# Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology

**ACL 2019**

---

Ran Zmigrod, **Sebastian J. Mielke**, Hanna Wallach, Ryan Cotterell

University of Cambridge // Johns Hopkins University // Microsoft Research
rz279@cam.ac.uk sjmielke@jhu.edu
wallach@microsoft.com rdc42@cam.ac.uk

Twitter: @RanZmigrod – paper and thread pinned! // @sjmielke

# Gender bias in NLP systems

Coreference resolution systems are biased:

*Even though the doctor reassured the nurse,* ***she*** *was worried.*

# Gender bias in NLP systems

Coreference resolution systems are biased:

*Even though the doctor reassured the nurse,* ***she*** *was worried.*

# Gender bias in NLP systems

Coreference resolution systems are biased:

*Even though the doctor reassured the nurse, **she** was worried.*

# Gender bias in NLP systems

Coreference resolution systems are biased:

*Even though the doctor reassured the nurse,* ***she*** *was worried.*

Both are possible...

# Gender bias in NLP systems

Coreference resolution systems are biased:

*Even though the doctor reassured the nurse,* ***she*** *was worried.*

Both are possible... but systems prefer nurse! (Rudinger et al., 2018; Zhao et al., 2018)

# Gender bias in NLP systems

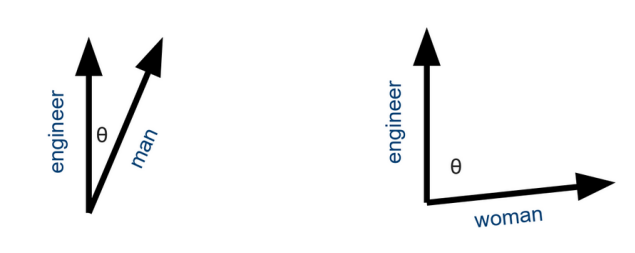
Coreference resolution systems are biased:

*Even though the doctor reassured the nurse,* ***she*** *was worried.*

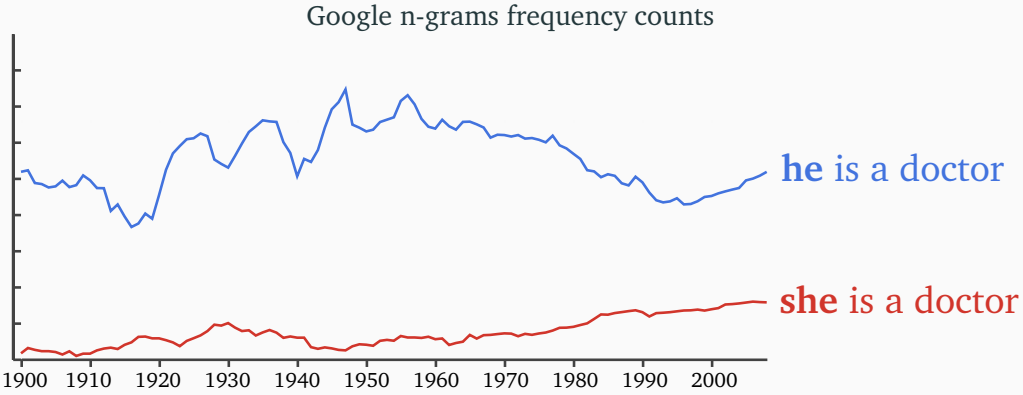
Both are possible... but systems prefer nurse! (Rudinger et al., 2018; Zhao et al., 2018)

---

Word embeddings carry biases:



# This shouldn't come as a surprise: our data is biased



# Our focus: stereotypes in language modeling (Lu et al., 2018)

Training data counts are visible as likelihoods under a language model:

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | He is a good doctor. | He is a good nurse. |
| | f | She is a good doctor. | She is a good nurse. |

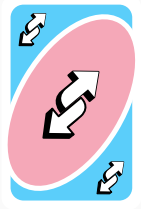
# Our focus: stereotypes in language modeling (Lu et al., 2018)

Training data counts are visible as likelihoods under a language model:

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | He is a good doctor. | He is a good nurse. |
| | f | She is a good doctor. | She is a good nurse. |

---

The solution:
**C**ounterfactual
**D**ata
**A**ugmentation
(Lu et al., 2018)



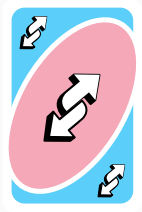
# Our focus: stereotypes in language modeling (Lu et al., 2018)

Training data counts are visible as likelihoods under a language model:

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | He is a good doctor. | He is a good nurse. |
| | f | She is a good doctor. | She is a good nurse. |

---

The solution: **C**ounterfactual **D**ata **A**ugmentation (Lu et al., 2018)



For every sentence with **she/he**:
e.g., “She is a nurse.”

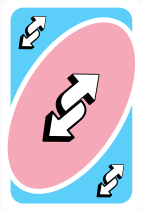
# Our focus: stereotypes in language modeling (Lu et al., 2018)

Training data counts are visible as likelihoods under a language model:

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | He is a good doctor. | He is a good nurse. |
| | f | She is a good doctor. | She is a good nurse. |

---

The solution:
**C**ounterfactual
**D**ata
**A**ugmentation
(Lu et al., 2018)



For every sentence with **she/he**:
e.g., “She is a nurse.”
add that sentence with **he/she** for training:
e.g., “He is a nurse.”

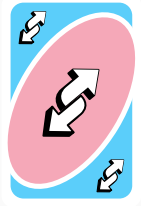
# Our focus: stereotypes in language modeling (Lu et al., 2018)

Training data counts are visible as likelihoods under a language model:

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | He is a good doctor. | He is a good nurse. |
| | f | She is a good doctor. | She is a good nurse. |

---

The solution: **C**ounterfactual **D**ata **A**ugmentation (Lu et al., 2018)



For every sentence with **she/he**:
e.g., “She is a nurse.”
add that sentence with **he/she** for training:
e.g., “He is a nurse.”
Now they should yield a balanced model!

**Sebastian J. Mielke**
@sjmielke

Follow

Reading #NLProc papers in the NYC subway, thinking about the #Benderrule 🙃


12:12 PM - 19 Jun 2019

**16** Likes

16

# “Agreement” or “what if: German”

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | **Er** ist ein guter Arzt. | **Er** ist ein guter Krankenpfleger. |
| | f | **Sie** ist eine gute Ärztin. | **Sie** ist eine gute Krankenpflegerin. |

# “Agreement” or “what if: German”

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | **Er** ist ein guter Arzt. | **Er** ist ein guter Krankenpfleger. |
| | f | **Sie** ist eine gute Ärztin. | **Sie** ist eine gute Krankenpflegerin. |

# “Agreement” or “what if: German”

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | **Er** ist ein guter Arzt. | **Er** ist ein guter Krankenpfleger. |
| | f | **Sie** ist eine gute Ärztin. | **Sie** ist eine gute Krankenpflegerin. |

# “Agreement” or “what if: German”

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | **Er** ist ein guter Arzt. | **Er** ist ein guter Krankenpfleger. |
| | f | **Sie** ist eine gute Ärztin. | **Sie** ist eine gute Krankenpflegerin. |

# “Agreement” or “what if: German”

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | **Er** ist ein guter Arzt. | **Er** ist ein guter Krankenpfleger. |
| | f | **Sie** ist eine gute Ärztin. | **Sie** ist eine gute Krankenpflegerin. |

So, uh, can we just… change all words’ grammatical gender?

# “Agreement” or “what if: German”

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | **Er** ist ein guter Arzt. | **Er** ist ein guter Krankenpfleger. |
| | f | **Sie** ist eine gute Ärztin. | **Sie** ist eine gute Krankenpflegerin. |

So, uh, can we just… change all words’ grammatical gender?

**Example:** Der Arzt sitzt auf einem Stuhl *(The male doctor sits on a chair)*

# “Agreement” or “what if: German”

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | **Er** ist ein guter Arzt. | **Er** ist ein guter Krankenpfleger. |
| | f | **Sie** ist eine gute Ärztin. | **Sie** ist eine gute Krankenpflegerin. |

So, uh, can we just… change all words’ grammatical gender?

**Example:** Der Arzt sitzt auf einem Stuhl *(The male doctor sits on a chair)*

**Swap all:** Die Ärztin sitzt auf einer Stuhl

# “Agreement” or “what if: German”

| | | stereotype | |
|---|---|---|---|
| | | m | f |
| pronoun | m | **Er** ist ein guter Arzt. | **Er** ist ein guter Krankenpfleger. |
| | f | **Sie** ist eine gute Ärztin. | **Sie** ist eine gute Krankenpflegerin. |

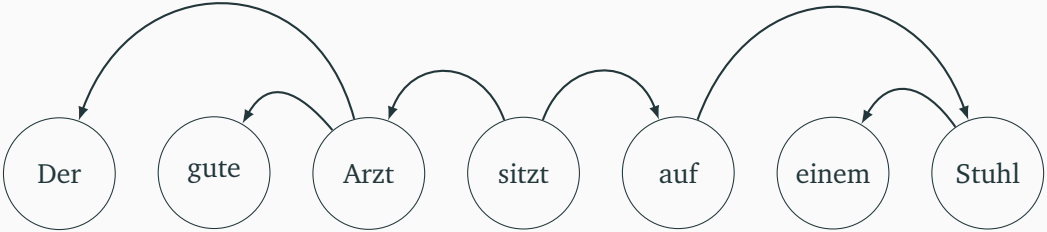
So, uh, can we just… change all words’ grammatical gender?

**Example:** Der Arzt sitzt auf einem Stuhl *(The male doctor sits on a chair)*

**Swap all:** Die Ärztin sitzt auf ~~einer Stuhl~~ *(The female doctor sits on a… what?)*

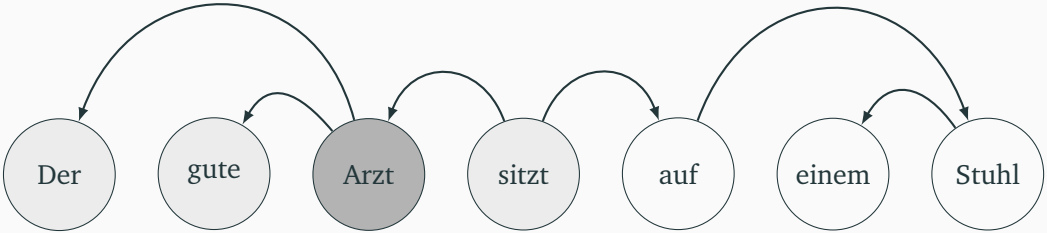
No, what we need is…

# Syntax to the rescue: use dependency parses



# Syntax to the rescue: use dependency parses

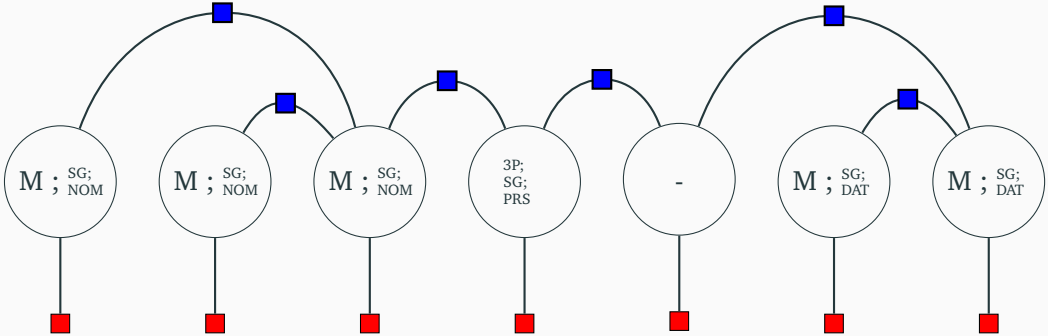
Only words “connected” in the dependency parse should change!



# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

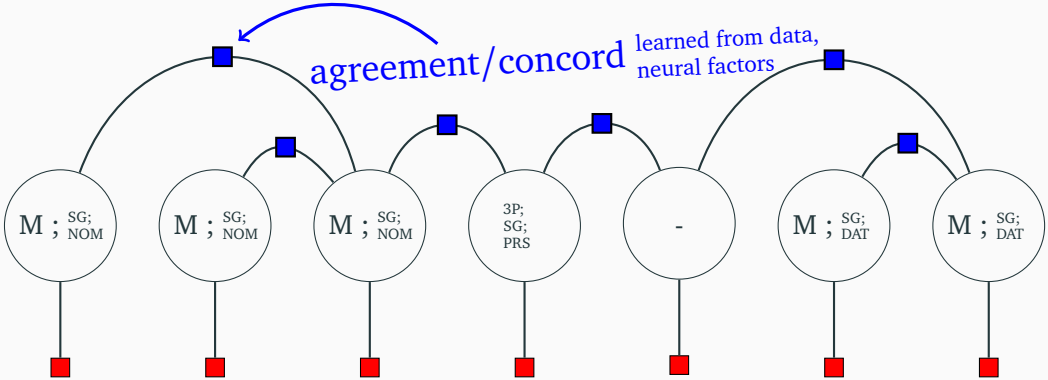
Build a MRF over morphological tags along the dependency parse!



# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

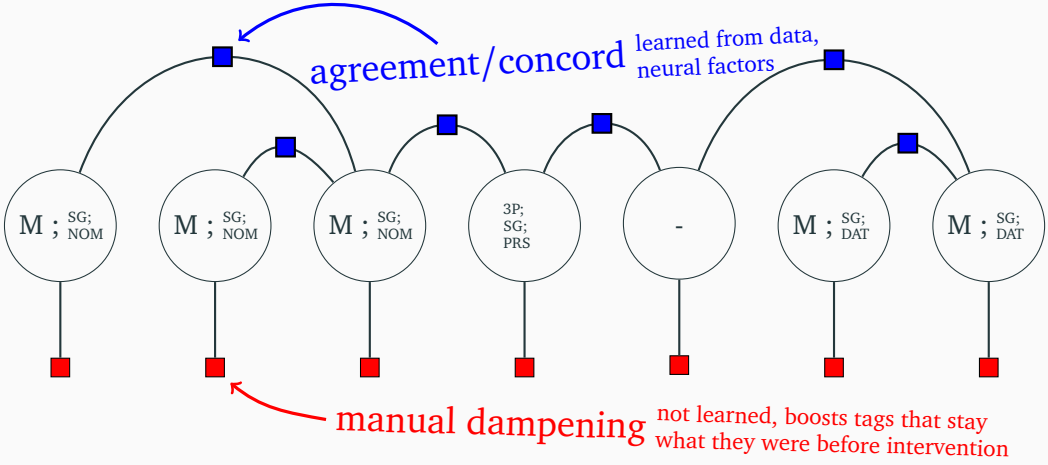
Build a MRF over morphological tags along the dependency parse!



# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

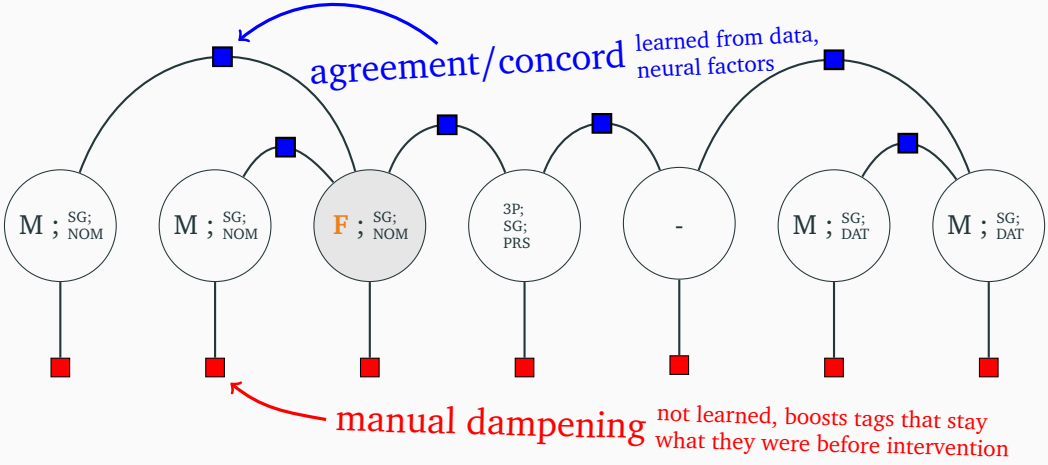
Build a MRF over morphological tags along the dependency parse!



# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

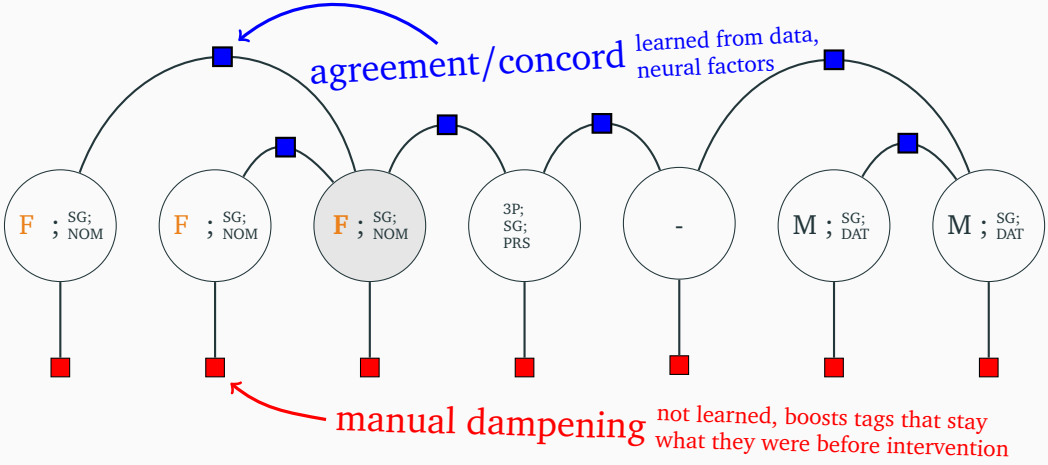
Build a MRF over morphological tags along the dependency parse!



# Syntax to the rescue: use dependency parses

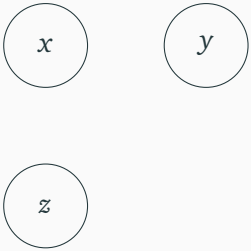
Only words “connected” in the dependency parse should change!

Build a MRF over morphological tags along the dependency parse!



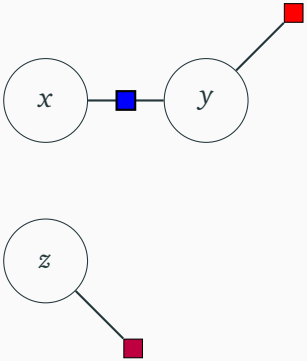
# **Recap:** what is a Markov Random Field (Koller and Friedman, 2009)?

Model $p(x, y, z)$ by decomposing into **factors** (■)!

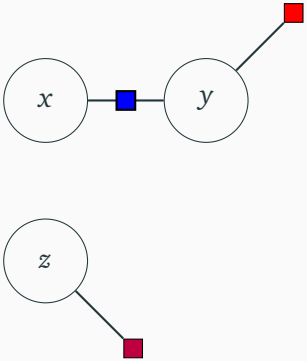


# **Recap:** what is a Markov Random Field (Koller and Friedman, 2009)?

Model $p(x, y, z)$ by decomposing into **factors** ( )!



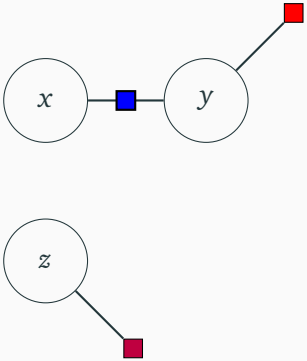
# **Recap:** what is a Markov Random Field (Koller and Friedman, 2009)?




Model $p(x, y, z)$ by decomposing into **factors** (■)!
Every factor gives a score to certain assignments:

- ■$(x = 2, y = 1) = 0.42$
- ■$(y = 1) = 1.3$
- ■$(z = 1) = -1$

# **Recap:** what is a Markov Random Field (Koller and Friedman, 2009)?



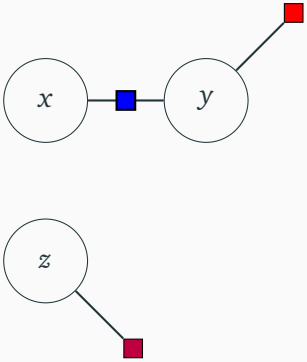

Model $p(x, y, z)$ by decomposing into **factors** (■)!
Every factor gives a score to certain assignments:

■$(x = 2, y = 1) = 0.42$
■$(y = 1) = 1.3$
■$(z = 1) = -1$

Add up all factors to obtain global score:

$\text{score}(x = 2, y = 1, z = 4) =$
■$(x = 2, y = 1) +$ ■$(y = 1) +$ ■$(z = 4)$

# **Recap:** what is a Markov Random Field (Koller and Friedman, 2009)?




Model $p(x, y, z)$ by decomposing into **factors** (■)!
Every factor gives a score to certain assignments:

■$(x = 2, y = 1) = 0.42$
■$(y = 1) = 1.3$
■$(z = 1) = -1$

Add up all factors to obtain global score:

$$\text{score}(x = 2, y = 1, z = 4) = \blacksquare(x = 2, y = 1) + \blacksquare(y = 1) + \blacksquare(z = 4)$$

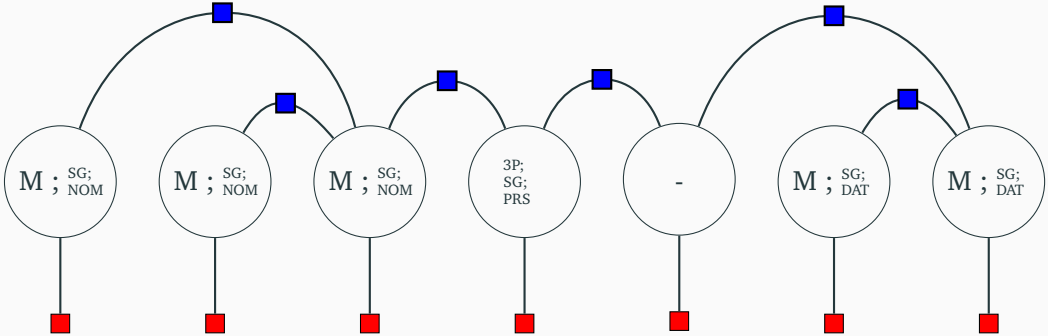
Get $p$ by global normalization (easy in trees):

$$p(x = 2, y = 1, z = 4) \propto \exp \text{score}(x = 2, y = 1, z = 4)$$

# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

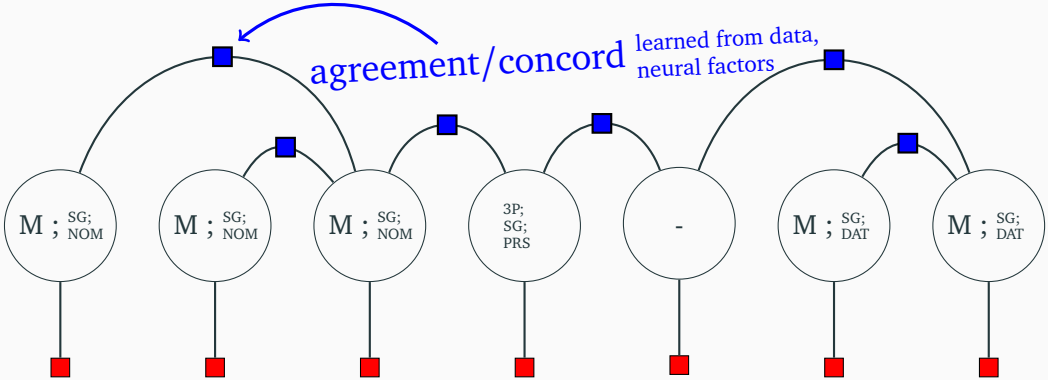
Build a MRF over morphological tags along the dependency parse!



# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

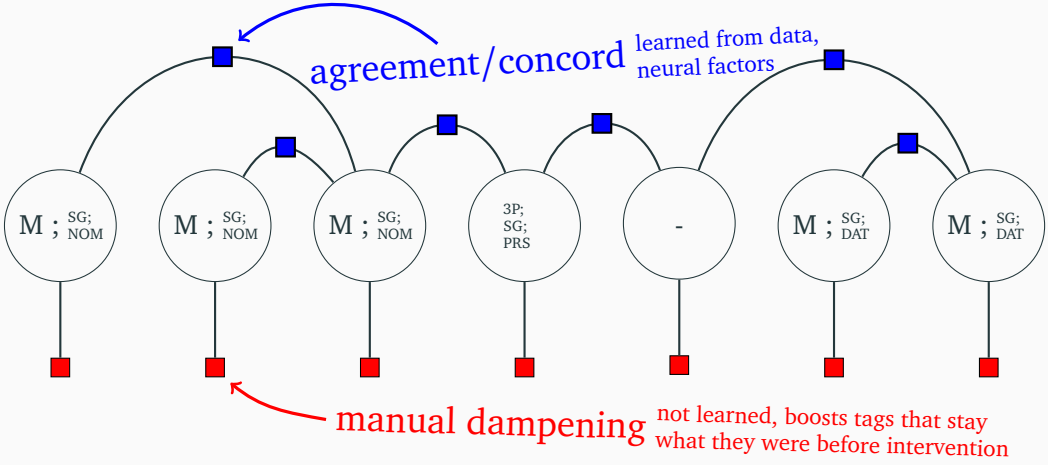
Build a MRF over morphological tags along the dependency parse!



# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

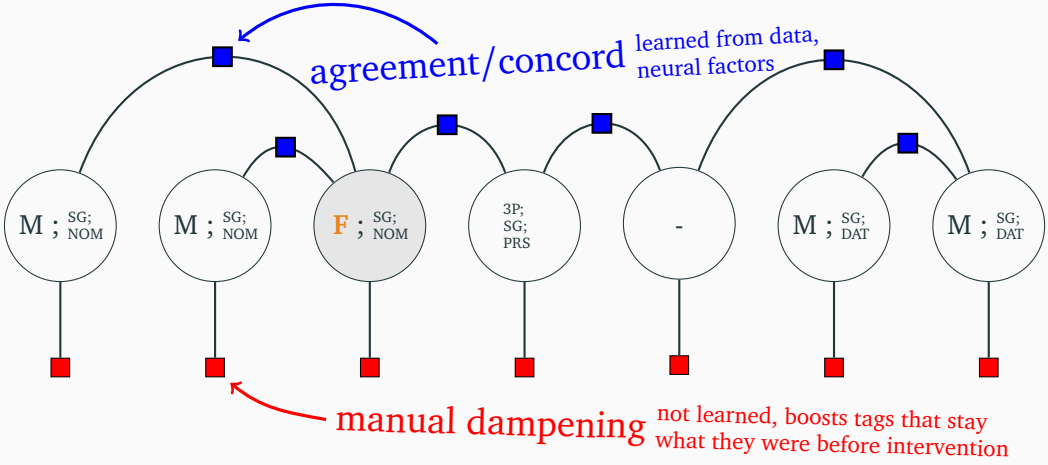
Build a MRF over morphological tags along the dependency parse!



# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

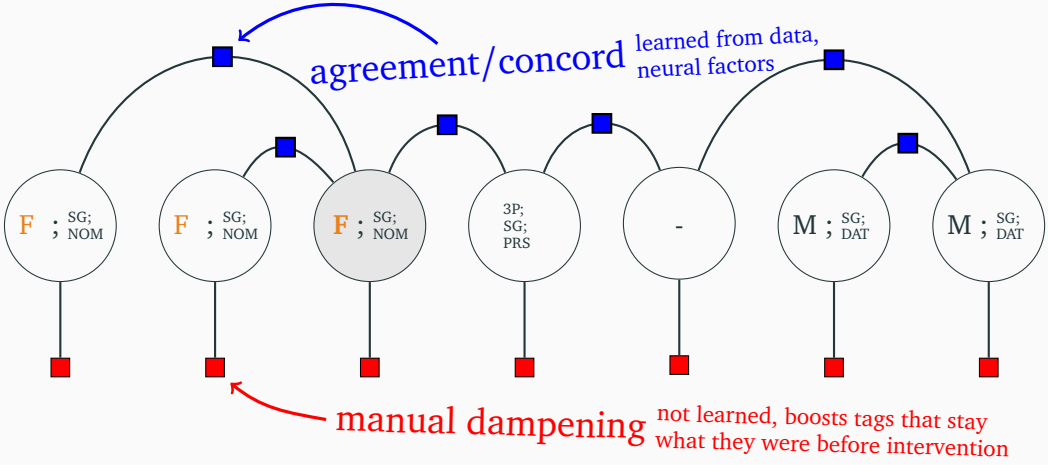
Build a MRF over morphological tags along the dependency parse!



# Syntax to the rescue: use dependency parses

Only words “connected” in the dependency parse should change!

Build a MRF over morphological tags along the dependency parse!



# Reinflect tokens to obtain the CDA sentence

Get the new sentence by performing **morphological reinflection** where tags changes:

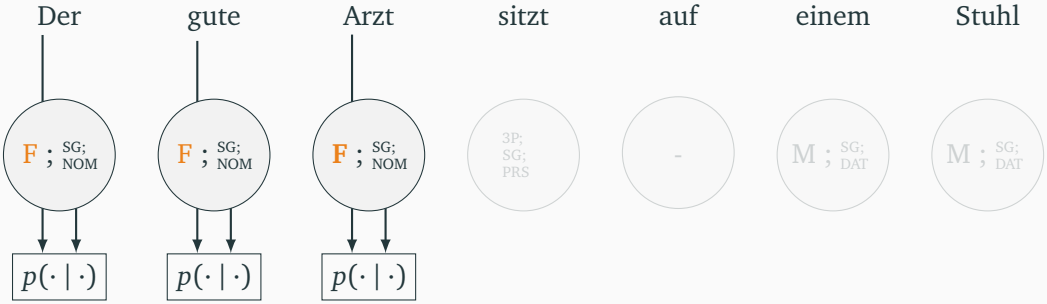
(this is a reasonably well-working procedure, established in three shared tasks at SIGMORPHON and CoNLL)

| Der | gute | Arzt | sitzt | auf | einem | Stuhl |
|---|---|---|---|---|---|---|
| F ; SG; NOM | F ; SG; NOM | F ; SG; NOM | 3P; SG; PRS | - | M ; SG; DAT | M ; SG; DAT |

# Reinflect tokens to obtain the CDA sentence

Get the new sentence by performing **morphological reinflection** where tags changes:

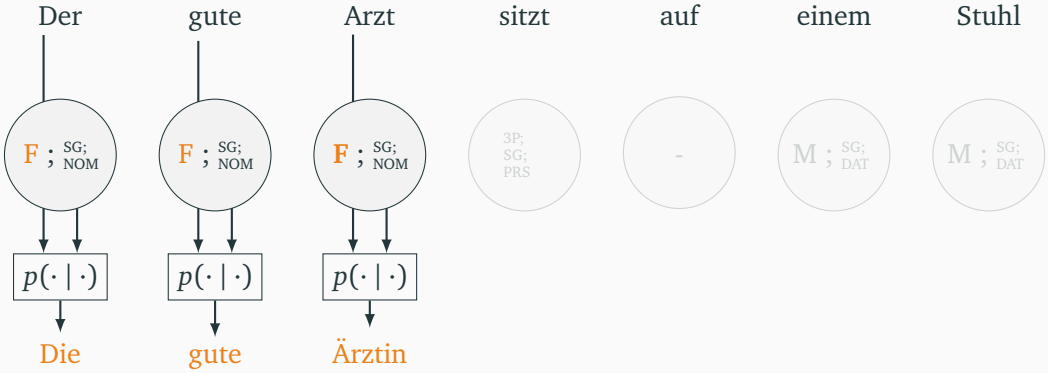
(this is a reasonably well-working procedure, established in three shared tasks at SIGMORPHON and CoNLL)



# Reinflect tokens to obtain the CDA sentence

Get the new sentence by performing **morphological reinflection** where tags changes:

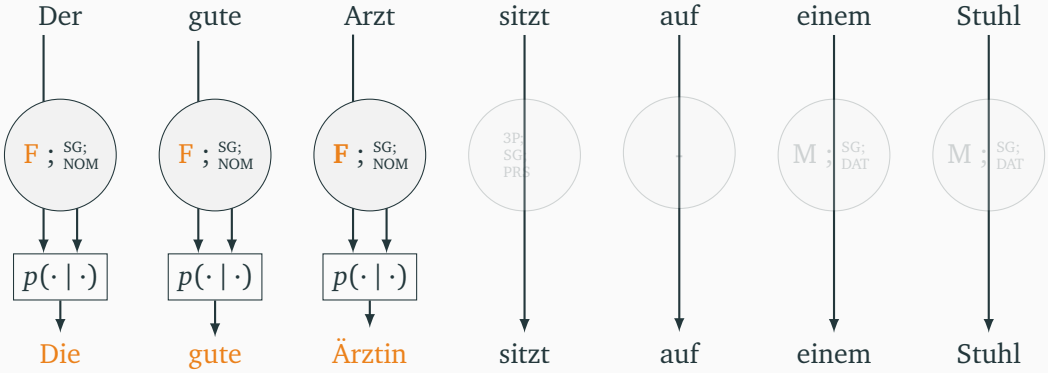
(this is a reasonably well-working procedure, established in three shared tasks at SIGMORPHON and CoNLL)



# Reinflect tokens to obtain the CDA sentence

Get the new sentence by performing **morphological reinflection** where tags changes:

(this is a reasonably well-working procedure, established in three shared tasks at SIGMORPHON and CoNLL)



# Intrinsic evaluation: how good are we at gender-swapping (Hebrew, Spanish)?

We manually annotated over 100 sentences for each language and checked performance:

| | Tag | | | | Form |
|---|---|---|---|---|---|
| | **P** | **R** | ***F*1** | **Acc** | **Acc** |

# Intrinsic evaluation: how good are we at gender-swapping (Hebrew, Spanish)?

We manually annotated over 100 sentences for each language and checked performance:

| | Tag | | | | Form |
|---|---|---|---|---|---|
| | **P** | **R** | ***F*1** | **Acc** | **Acc** |
| Hebrew: hardcoded factors | **89.04** | 40.12 | 55.32 | 86.88 | 83.63 |

# Intrinsic evaluation: how good are we at gender-swapping (Hebrew, Spanish)?

We manually annotated over 100 sentences for each language and checked performance:

| | Tag | | | | Form |
|---|---|---|---|---|---|
| | **P** | **R** | ***F*1** | **Acc** | **Acc** |
| Hebrew: hardcoded factors | **89.04** | 40.12 | 55.32 | 86.88 | 83.63 |
| Hebrew: linear factors | 87.07 | 62.35 | 72.66 | 90.5 | **86.75** |

# Intrinsic evaluation: how good are we at gender-swapping (Hebrew, Spanish)?

We manually annotated over 100 sentences for each language and checked performance:

| | Tag | | | | Form |
|---|---|---|---|---|---|
| | **P** | **R** | ***F*1** | **Acc** | **Acc** |
| Hebrew: hardcoded factors | **89.04** | 40.12 | 55.32 | 86.88 | 83.63 |
| Hebrew: linear factors | 87.07 | 62.35 | 72.66 | 90.5 | **86.75** |
| Hebrew: neural factors | 87.18 | **62.96** | **73.12** | **90.62** | 86.25 |

# Intrinsic evaluation: how good are we at gender-swapping (Hebrew, Spanish)?

We manually annotated over 100 sentences for each language and checked performance:

| | Tag | | | | Form |
|---|---|---|---|---|---|
| | **P** | **R** | ***F*1** | **Acc** | **Acc** |
| Hebrew: hardcoded factors | **89.04** | 40.12 | 55.32 | 86.88 | 83.63 |
| Hebrew: linear factors | 87.07 | 62.35 | 72.66 | 90.5 | **86.75** |
| Hebrew: neural factors | 87.18 | **62.96** | **73.12** | **90.62** | 86.25 |
| Spanish: hardcoded factors | **96.97** | 51.45 | 67.23 | 90.21 | 86.32 |
| Spanish: linear factors | 92.74 | **73.95** | **82.29** | 93.79 | 89.52 |
| Spanish: neural factors | 95.34 | 72.35 | 82.27 | **93.91** | **89.65** |

# Extrinsic evaluation: train language models on CDA-balanced data, then evaluate:

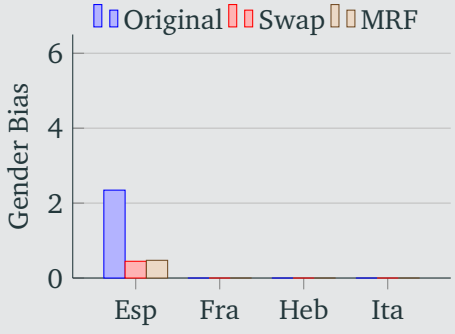
**Bias**

$$\log \frac{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Der gute Arzt } \mathbf{x})}{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Die gute Ärztin } \mathbf{x})} \quad \frac{\text{m}}{\text{f}}$$

# Extrinsic evaluation: train language models on CDA-balanced data, then evaluate:

**Bias**

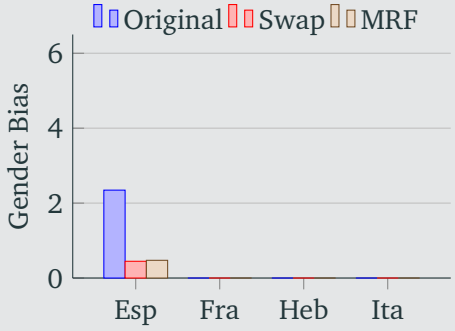
$$\log \frac{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Der gute Arzt } \mathbf{x})}{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Die gute Ärztin } \mathbf{x})} \quad \frac{\text{m}}{\text{f}}$$



# Extrinsic evaluation: train language models on CDA-balanced data, then evaluate:

## Bias

$$\log \frac{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Der gute Arzt } \mathbf{x})}{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Die gute Ärztin } \mathbf{x})} \quad \frac{\text{m}}{\text{f}}$$



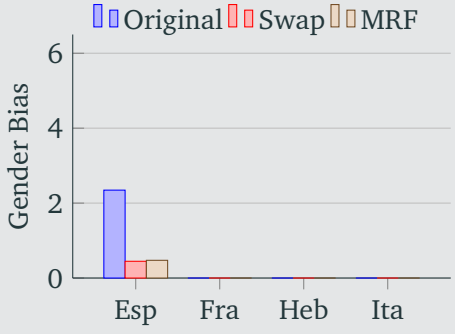

## Grammaticality

$$\log \frac{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Die gute Ärztin } \mathbf{x})}{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Der gute Ärztin } \mathbf{x})} \quad \frac{\text{ok}}{\text{bad}}$$

# Extrinsic evaluation: train language models on CDA-balanced data, then evaluate:

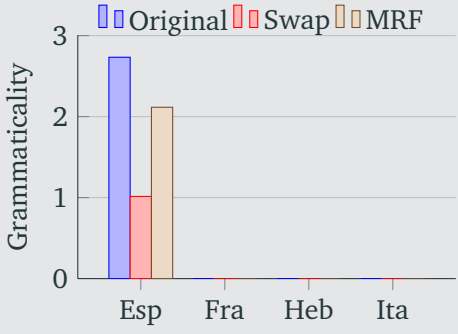
## Bias

$$\log \frac{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Der gute Arzt } \mathbf{x})}{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Die gute Ärztin } \mathbf{x})} \quad \frac{\text{m}}{\text{f}}$$




## Grammaticality

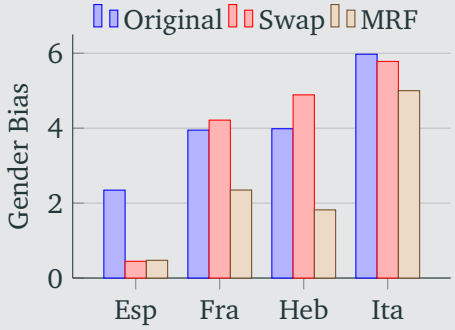
$$\log \frac{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Die gute Ärztin } \mathbf{x})}{\sum_{\mathbf{x} \in \Sigma^*} p(\text{Der gute Ärztin } \mathbf{x})} \quad \frac{\text{ok}}{\text{bad}}$$



# Extrinsic evaluation: train language models on CDA-balanced data, then evaluate:

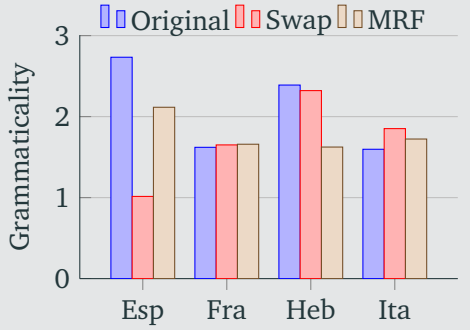
## Bias

$$\log \frac{\sum_{\mathbf{x}\in\Sigma^*} p(\text{Der gute Arzt } \mathbf{x})}{\sum_{\mathbf{x}\in\Sigma^*} p(\text{Die gute Ärztin } \mathbf{x})} \quad \frac{\text{m}}{\text{f}}$$




## Grammaticality

$$\log \frac{\sum_{\mathbf{x}\in\Sigma^*} p(\text{Die gute Ärztin } \mathbf{x})}{\sum_{\mathbf{x}\in\Sigma^*} p(\text{Der gute Ärztin } \mathbf{x})} \quad \frac{\text{ok}}{\text{bad}}$$



# Conclusion

# Conclusion

1. As so often, things that are easy in English...
   ...become surprisingly hard in other languages.

# Conclusion

1. As so often, things that are easy in English...
   ...become surprisingly hard in other languages.
2. Old-school probabilistic models often work well enough™

# Conclusion

1. As so often, things that are easy in English...
   ...become surprisingly hard in other languages.
2. Old-school probabilistic models often work well enough$^{TM}$
3. And, always, careful with your training data, Eugene!

# Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology

**ACL 2019**

---

Ran Zmigrod, **Sebastian J. Mielke**, Hanna Wallach, Ryan Cotterell

University of Cambridge // Johns Hopkins University // Microsoft Research
rz279@cam.ac.uk sjmielke@jhu.edu
wallach@microsoft.com rdc42@cam.ac.uk

Twitter: @RanZmigrod – paper and thread pinned! // @sjmielke