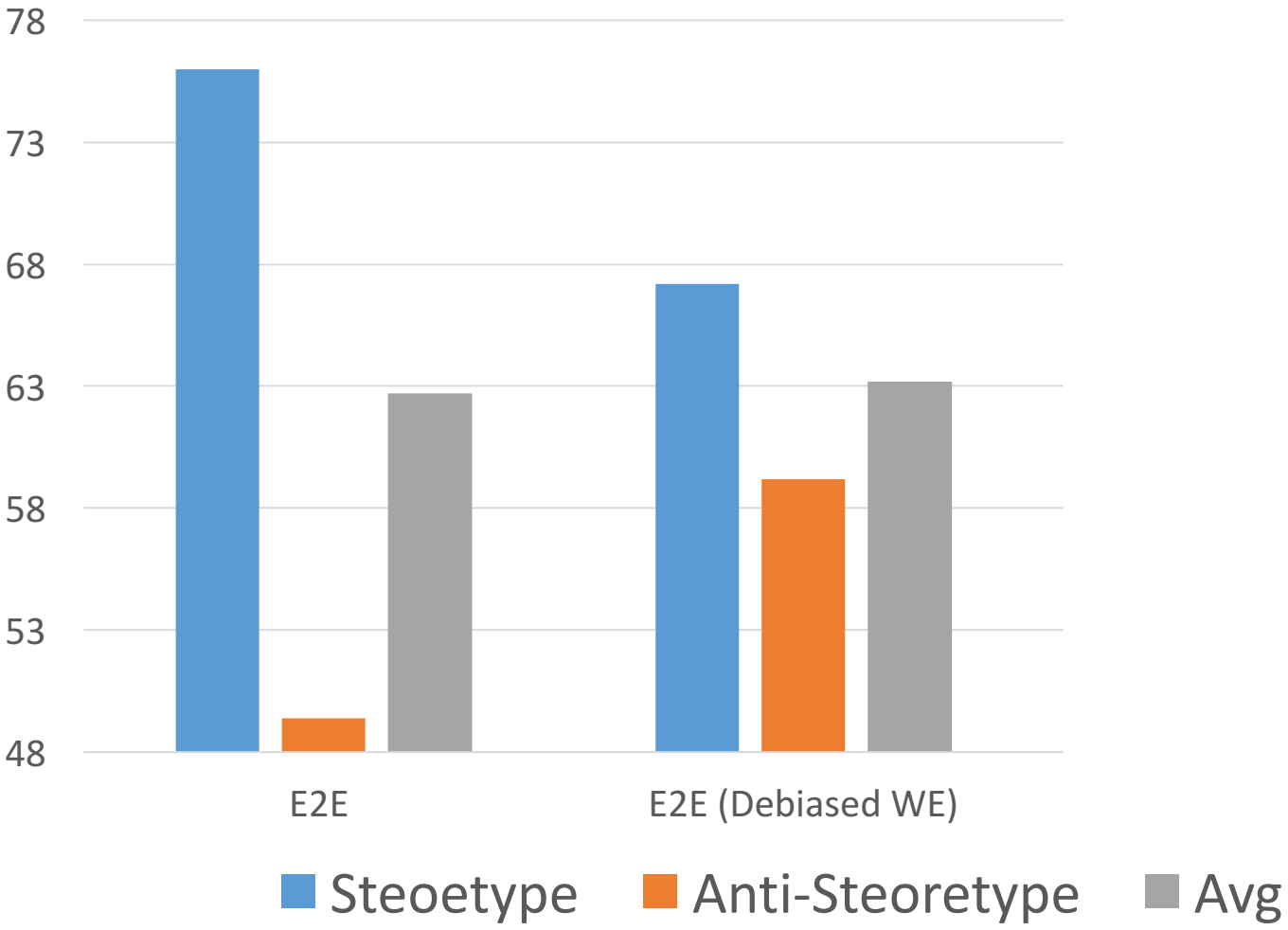
Make Gender Information Transparent in Word Embedding

Learning Gender-Neutral Word Embeddings [ZZLWC; EMNLP18]





Gender bias in Coref System

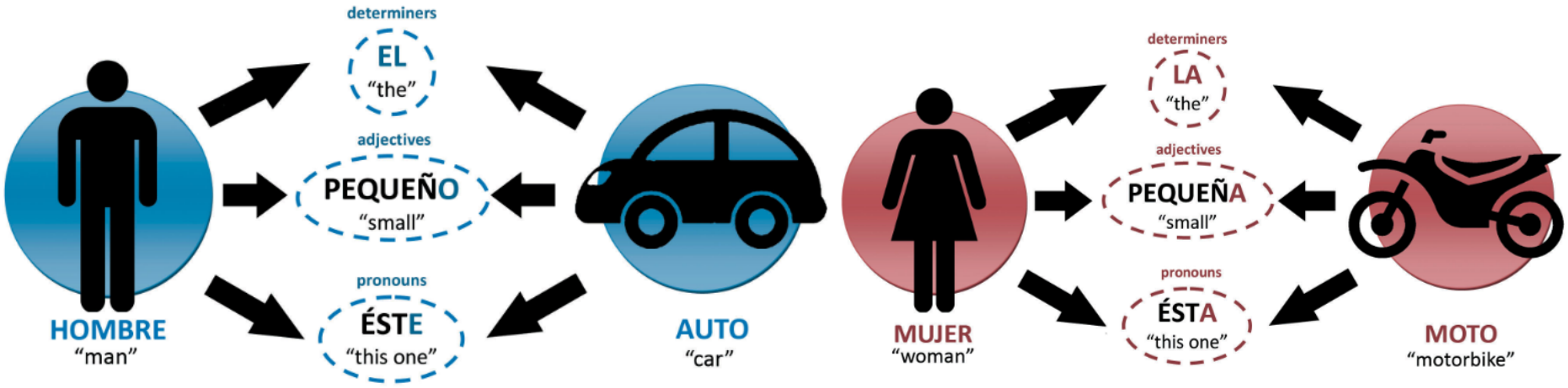


How about…

- language with grammatical gender
- bilingual word embedding
- contextulaized embedding

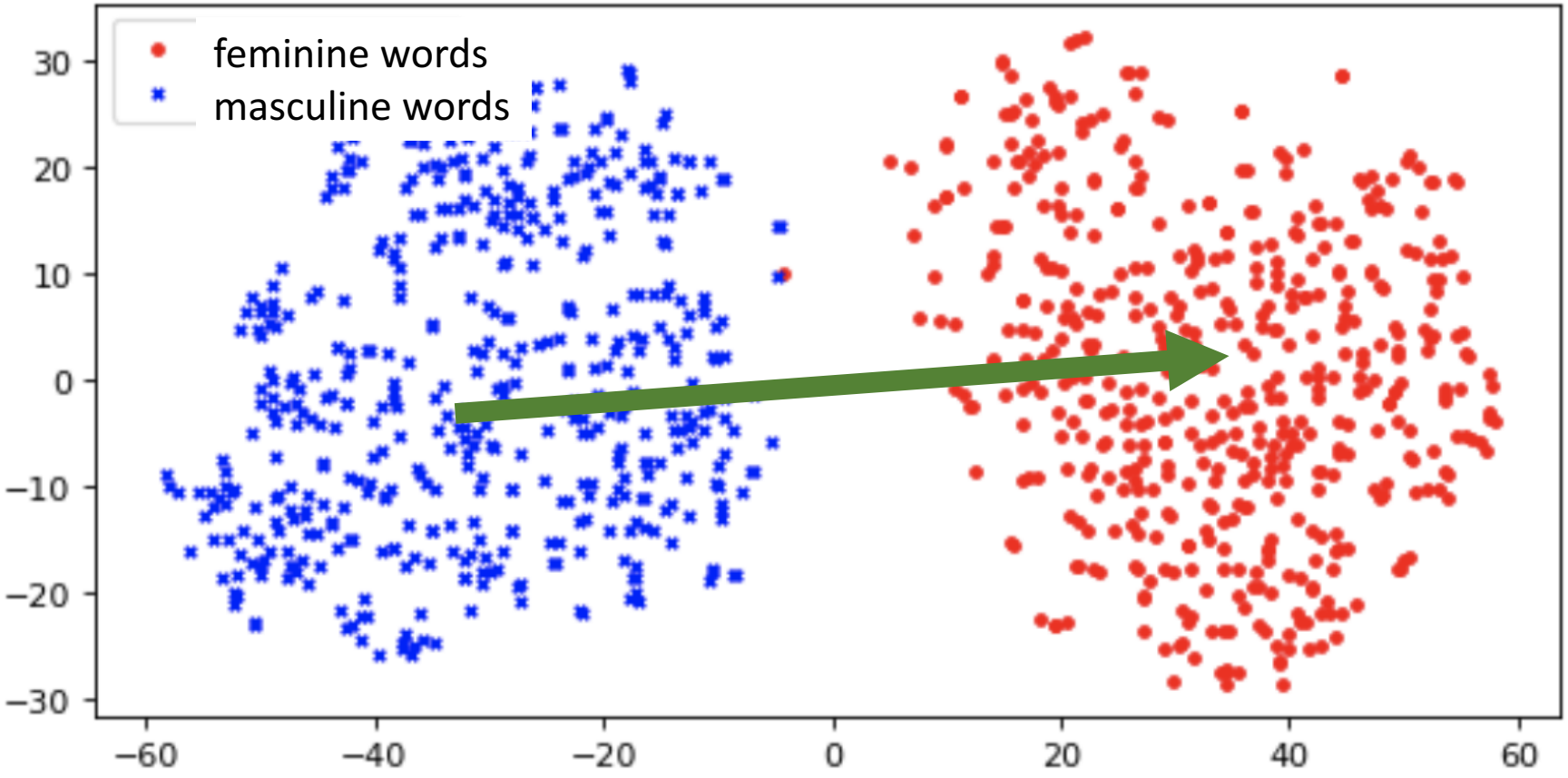
How about other languages?
[ZSZHCCC EMNLP19]

- Language with grammatical gender
  - Morphological agreement

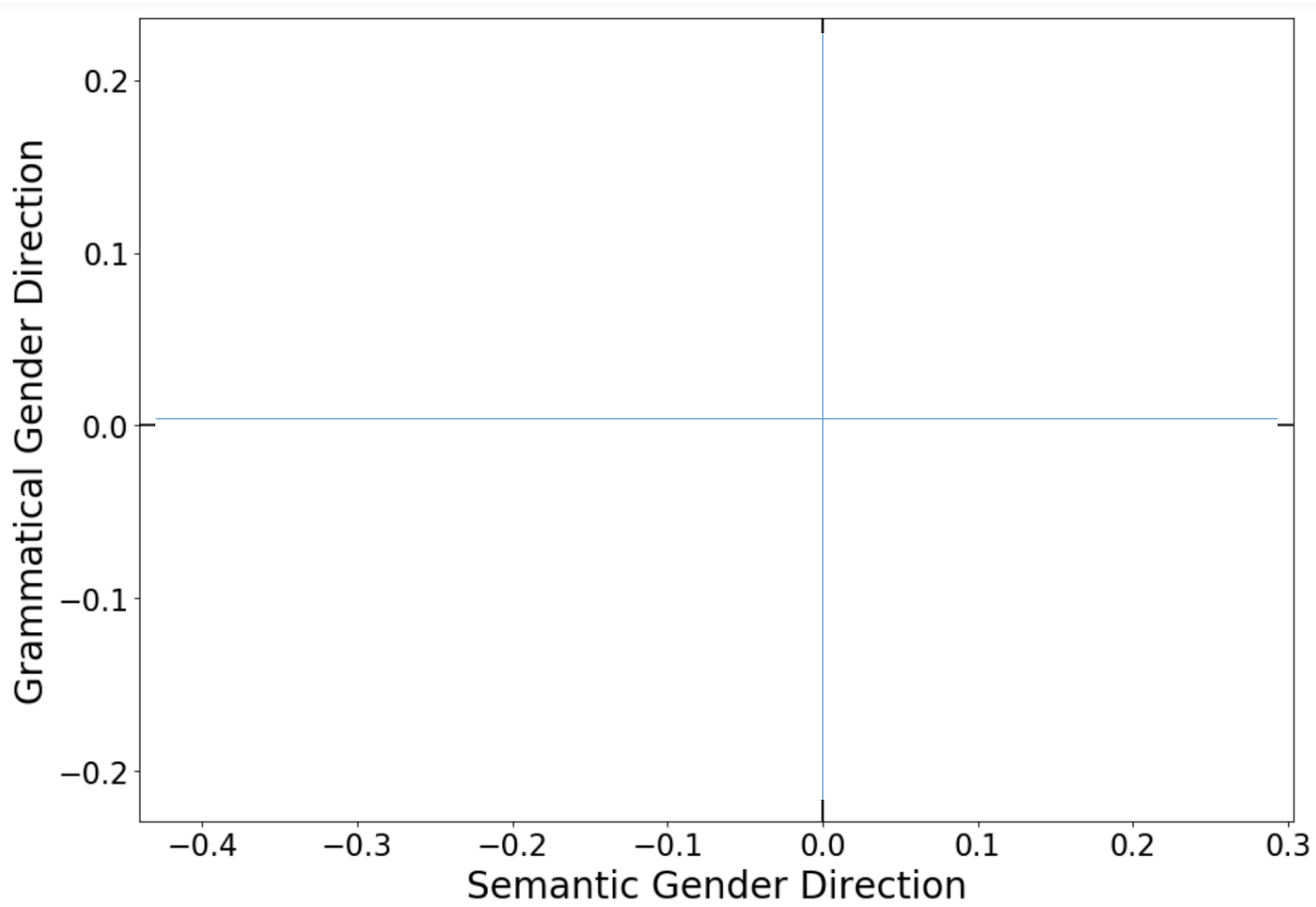


❖ Linear Discriminative Analysis (LDA)

❖ Identify grammatical gender direction



masculine

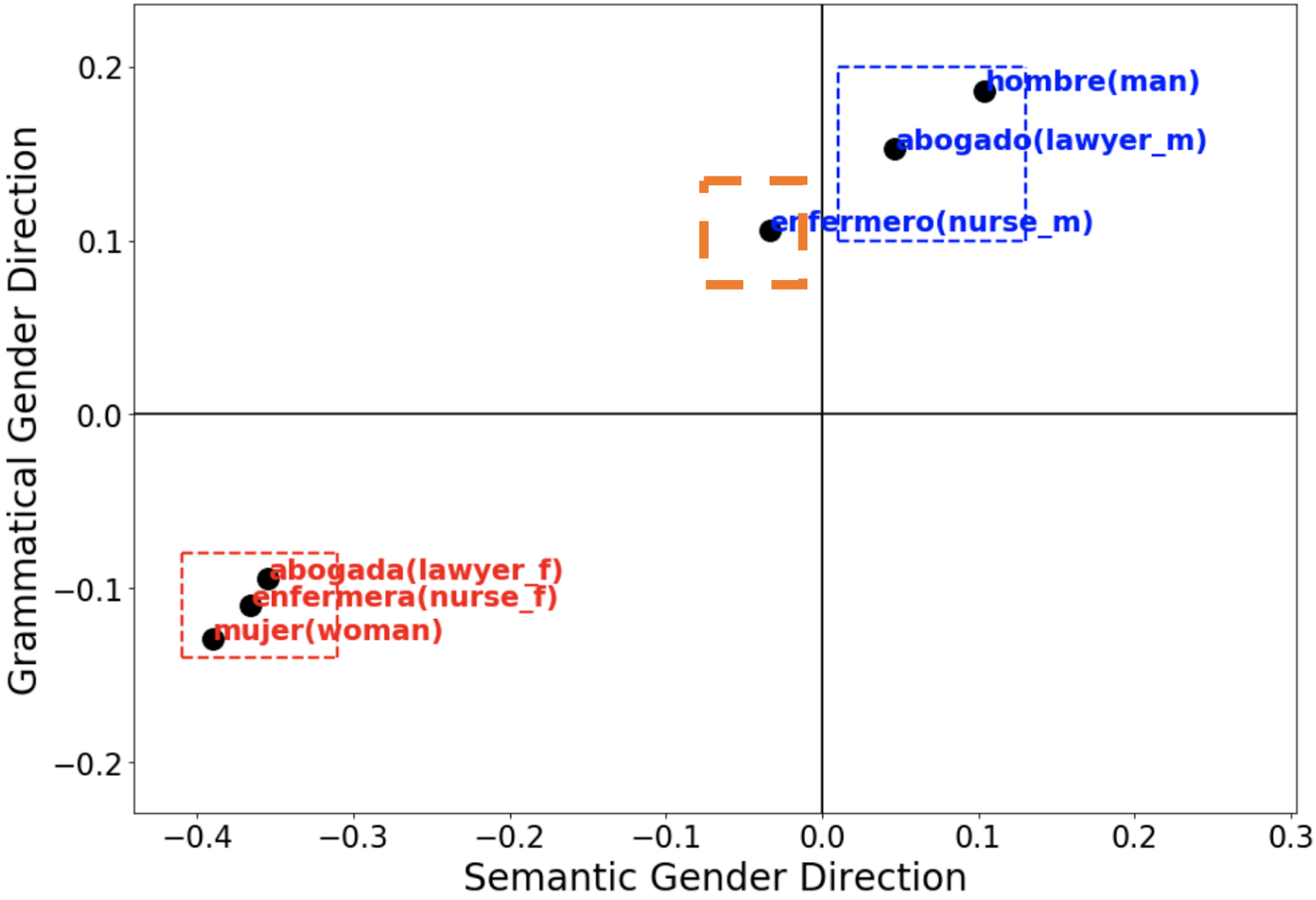


Female

Male

feminine

masculine

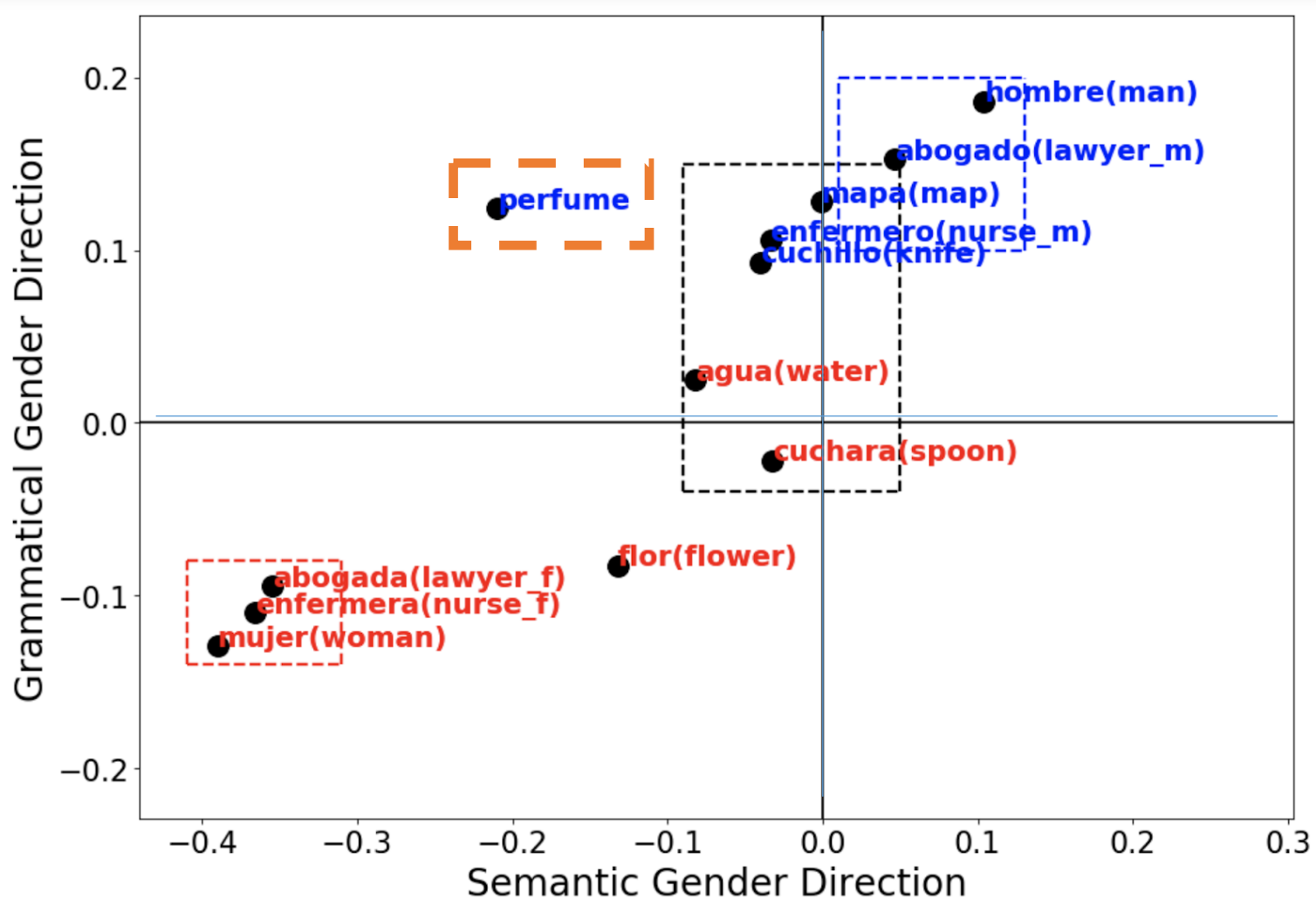


Female

Male

feminine

masculine

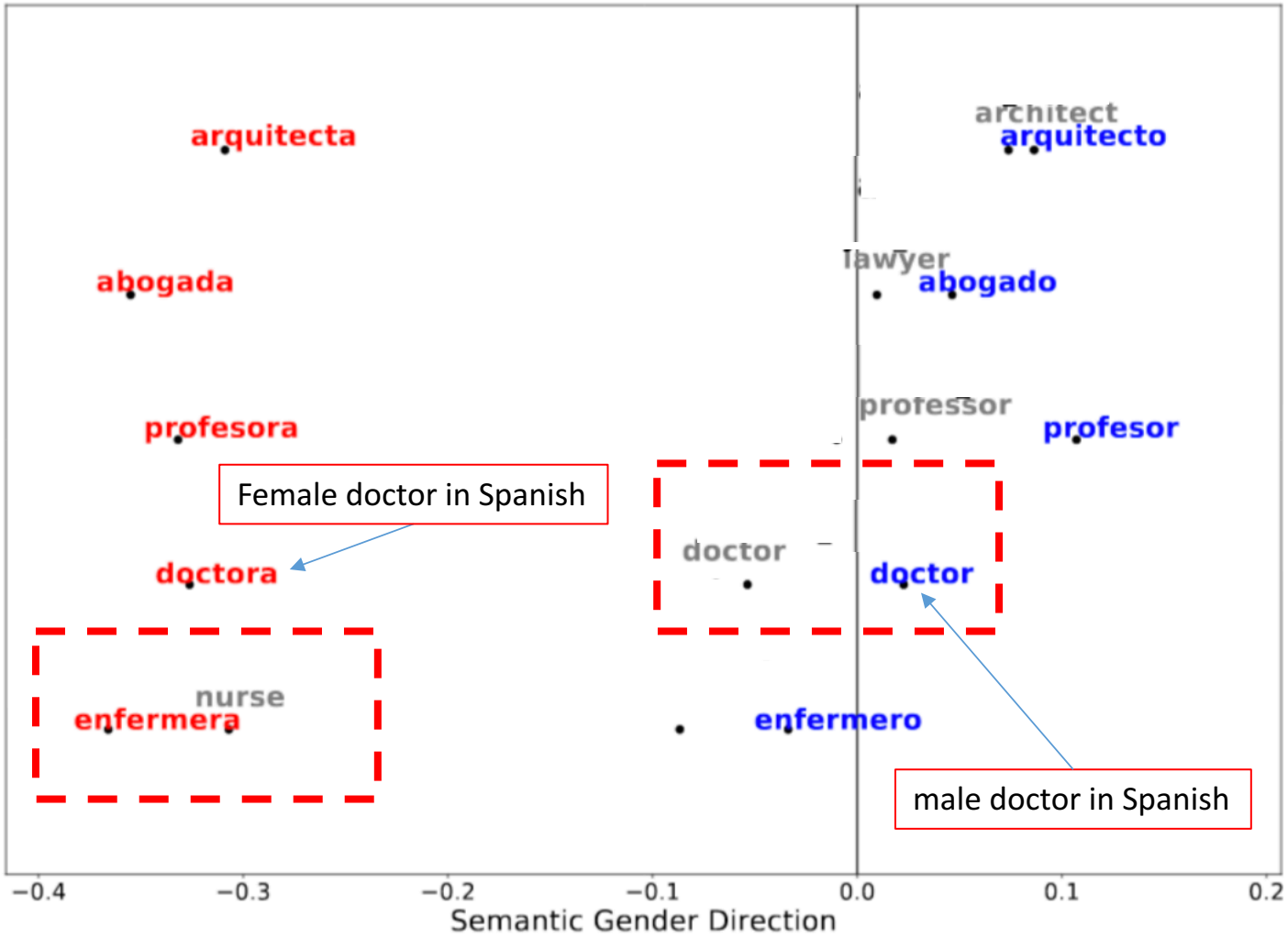


Female

Male

feminine

How about bilingual embedding?
[ZSZHCCC EMNLP19]



How about Contextualized Representation?

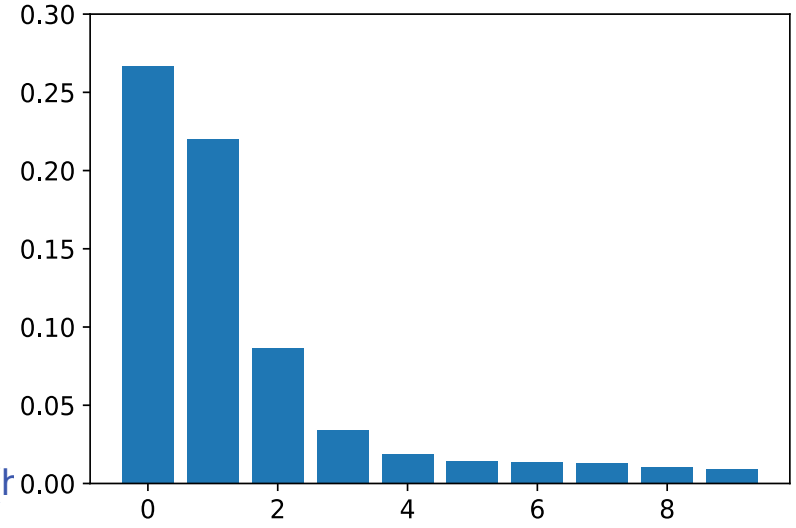
Gender Bias in Contextualized Word Embeddings
[ZWYCOC; NAACL19]

- First two components explain more variance than others

(Feminine) The driver stopped the car at the hospital because **she** was paid to do so

(Masculine) The driver stopped the car at the hospital because **he** was paid to do so

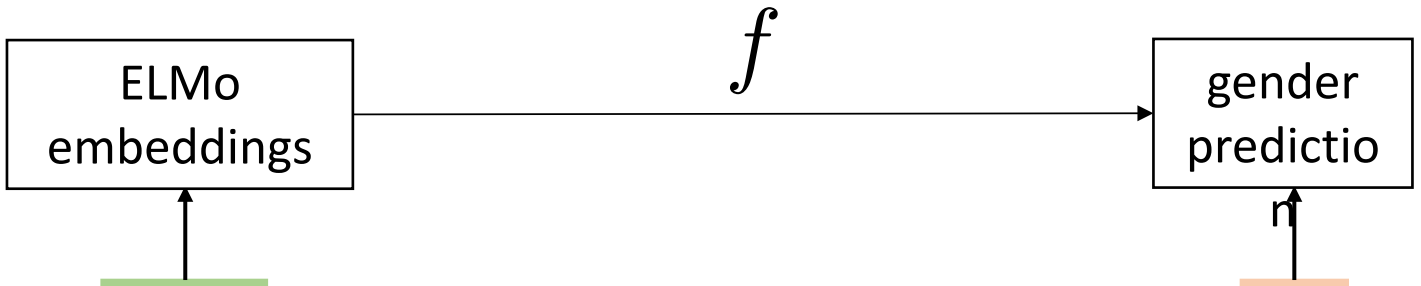
gender direction: ELMo(driver) – ELMo(driver)



Unequal Treatment of Gender

- Classifier

f : ELMo(occupation) $\longrightarrow$ context gender



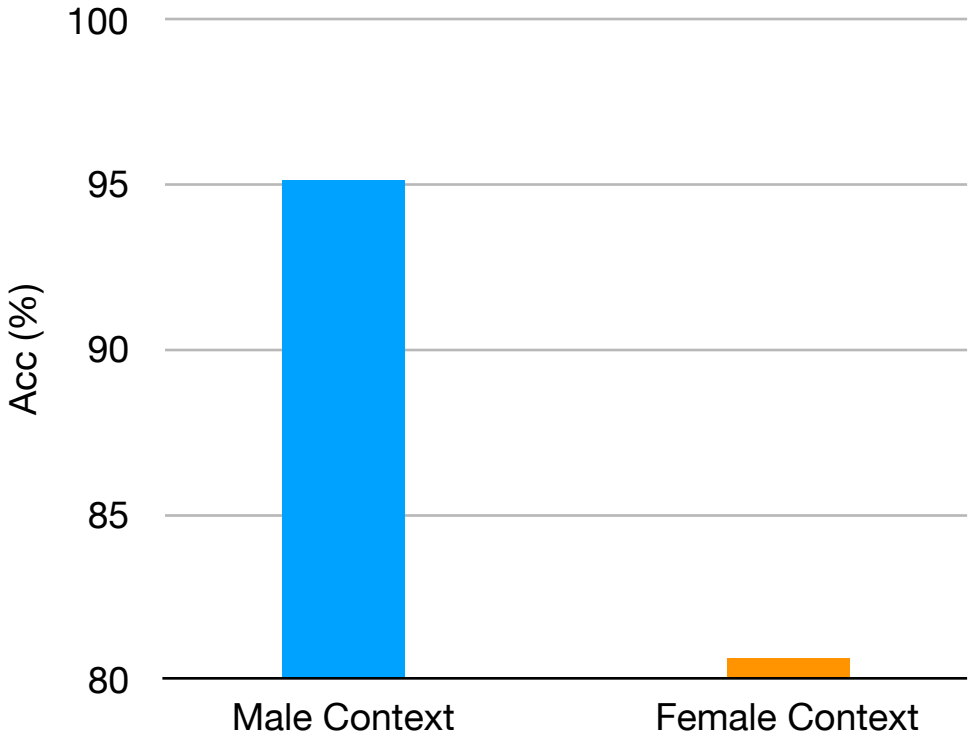
Unequal Treatment of Gender

The writer taught himself to play violin .

- Classifier

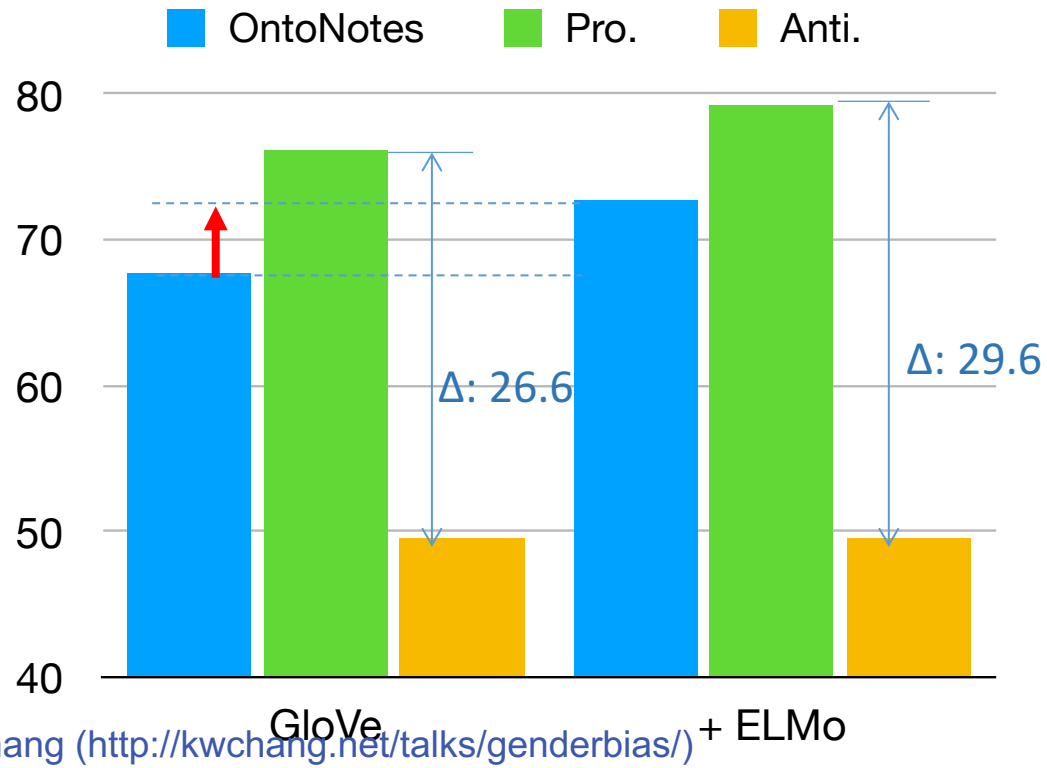
f : ELMo(occupation) $\longrightarrow$ context gender

- ELMo propagates gender information to other words
- Male information is 14% more accurately propagated than female



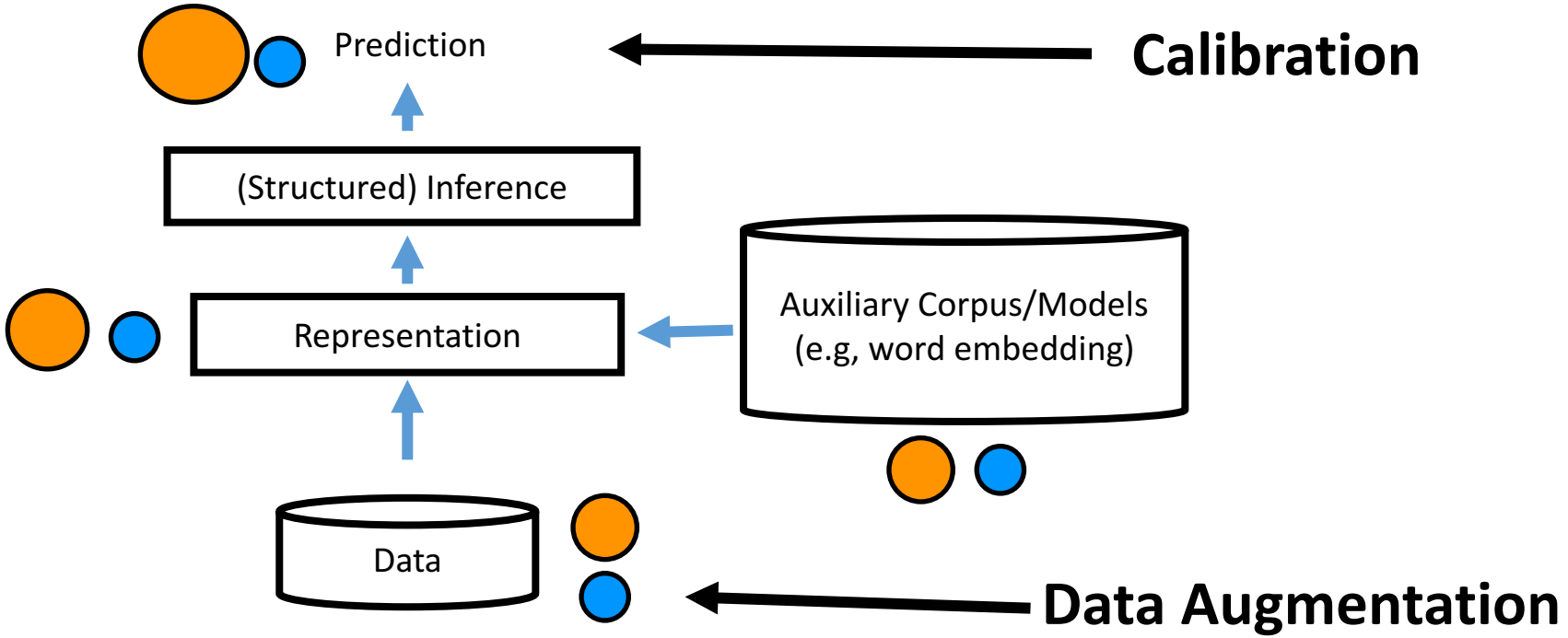
Coreference with contextualized embedding

- ELMo boosts the performance
- However, enlarge the bias (Δ)



Should We Debias Word Embedding?

- Awareness is better than blindness (Caliskan et. al. 17)
- Completely removing bias from embedding is hard if not impossible (Gonen&Goldberg 19)



Data Augmentation-- Balance the data

- Gender Swapping -- simulate sentence in opposite gender

John went to ~~his~~ house

F2 went to her house

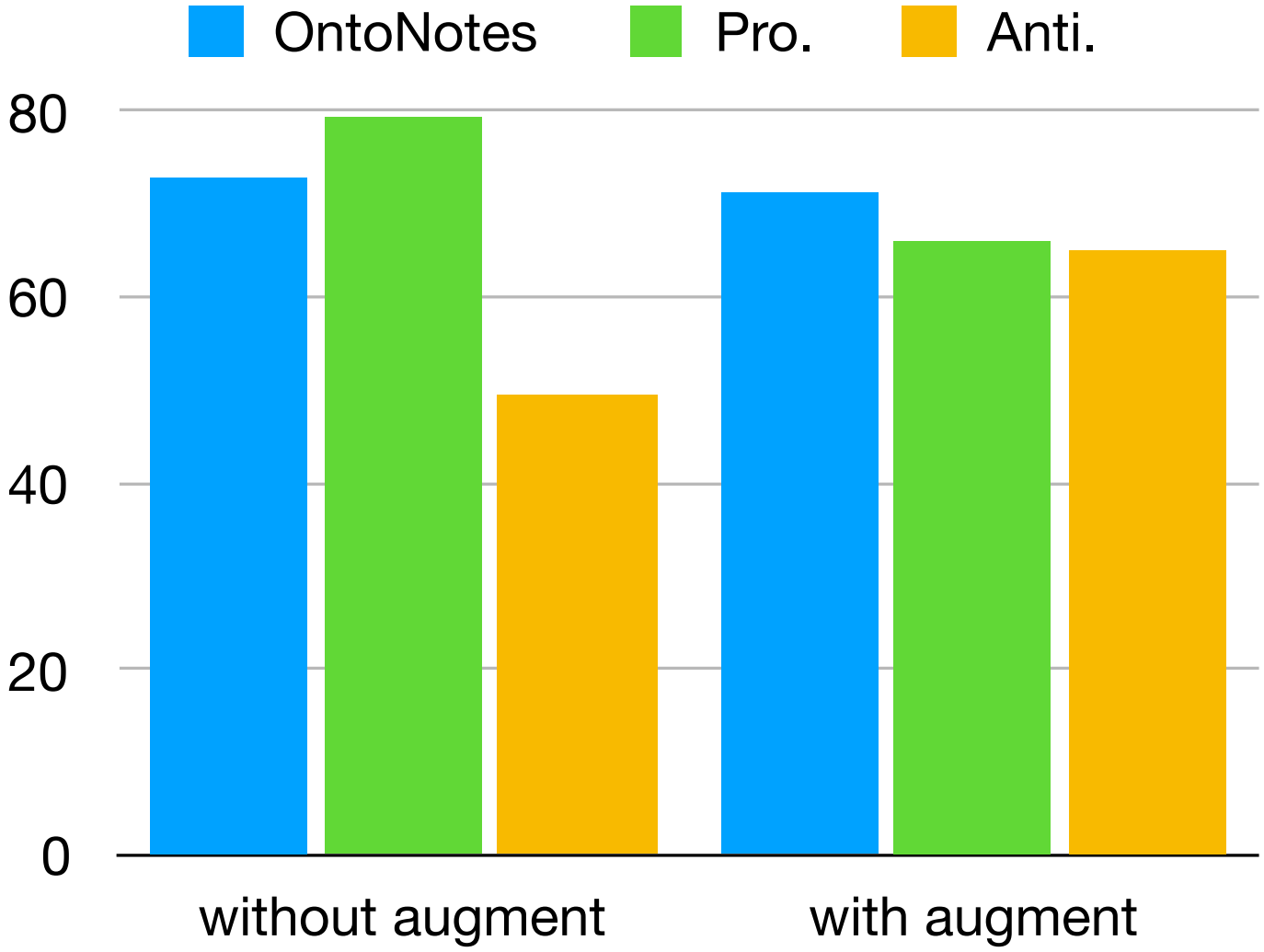
Named Entity are anonymized

Gender words are swapped

Better than down/up sampling

This idea has been used in computer vision as well

Reduce Bias via Data Augmentation in Coreference Resolution



Bias Calibration -- Visual-and-Language Models

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

[EMNLP 17*] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang

What's the agent for this image?

Cooking	
Role	Object
agent	?
food	vegetable
container	bowl
tool	knife
place	kitchen

An example from a vSRL (visual Semantic Role Labeling) system

*Best Long Paper Award at EMNLP 17

Dataset Gender Bias

33%

66%

Male

Female

imsitu.org

2

Model Bias After Training

16%

84%

Male

Female

imsitu.org

3

imSitu Visual Semantic Role Labeling (vSRL)

Convolutional Neural Network

COOKING (events)	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula

Regression

Conditional Random Field

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)

Yatskar et al. CVPR '16, Yang et al. NAACL '16, Gupta and Malik arXiv '16

Leakage of Gender

Adversarial Removal of Gender from Deep Image Representations [WZYCO19; ICCV 19]

Defining Dataset Bias (events)

Training Gender Ratio (verb)

Training Set

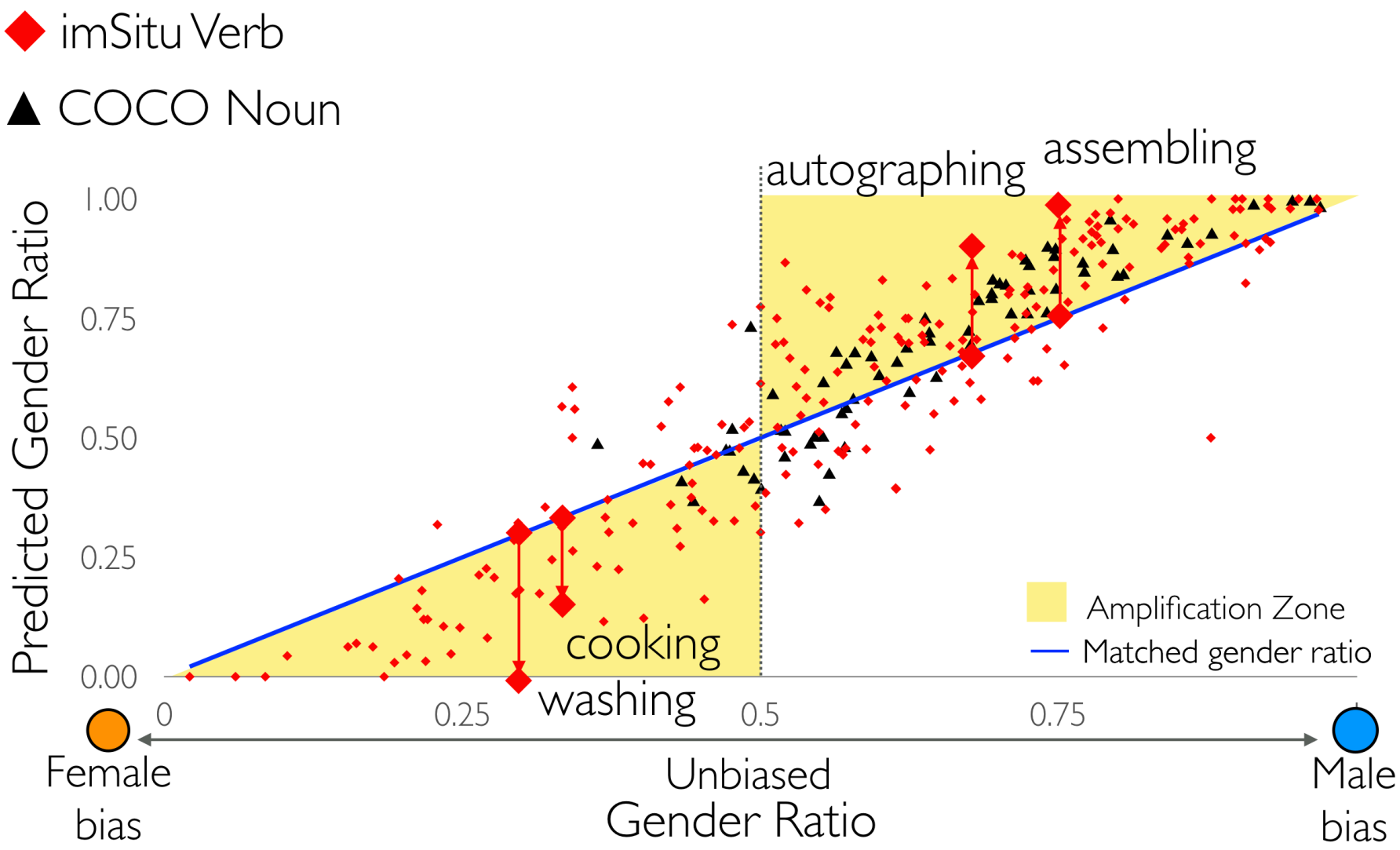
- cooking
- woman
- man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

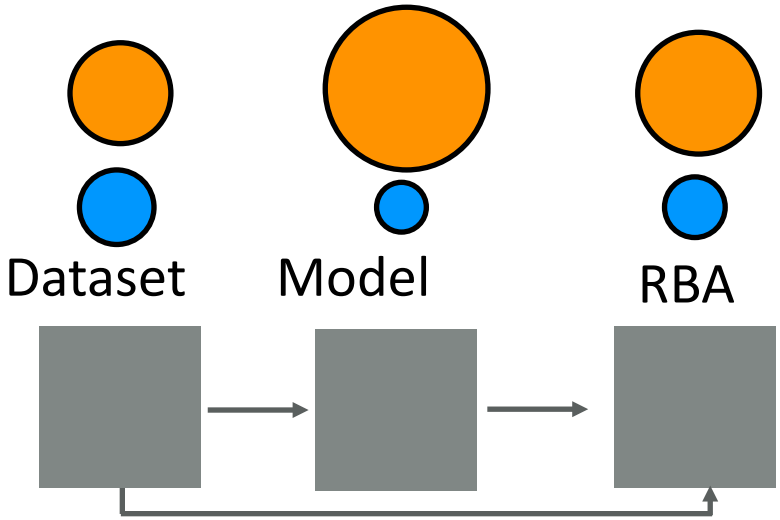
COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

$$\frac{\#(\text{cooking}, \text{man})}{\#(\text{cooking}, \text{man}) + \#(\text{cooking}, \text{woman})} = 1/3$$

Model Bias Amplification



Reducing Bias Amplification (RBA)



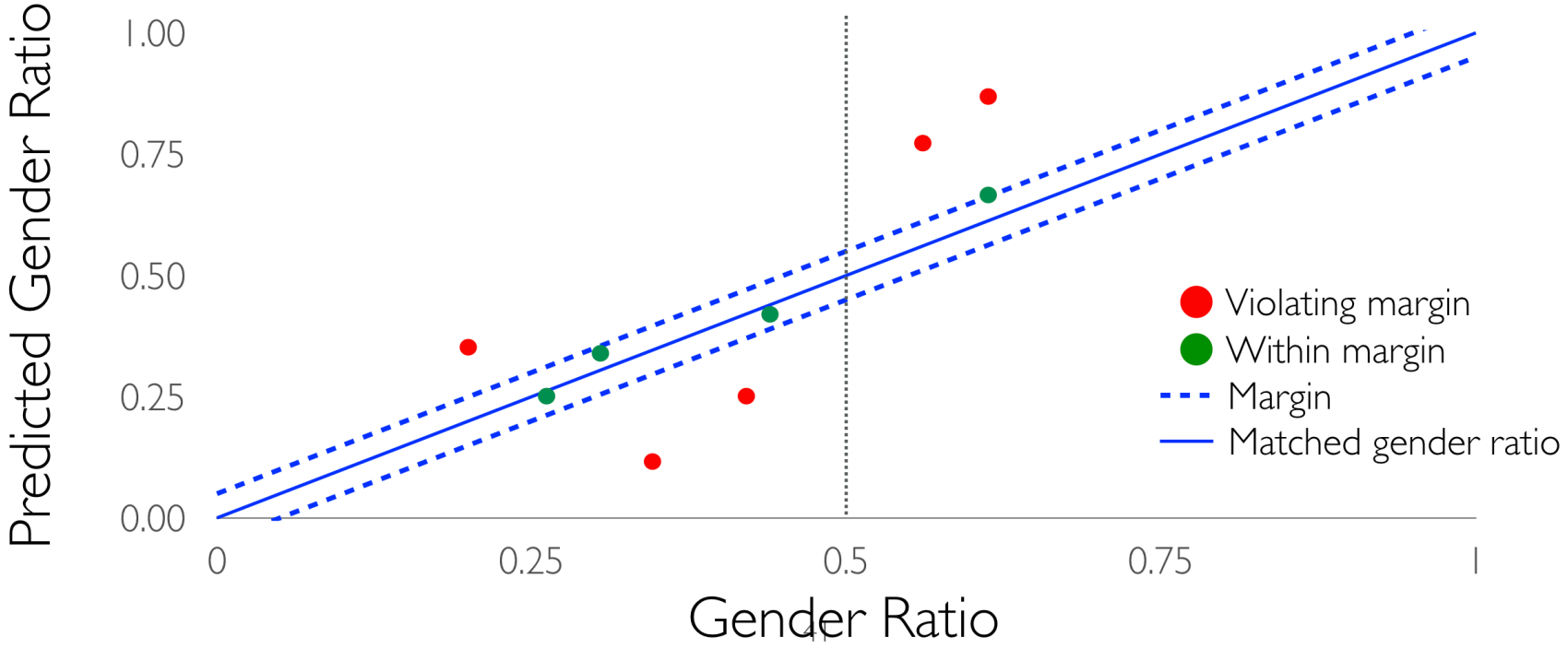
- Corpus-level constraints on model output (ILP)
  - Doesn't require model retraining
- Reuse model inference through Lagrangian relaxation
  - Can be applied to any structured model

Reducing Bias Amplification (RBA)

Integer Linear Program

$$\sum_{i} \max_{y_i} s(y_i, \text{image})$$

$$\forall \text{ points} \quad \left| \text{Training Ratio} - \underset{f(y_1 \ldots y_n)}{\text{Predicted Ratio}} \right| <= \text{margin}$$

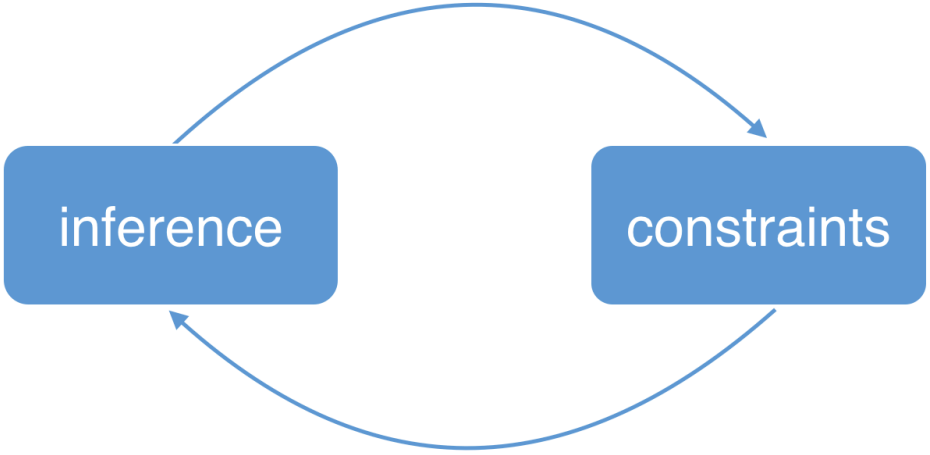


Reducing Bias Amplification (RBA)

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\forall \text{ points} \quad \left| \text{Training Ratio} - \text{Predicted Ratio}_{f(y_1 \ldots y_n)} \right| \quad <= \quad \text{margin}$$

Lagrangian Relaxation



Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

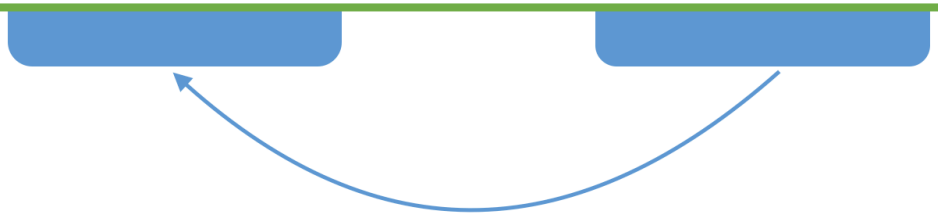
Reducing Bias Amplification (RBA)

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$\forall$ points | Training Ratio - Predicted Ratio | <= margin

$$\max_{\{y^i\} \in \{Y^i\}} \sum_i f_\theta(y^i, i), \quad \text{s.t.} \quad A \sum_i y^i - b \leq 0$$

Lagrangian : $\sum_i f_\theta(y^i) - \sum_{j=1}^{l} \lambda_j (A_j \sum_i y^i - b_j) \quad \lambda_j \geq 0$



Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Lagrangian Relaxation

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| <= \text{margin}$$

$$(1/2)$$

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Lagrangian Relaxation

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| <= \text{margin}$$

(1/2)

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Lagrangian Relaxation

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio}^{(1/2)} - \text{Predicted Ratio} \right| <= \text{margin}$$

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Lagrangian Relaxation

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	man
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| <= \text{margin}$$

$$(1/2)$$

- Lagrange Multiplier (λ) Per Constraint

inference

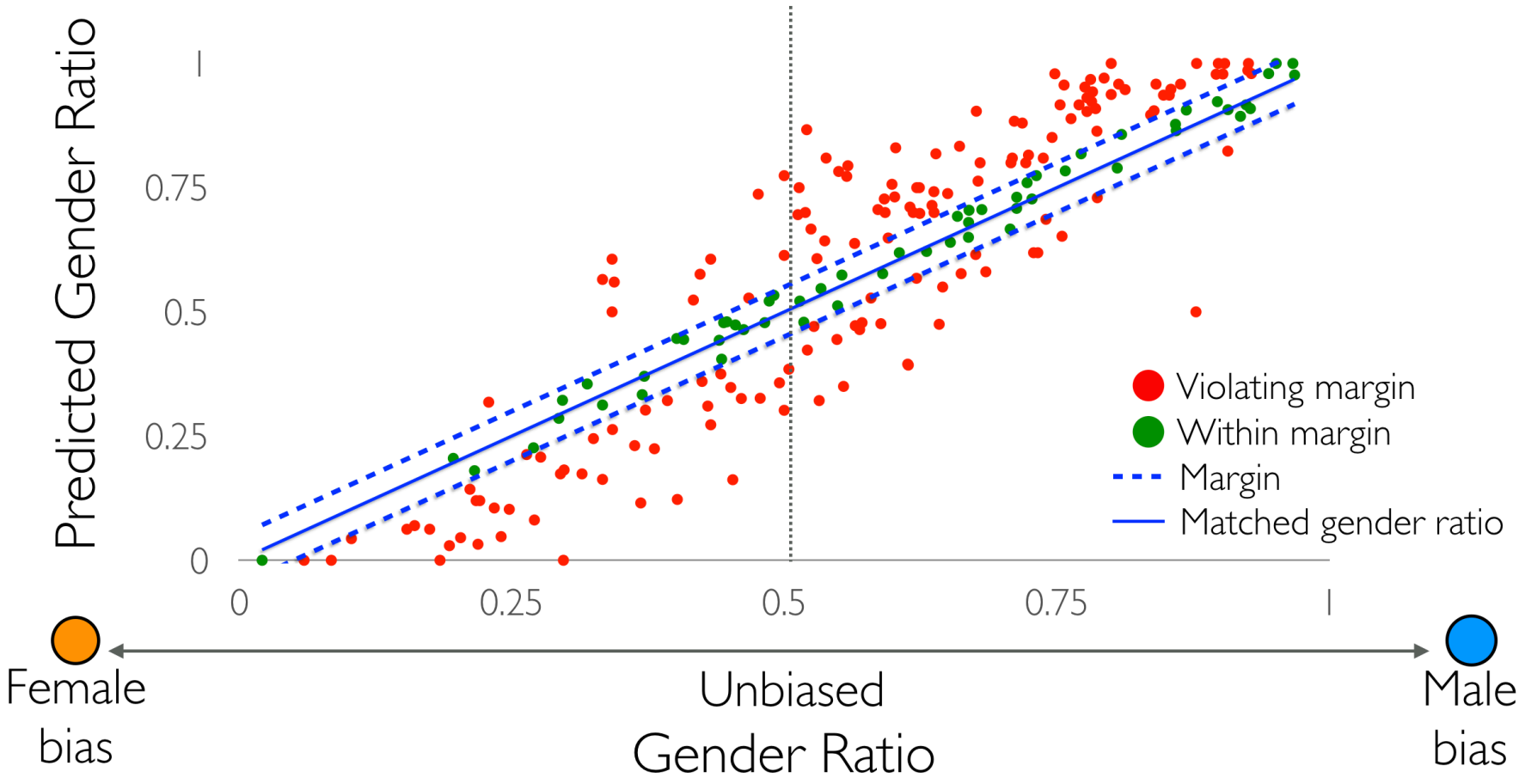
update λ

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

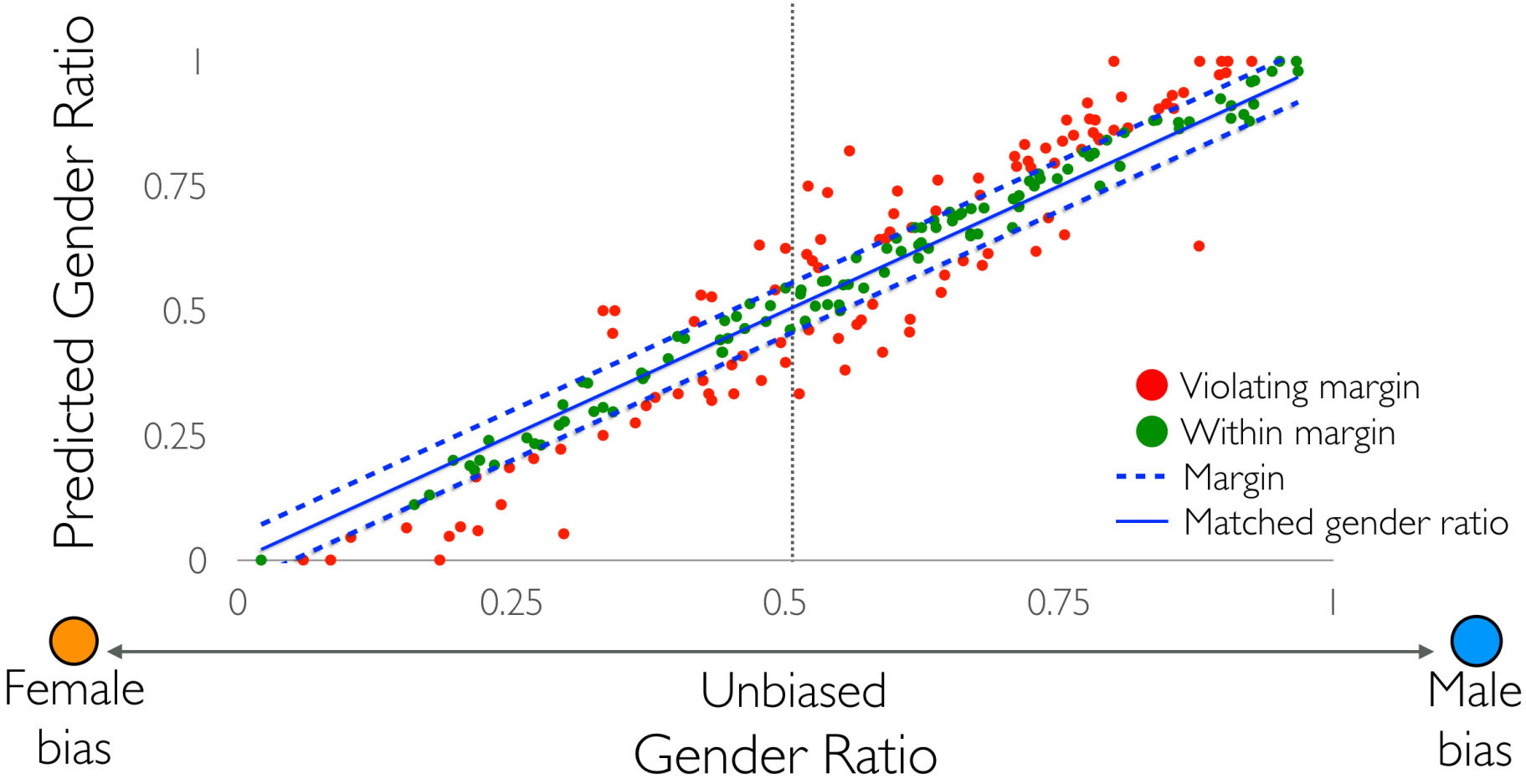
Gender Bias De-amplification in imSitu

imSitu Verb Violation: 72.6% .050 |bias↑| 24.07 acc.



Gender Bias De-amplification in imSitu

imSitu Verb	Violation: 72.6%	.050 |bias↑|	24.07	acc.
w/ RBA	Violation: 50.5%	.024 |bias↑|	23.97	acc.



Outline

- Gender Bias in NLP
  - Representational harm
  - Performance gap in downstream applications

arXiv.org > cs > arXiv:1906.08976

Computer Science > Computation and Language

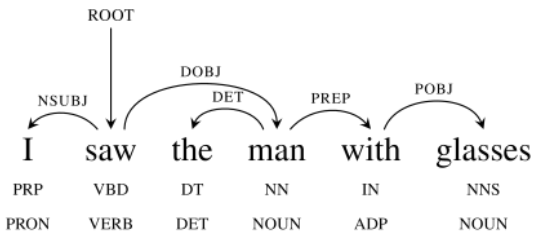
Mitigating Gender Bias in Natural Language Processing: Literature Review

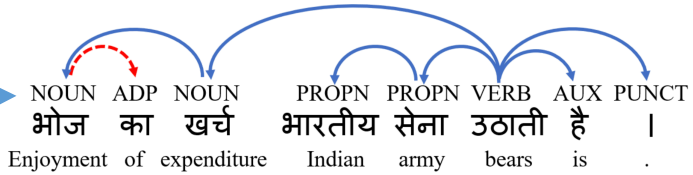
Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, William Yang Wang

(Submitted on 21 Jun 2019)

[ACL 2019]

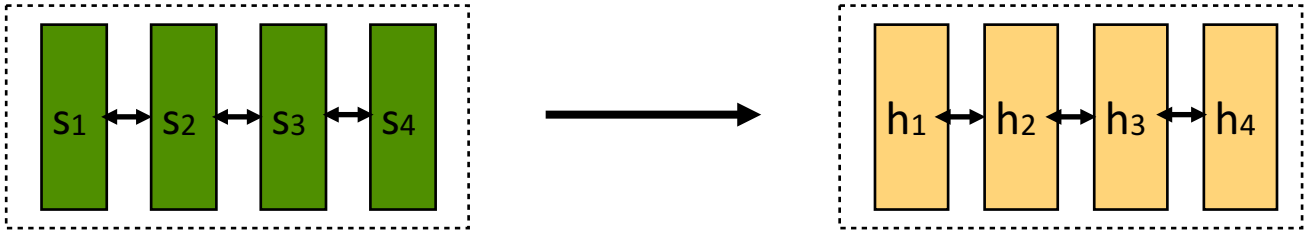
- Cross-lingual Dependency Parsing



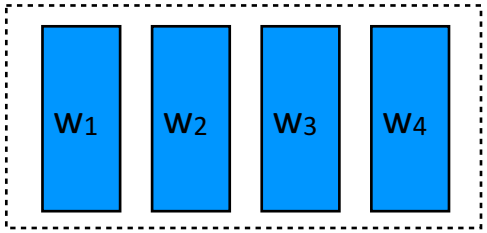


Standard Neural Architectures for NLP

An encoder to produce contextualized representations



Embeddings for the input sentence



A decoder that makes (structured) predictions

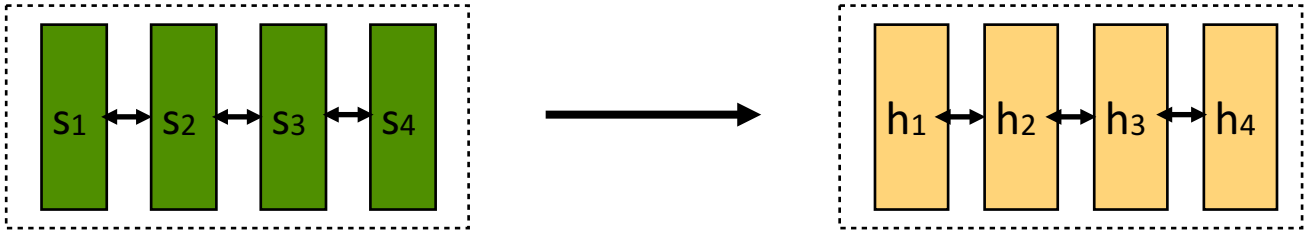
$$\mathbf{y} = \{p_1, \ldots, p_n\}$$

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)

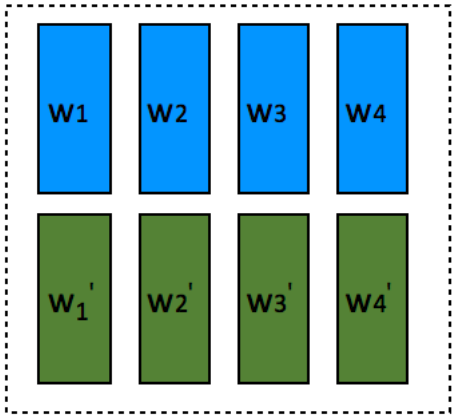
Credit: Nanyun Peng

Standard Neural Architectures for NLP

An encoder to produce contextualized representations



Multi-Lingual Embeddings for the input sentence



A decoder that makes (structured) predictions

$$\mathbf{y} = \{p_1, \ldots, p_n\}$$

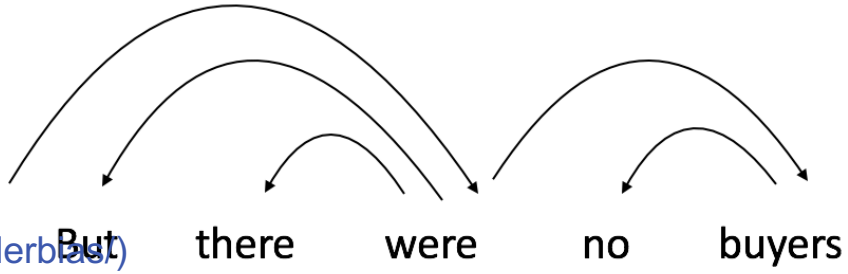
Zero-shot multi-lingual transfer

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)

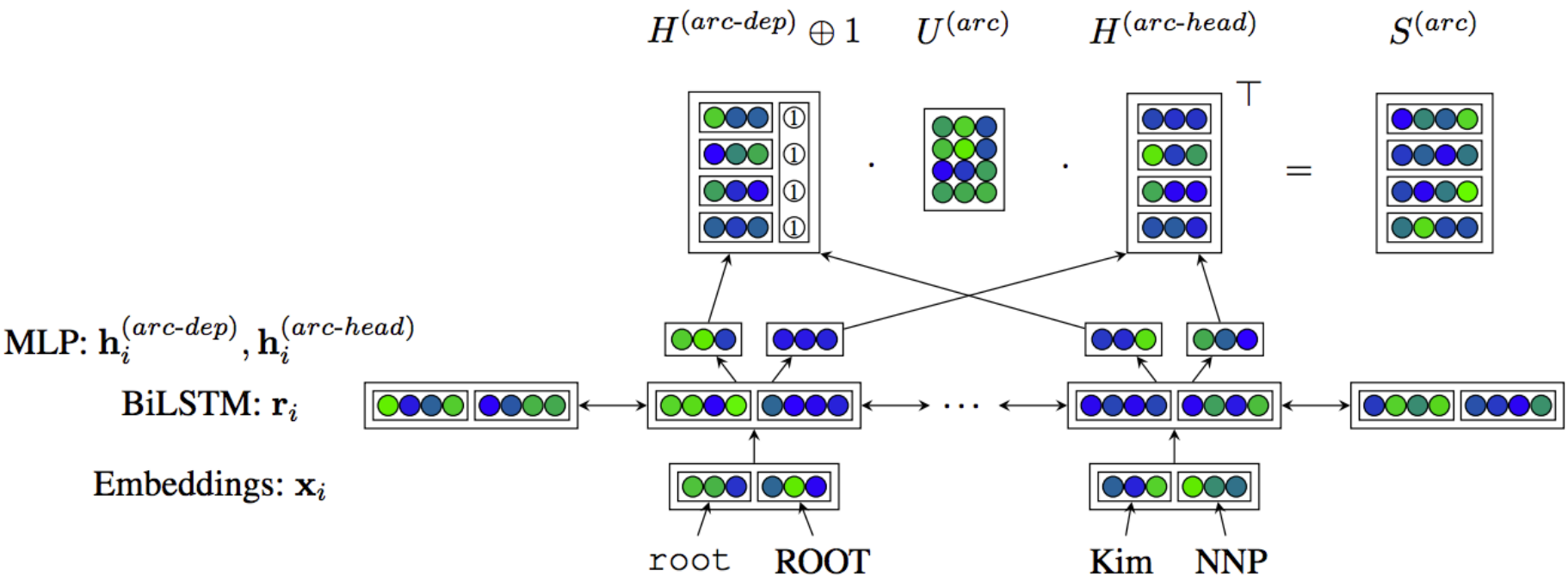
Credit: Nanyun Peng

Cross-lingual Transfer for Decency Parsing

- Examine and verify our hypothesis on cross-lingual dependency parsing
  - UD annotation for over 70 languages
  - Parser is a low-level task that reflects the problems
- Remove language-specific knowledge (e.g., word order) from encoder
- Add language-specific knowledge to decoder



Background: Deep Biaffine Parser

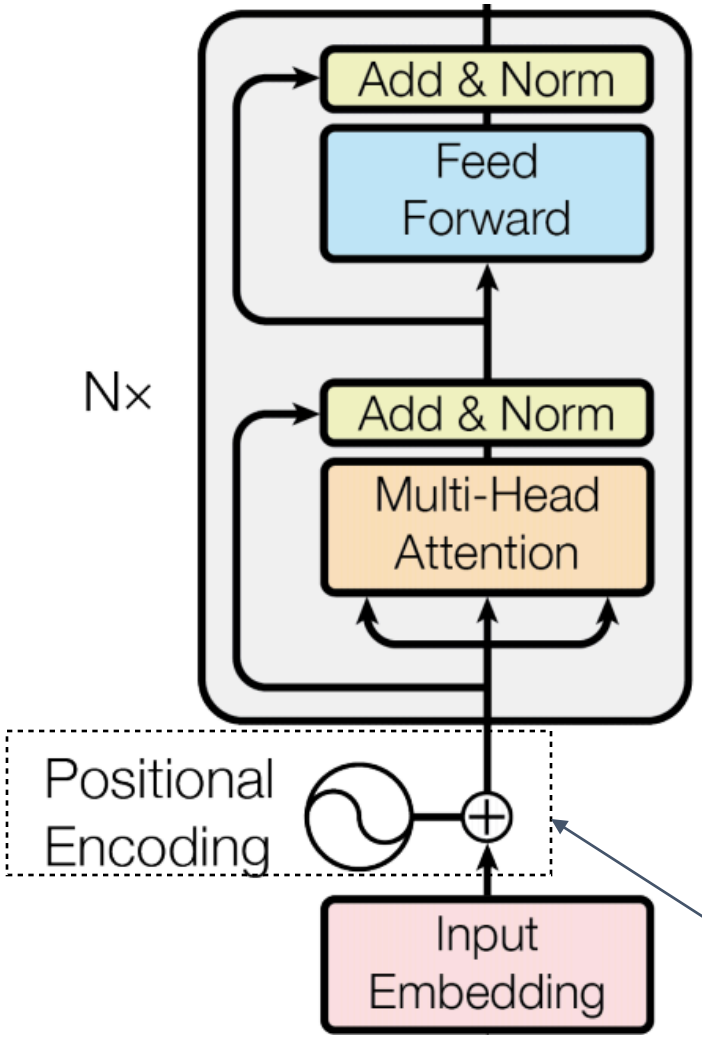


- **Graph-based** parser
- Encoder: RNN (Order-sensitive); Decoder: Graph (Order-free)

Dozat and Manning (ICLR2017)

Remove Word Order information -- Multi-Head Self-Attention with Relative Position

[WZMCN NAACL 19]



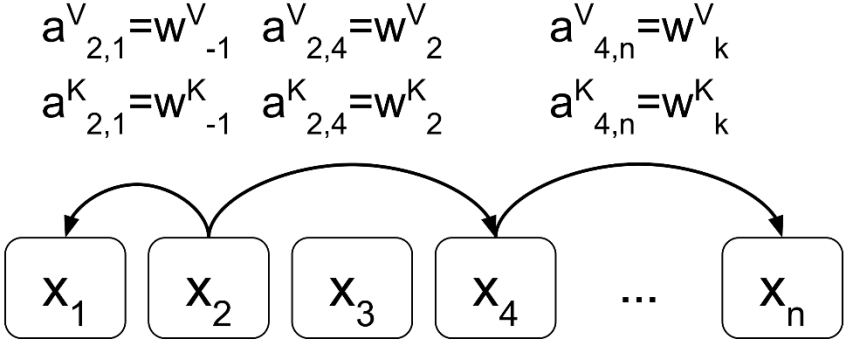
- In the original paper:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Vaswani et. al. (NIPS 2017)

- Encoder absolute distance



Shaw et. al. (NAACL2018)

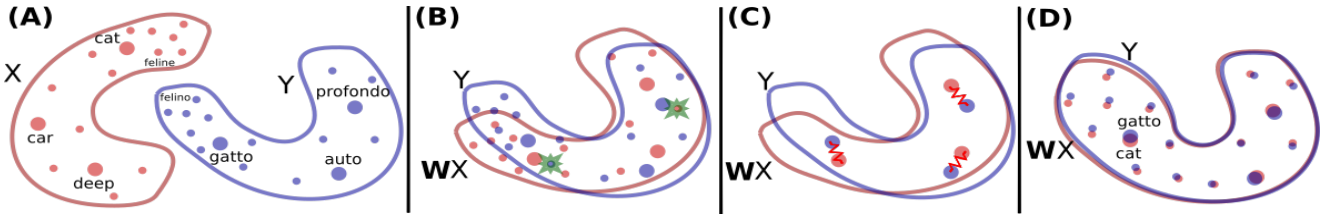
Flexible positional encoding (order-free)

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)

Architectures for Cross-lingual Parser

- Embedding

Facebook MUSE



Conneau et. al. ICLR2018

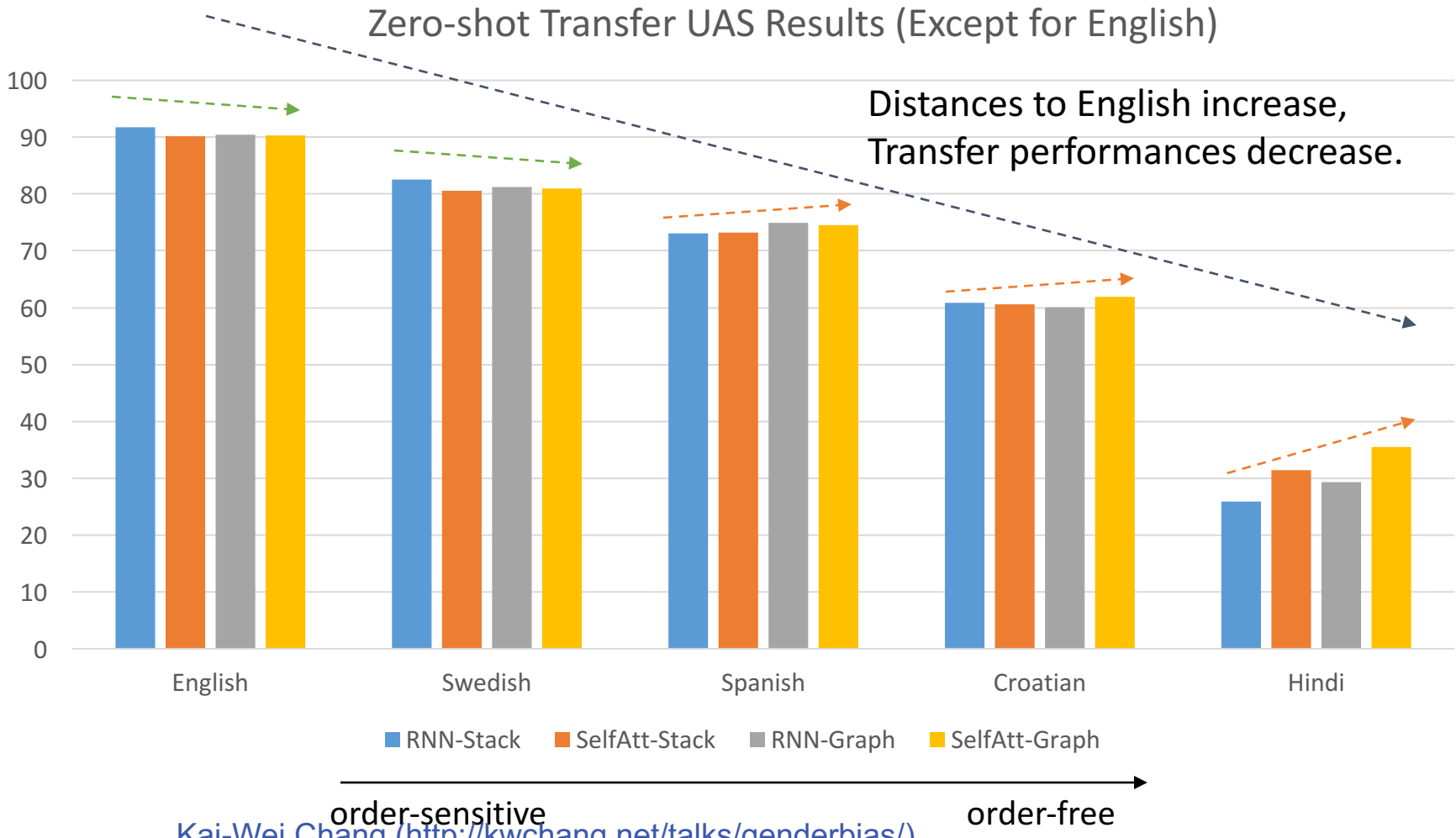
- Encoders
  - BiLSTMs (order-sensitive) v.s.
  - Multi-Head Self-Attention with Absolute Relative Positional Encoding (order-free)
- Decoders
  - Pointer Network (order-sensitive) v.s.
  - BiAffine Attention (order-free)

Experiments

- Datasets:
  - UD (V2.2)
  - 31 languages, 12 families
- Setting:
  - Train/Dev on English
  - Directly predict on the res
    30 languages (zero-shot)

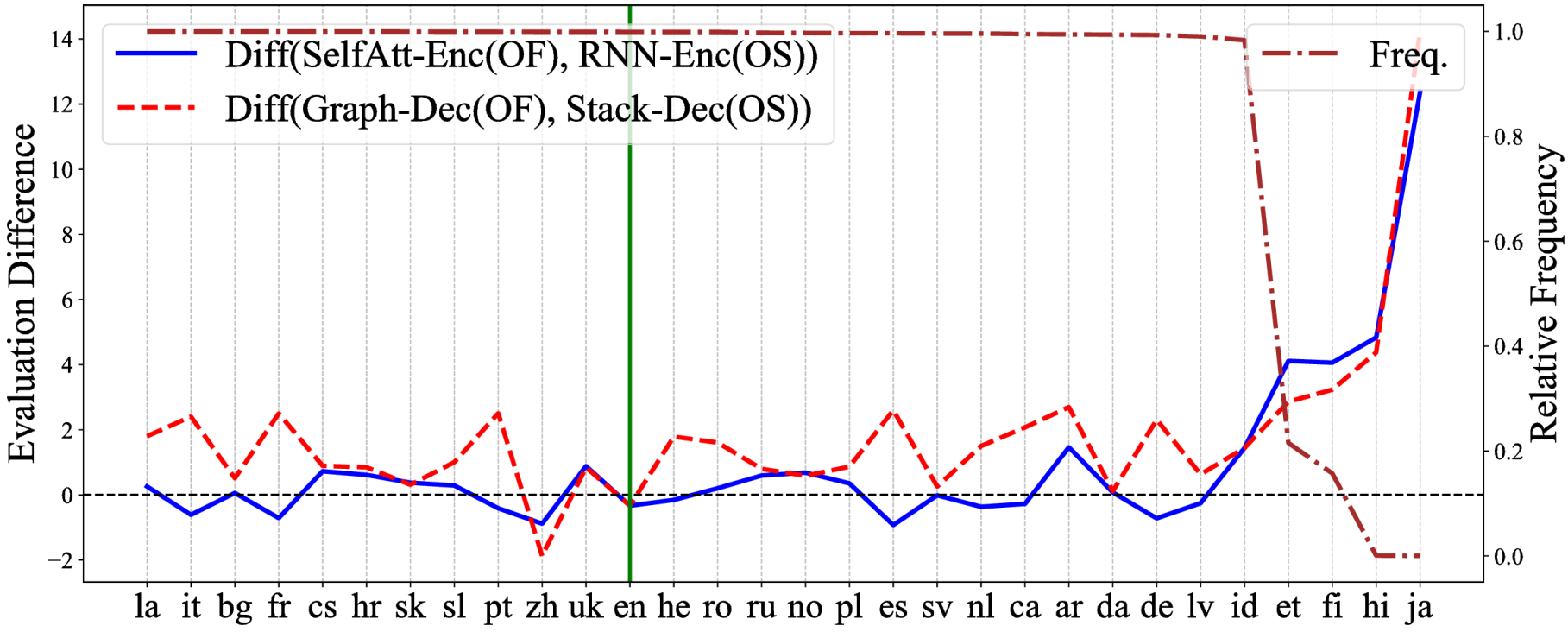
Language Families	Languages
Afro-Asiatic	Arabic (ar), Hebrew (he)
Austronesian	Indonesian (id)
IE.Baltic	Latvian (lv)
IE.Germanic	Danish (da), Dutch (nl), English (en), German (de), Norwegian (no), Swedish (sv)
IE.Indic	Hindi (hi)
IE.Latin	Latin (la)
IE.Romance	Catalan (ca), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Spanish (es)
IE.Slavic	Bulgarian (bg), Croatian (hr), Czech (cs), Polish (pl), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)
Japanese	Japanese (ja)
Korean	Korean (ko)
Sino-Tibetan	Chinese (zh)
Uralic	Estonian (et), Finnish (fi)

Selected Transfer Results of Different Architectures



Kai-Wei Chang (http://kwchang.net/talks/genderbias/)

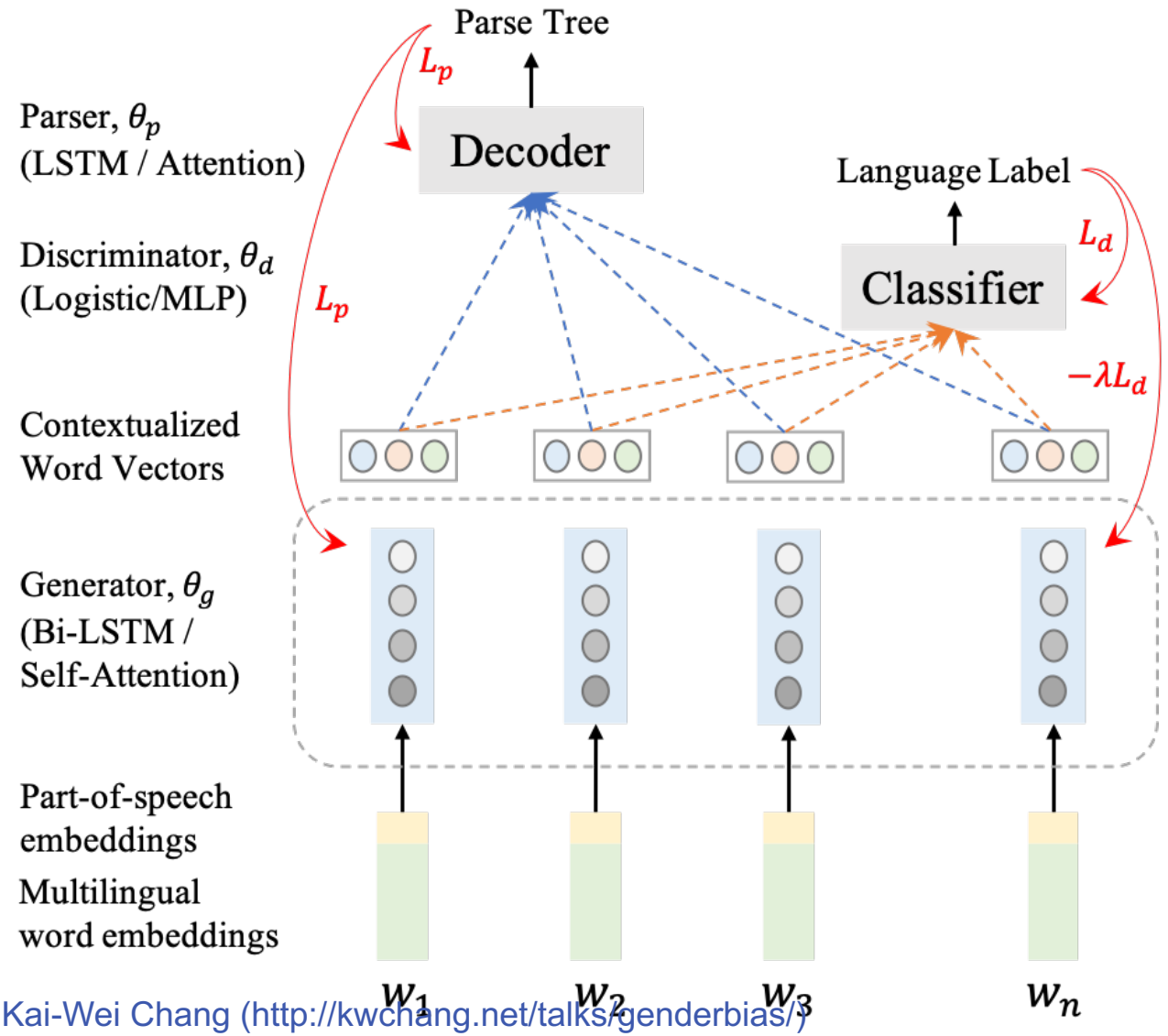
Case Study -- Adposition
Preposition v.s. postposition



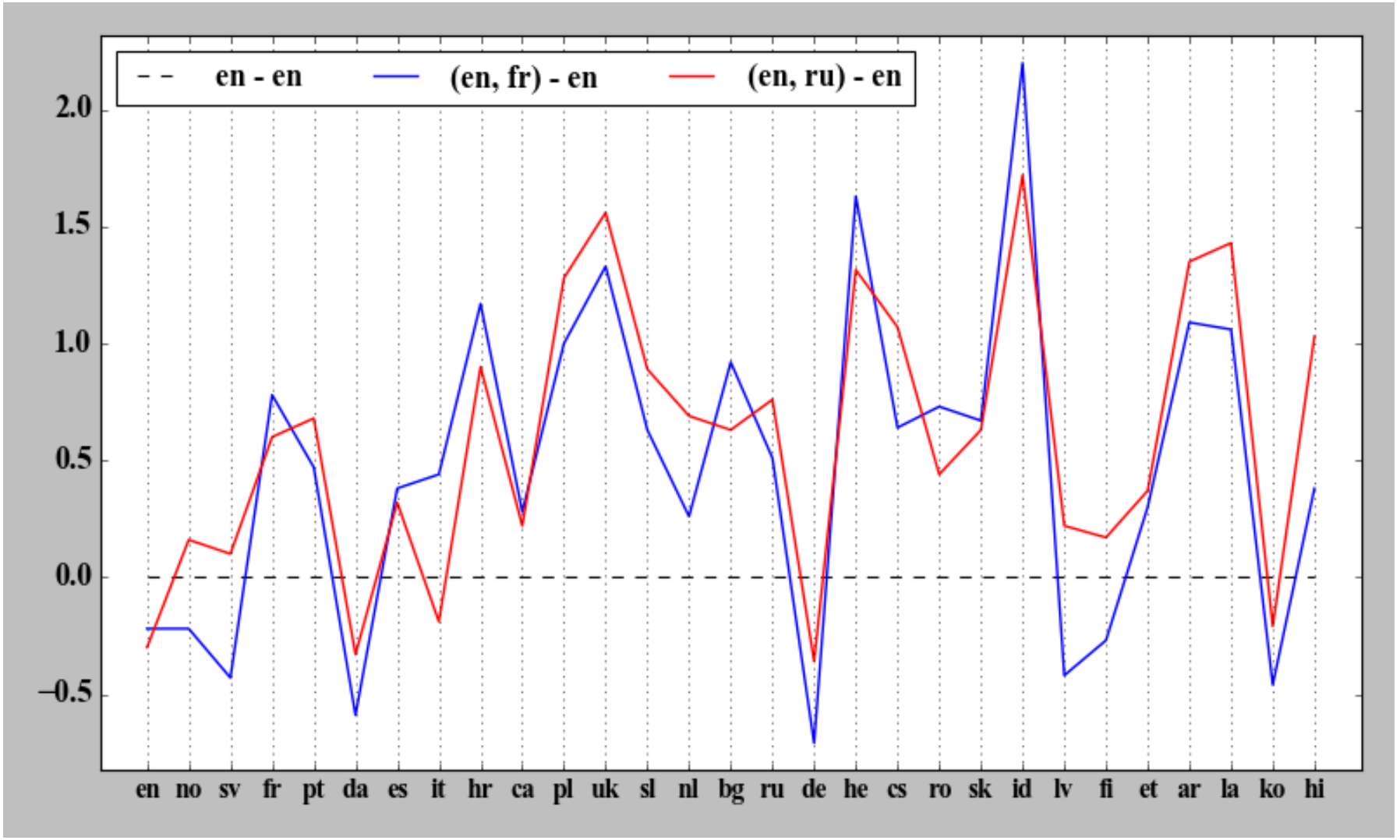
The languages (x-axis) are sorted by this relative frequency from high to low

Adversarial Learning for Removing Language-specific Information

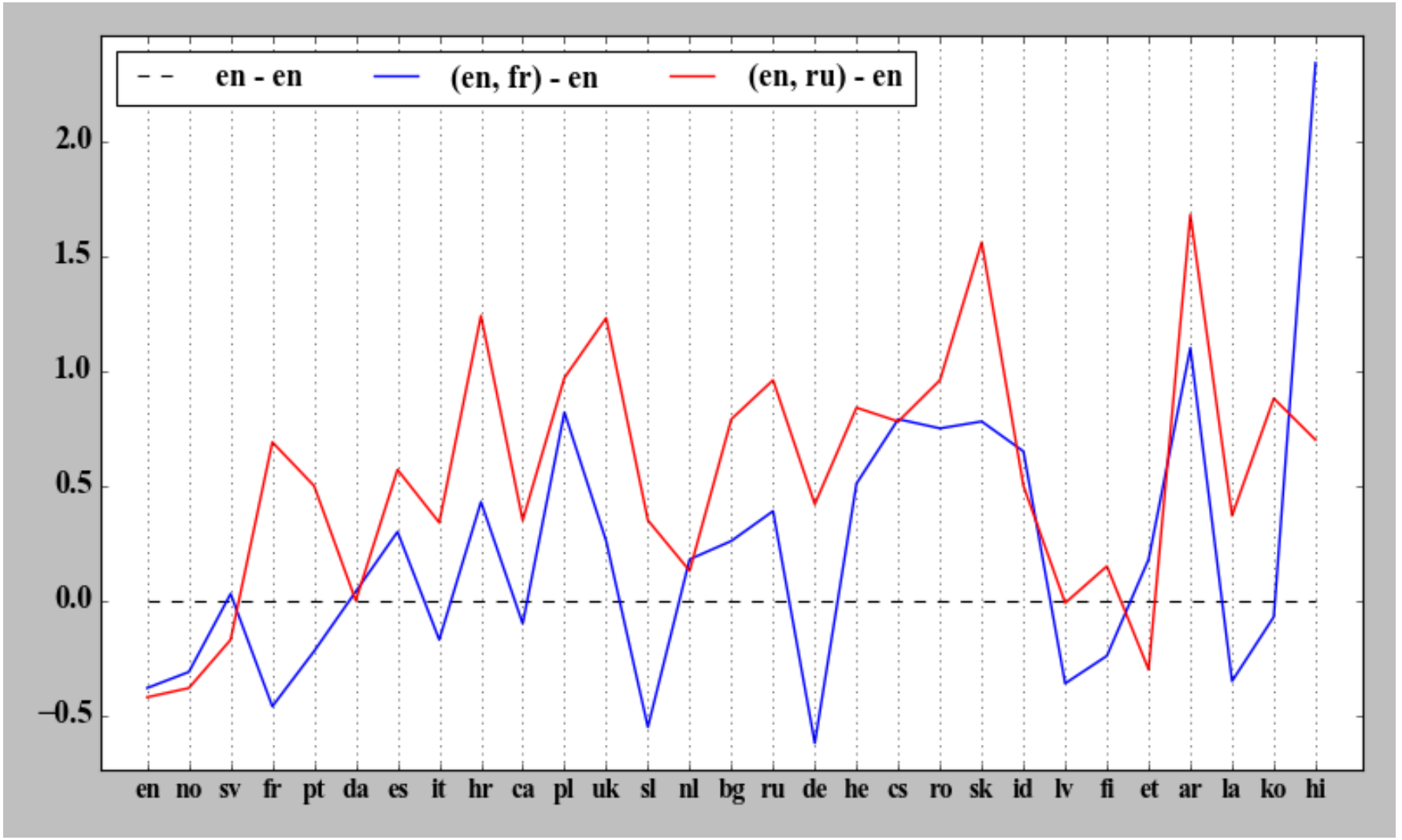
[WZMCN CoNLL 19]



Cross-lingual transfer with Mulitilingual embedding



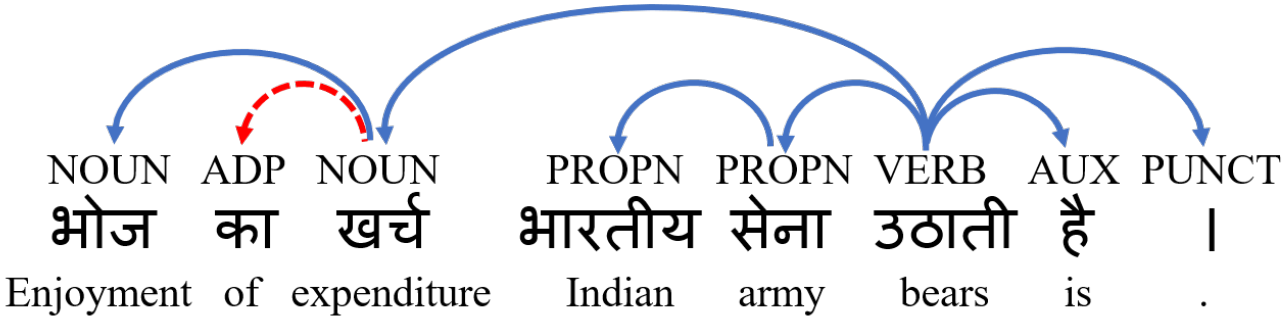
Cross-lingual transfer with Mulitilingual BERT



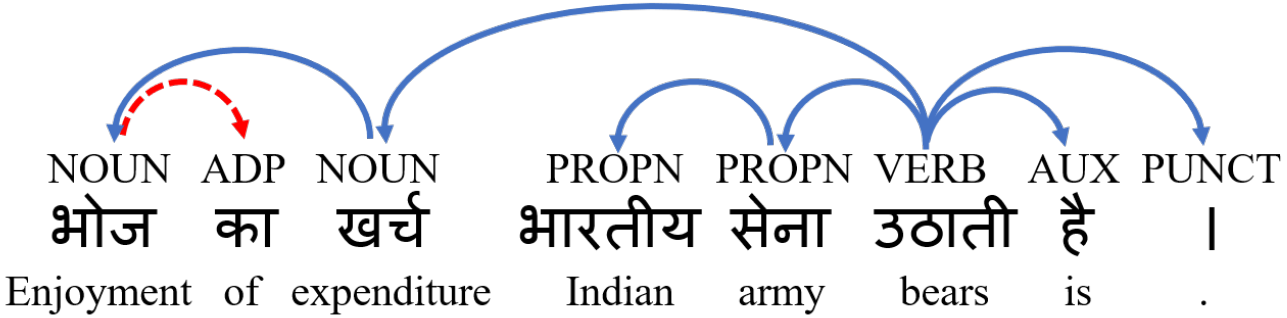
Decoding with Language-Specific Knowledge

[TPC EMNLP 19]

- Leveraging linguistics knowledge (WALS features) in decoding



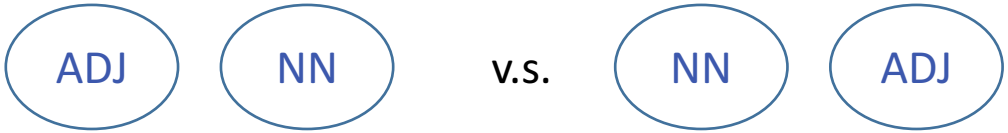
Constraint: In an ADP-NOUN arc in Hindi, ADP is more likely to be on the right.



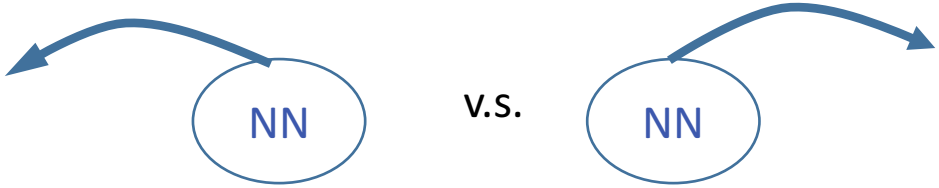
Corpus-Statistics Constraints

[TPC EMNLP 19]

- Consider constraints in two forms:
  - specifies the ratio r of POS1 being on the left in all POS1-POS2 arcs

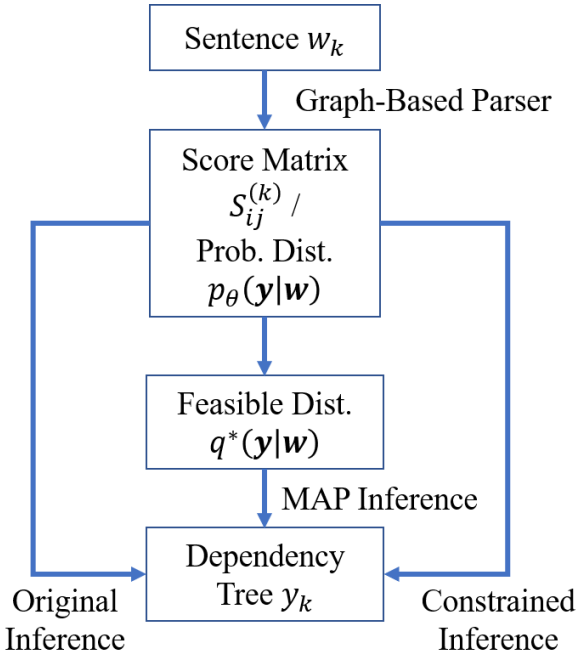


  - specifies the ratio r of the heads of a particular POS appears on the left of that POS

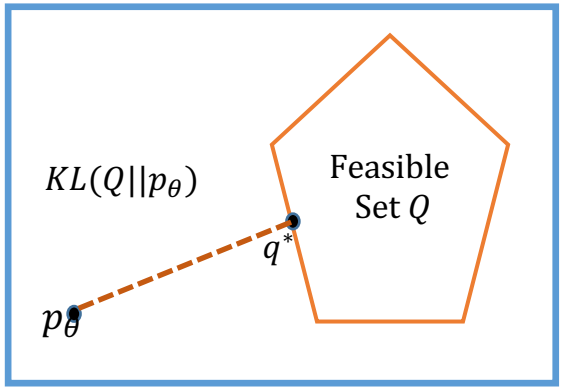


- Compiling from WALS features:
  - Dominant order: e.g. 85A – Binary constraint (ADP, NN)
    - Prepositions: $r \in (0,0.25)$
    - No dominant order: $r \in (0.25,0.75)$
    - Postpositions: $r \in (0.75,1)$

Constrained Inference



- Lagrangian Relaxation
  - Introduce Lagrangian multipliers for each constraints.
  - Apply sub-gradient descent method to solve the dual form.
- Posterior Regularization
  - Use constraints to define a feasible distribution set Q.
  - Find the closest distribution $q \in Q$ from p_θ, and do MAP inference on q.



Experiments Results

[TPC EMNLP 19]

Lang.		Baseline	Lagrangian Relaxation			Posterior Regularization		
			Oracle	WALS	ΔWALS	Oracle	WALS	ΔWALS
	en	90.5	90.3	90.4	-0.1	90.4	90.6	+0.1
IE.Indic	ur	18.3	35.2	34.0	+15.7	35.0	33.7	+15.4
Dravidian	ta	36.1	42.8	43.4	+7.3	43.1	43.0	+6.9
Turkic	tr	31.2	35.2	37.1	+5.9	35.1	36.3	+5.1
Austronesian	id	49.3	53.1	52.3	+3.0	52.3	51.9	+2.6
Average Performance		54.3	58.4	57.8	+3.5	58.1	57.5	+3.1

- LR, PR get improvements in 15,17 out of 19 target languages from variant of language families, respectively
- The improvements are closely related to the ratio gap in constraints
- LR has greater average improvement, while PR is a more robust inference algorithm

Conclusions

- NLP systems affect by societal bias present in data
- How to learn/unlearn/control a model
- The issues are not new
  - Domain adaptation / Constraint Inference
- References: http://kwchang.net

Students: Jieyu Zhao, Tianlu Wang, Pei Zhou, Weijia Shi, Wasi Ahmad, Meng Tao, Moustafa Alzantot, Emily Sheng, Tony Sun, Andrew, Gaut

Collaborators: Vicente Ordonez, Nanyun Peng, Muhao Chen, Mark Yatskar, Edward Hovy, Premkumar Natarajan, Wei Wang, Mani Srivastava, Tolga Bolukbasi, James Zou, Venkatesh Saligrama, Adam Kalai, William Wang