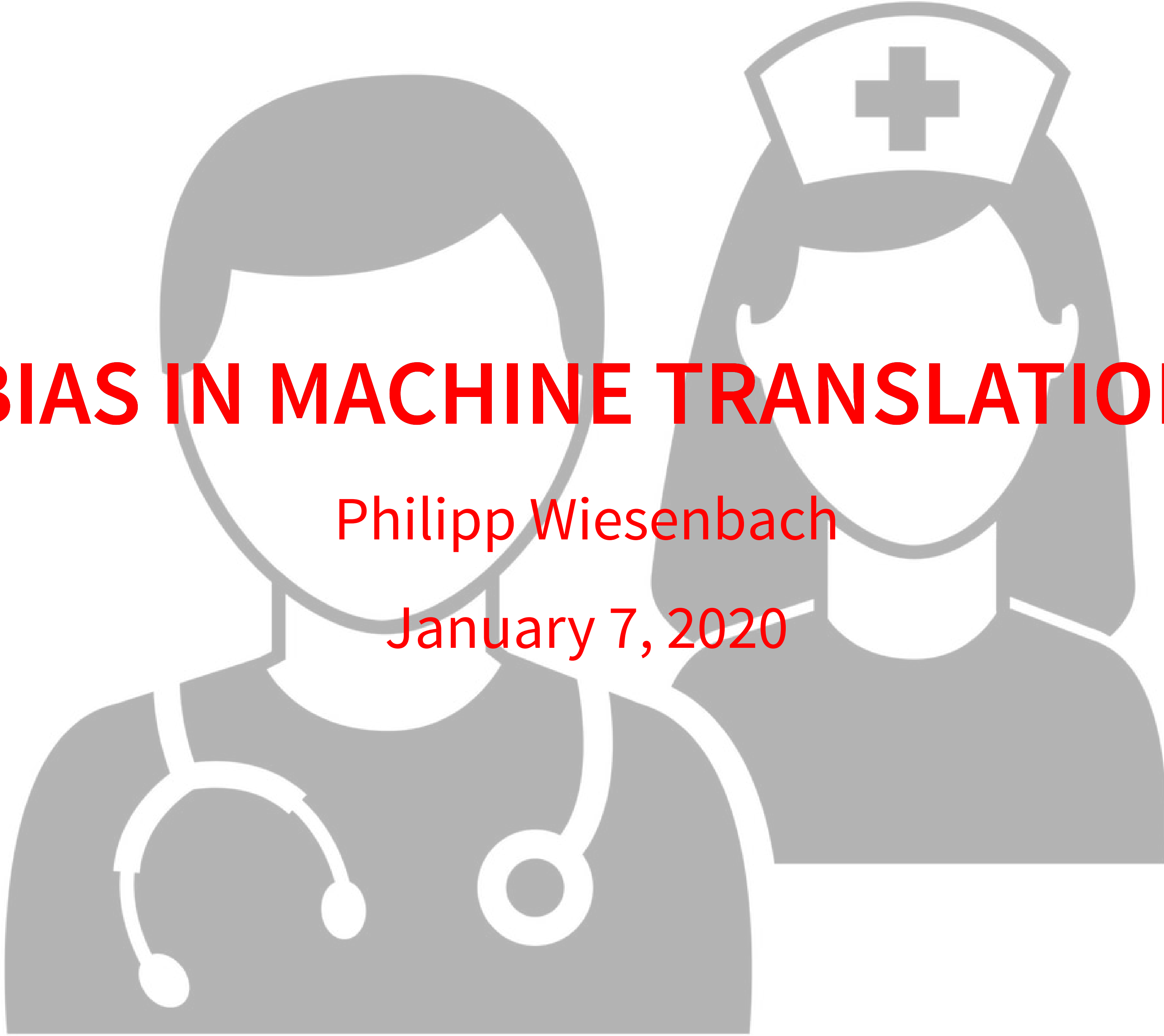


# BIAS IN MACHINE TRANSLATION

Philipp Wiesenbach

January 7, 2020

# OVERVIEW

1. Recap of bias forms
2. Examples of bias in MT
3. Learning gender-neutral word embeddings
4. Examining impact of embeddings in MT:
    - GloVe
    - Hard-GloVe
    - GN-GloVe
5. Summary
6. Critique

# RECAP

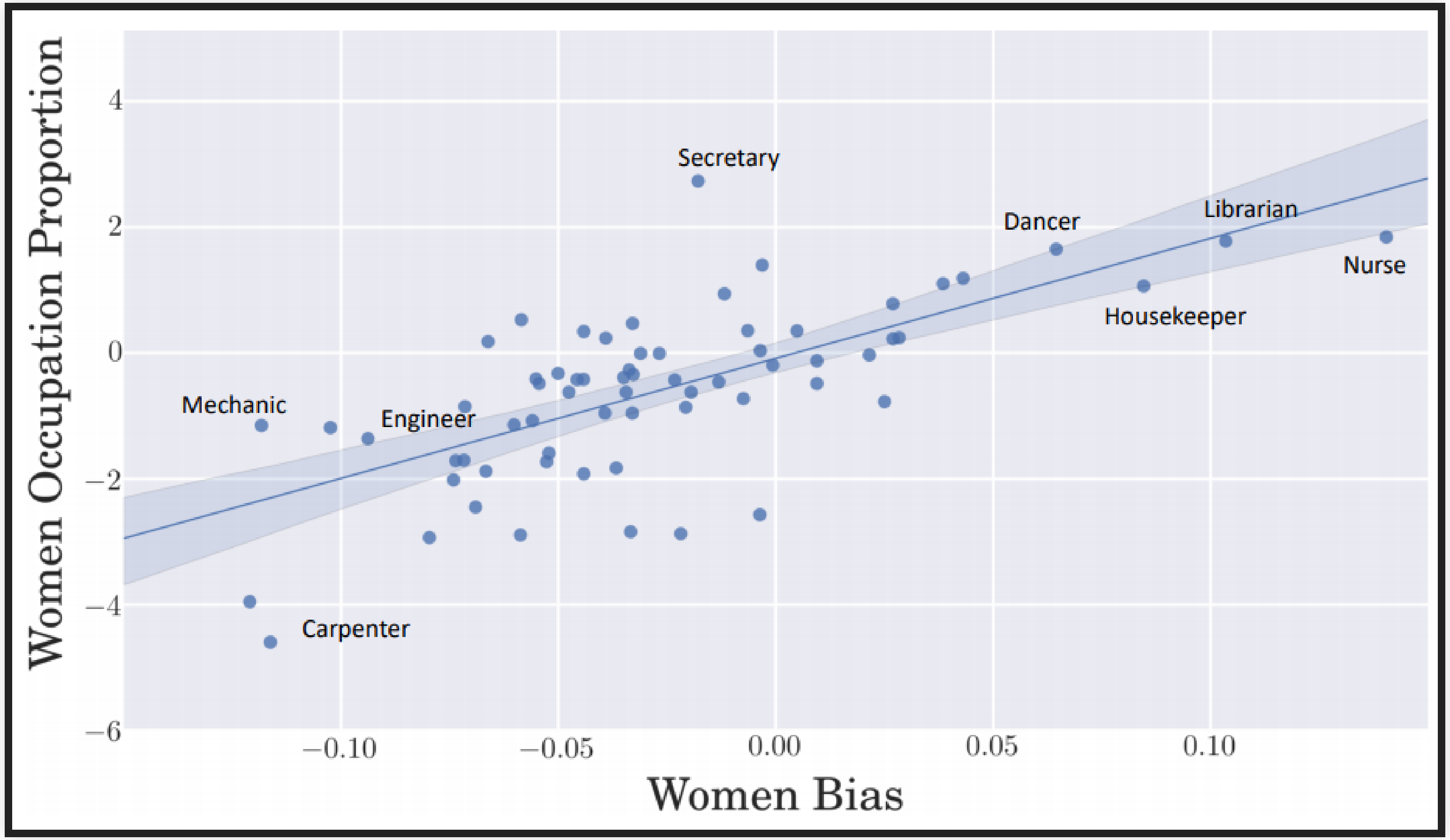
# HOW CAN BIAS BE EXHIBITED AGAIN?

- Disparate Treatment: *Choices made directly on a protected attribute*
  - Does favouring minority groups create new bias?
- Disparate Impact: *Choices are fair for all classes, but outcome still favours a certain class*
  - What about $\frac{p(yes|Minority)}{p(yes|Majority)} \leq 0.8$?

# HOW CAN WE EQUALIZE?

- Demographical/Statistical Parity:
  - $p(\tilde{y} = 1 | A = 0) = p(\tilde{y} = 1 | A = 1)$
- Equal Opportunity:
  - $p(\tilde{y} = 1 | A = 0, y = 1) = p(\tilde{y} = 1 | A = 1, y = 1)$
- Fairness through unawareness:
  - An algorithm is fair (is it?) as long as any protected attributes A are not explicitly used in decision-making process.

# REMEMBER?

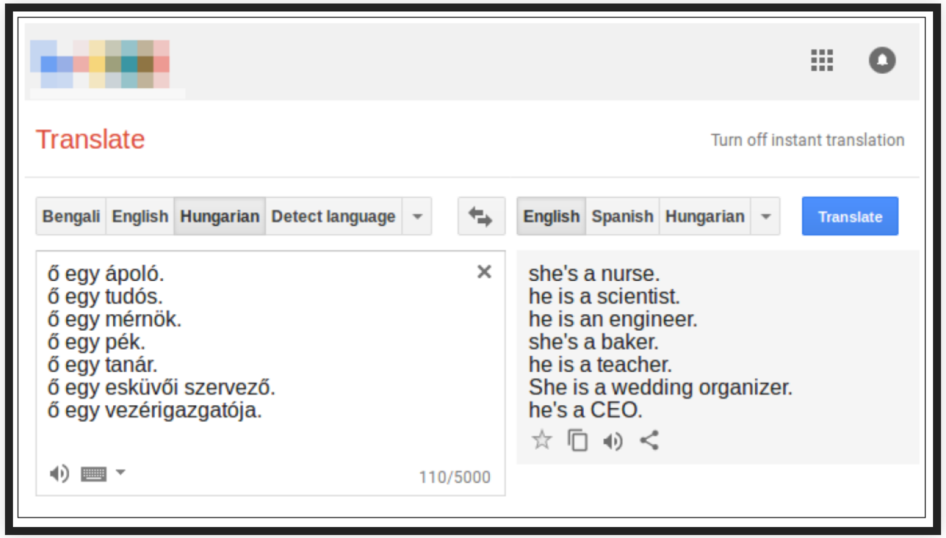



*Woman occupation proportion vs embedding bias in Google News vectors (Garg et al. 2018)*

# ASSESSING GENDER BIAS IN MACHINE TRANSLATION (PRATES ET AL. 2019)

# ASSESSING GENDER BIAS IN MT

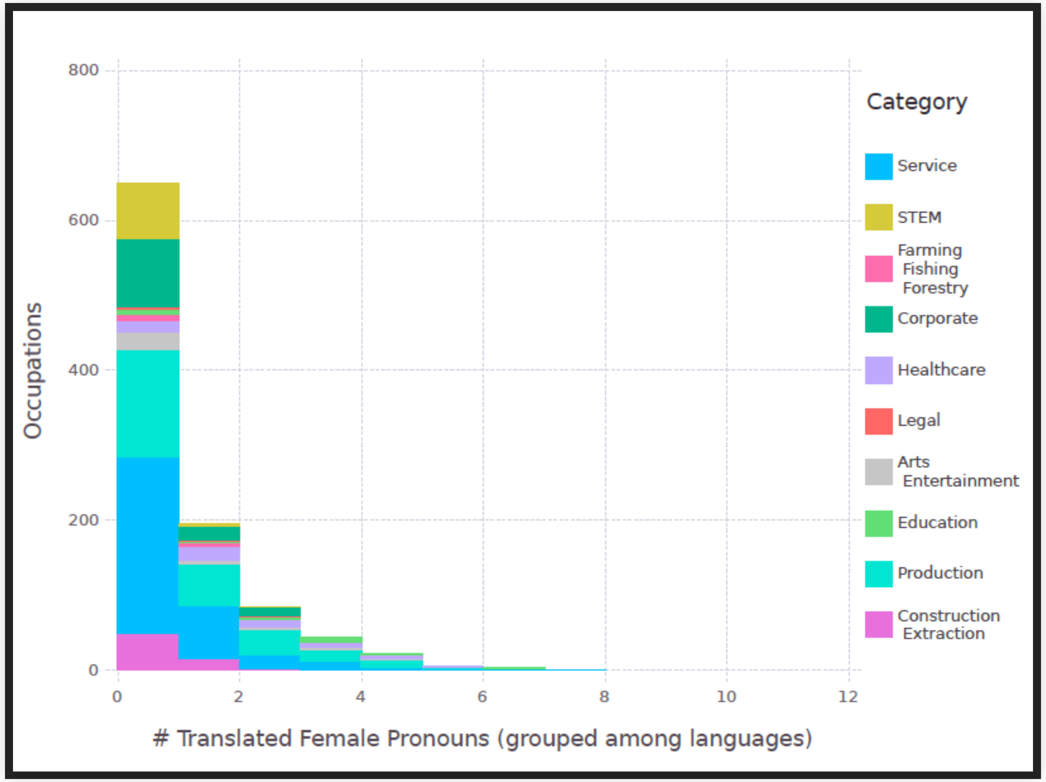
- Exploitation of Google Translates shows that translating from gender-neutral languages to English shows strong bias towards male pronouns.



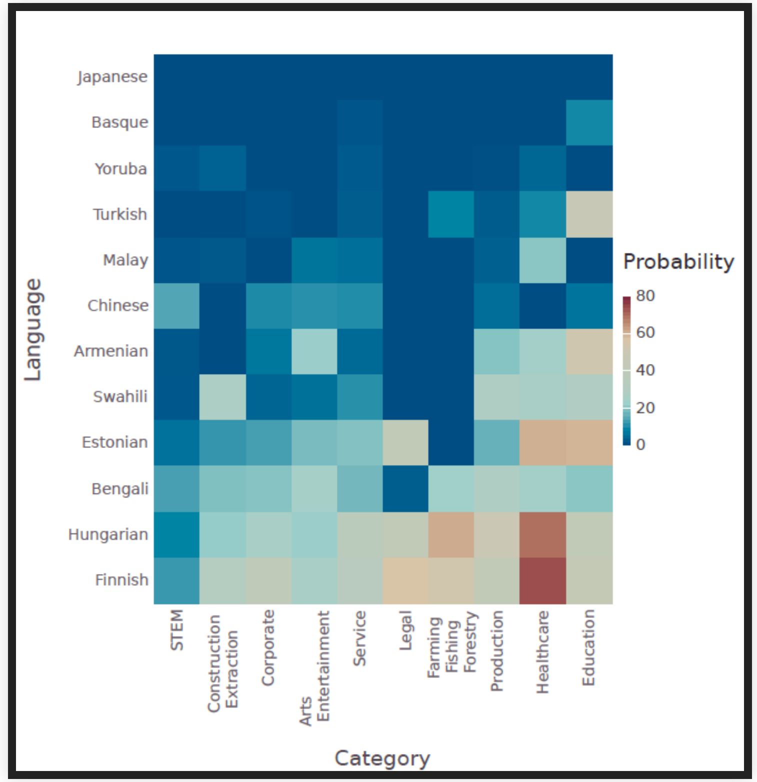

*Translating from Hungarian to English (Prates et al. 2019)*

# EXPERIMENT SERIES

- Assessing the distribution of translated gender pronouns per occupation across 12 languages ⟶ English




*Female translation count per occupation*




*Female translation probabilities per language*

# GOOGLE'S REACTION

*Using Google Translate Turkish $\longrightarrow$ English (16.12.2019)*

*Using Google Translate Turkish $\longrightarrow$ German (16.12.2019)*

# LEARNING GENDER-NEUTRAL EMBEDDINGS (ZHAO ET AL. 2018)

# MAIN IDEA:

- Instead of post-processing - learn gender neutral embeddings with GloVe
- Features of *GN-GloVe*:
  - End-to-end
  - Interpretability
  - Preserves word proximity

# MINI-WALKTHROUGH: GLOVE

- Starting point: Predict ratios of co-occurrences between a source word $\tilde{w}_k$ and two context words $w_i$ and $w_j$:
  - $F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
- Calculating ratios effictively redues noise

| Probability and Ratio | $k = solid$ | $k = gas$ | $k = water$ | $k = fashion$ |
|---|---|---|---|---|
| $P(k\|ice)$ | $1.9 \times 10^{-4}$ | $6.6 \times 10^{-5}$ | $3.0 \times 10^{-3}$ | $1.7 \times 10^{-5}$ |
| $P(k\|steam)$ | $2.2 \times 10^{-5}$ | $7.8 \times 10^{-4}$ | $2.2 \times 10^{-3}$ | $1.8 \times 10^{-5}$ |
| $P(k\|ice)/P(k\|steam)$ | 8.9 | $8.5 \times 10^{-2}$ | 1.36 | 0.96 |

*Co-occurrence probabilities for target words ice and steam with selected context words from a 6 billion token corpus (Pennington et al. 2014)*

- Transformations

- Transformations
  - $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$

- Transformations
  - $F(w_i - w_j, \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
  - $F((w_i - w_j)^\intercal \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$

- Transformations
  - $F(w_i - w_j, \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
  - $F((w_i - w_j)^\intercal \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
- Requirements:

- Transformations
  - $F(w_i - w_j, \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
  - $F((w_i - w_j)^\intercal \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
- Requirements:
  - Homomorphism between ($\mathcal{R}$, +) and ($\mathcal{R}_{>0}, \times$):

- Transformations
  - $F(w_i - w_j, \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
  - $F((w_i - w_j)^\intercal \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
- Requirements:
  - Homomorphism between ($\mathcal{R}$, +) and ($\mathcal{R}_{>0}, \times$):
  - $\rightarrow F(w_i^\intercal \tilde{w_k} - w_j^\intercal \tilde{w_k}) = \frac{F(w_i^\intercal \tilde{w_k})}{F(w_j^\intercal \tilde{w_k})}$

- Transformations
  - $F(w_i - w_j, \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
  - $F((w_i - w_j)^\intercal \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
- Requirements:
  - Homomorphism between ($\mathcal{R}$, +) and ($\mathcal{R}_{>0}, \times$):
  - $\rightarrow F(w_i^\intercal \tilde{w_k} - w_j^\intercal \tilde{w_k}) = \frac{F(w_i^\intercal \tilde{w_k})}{F(w_j^\intercal \tilde{w_k})}$
  - $\rightarrow F(w_i^\intercal \tilde{w_k}) = P_{ik} = \frac{X_{ik}}{X_i}$

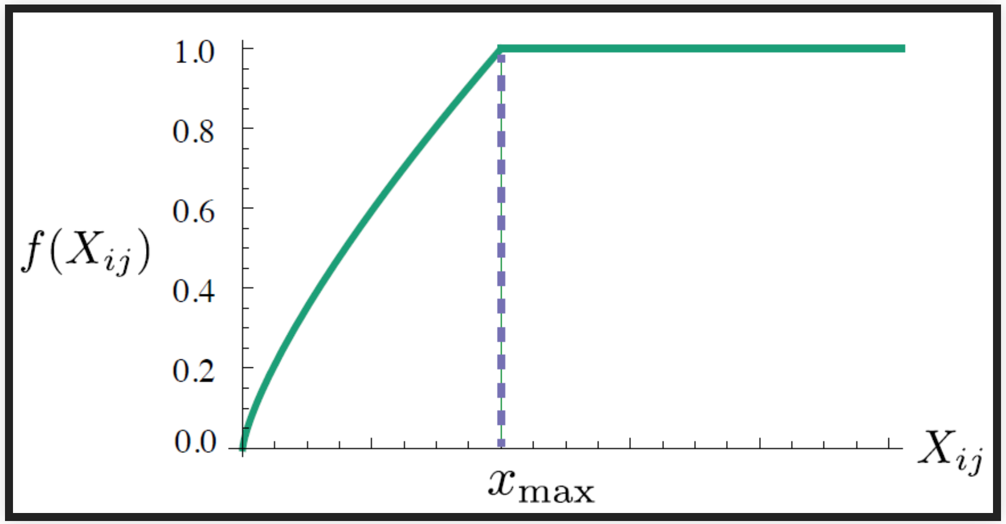
- Transformations
  - $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
  - $F((w_i - w_j)^\intercal \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
- Requirements:
  - Homomorphism between ($\mathcal{R}$, +) and ($\mathcal{R}_{>0}, \times$):
  - $\rightarrow F(w_i^\intercal \tilde{w}_k - w_j^\intercal \tilde{w}_k) = \frac{F(w_i^\intercal \tilde{w}_k)}{F(w_j^\intercal \tilde{w}_k)}$
  - $\rightarrow F(w_i^\intercal \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$
- Solution with $F(x) = e^x$

- Transformations
  - $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
  - $F((w_i - w_j)^\top \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
- Requirements:
  - Homomorphism between ($\mathcal{R}$, +) and ($\mathcal{R}_{>0}, \times$):
  - $\rightarrow F(w_i^\top \tilde{w}_k - w_j^\top \tilde{w}_k) = \frac{F(w_i^\top \tilde{w_k})}{F(w_j^\top \tilde{w}_k)}$
  - $\rightarrow F(w_i^\top \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$
- Solution with $F(x) = e^x$
  - $w_i^\top \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$

- Transformations
  - $F(w_i - w_j, \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
  - $F((w_i - w_j)^\intercal \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$
- Requirements:
  - Homomorphism between ($\mathcal{R}$, +) and ($\mathcal{R}_{>0}, \times$):
  - $\rightarrow F(w_i^\intercal \tilde{w}_k - w_j^\intercal \tilde{w}_k) = \frac{F(w_i^\intercal \tilde{w_k})}{F(w_j^\intercal \tilde{w}_k)}$
  - $\rightarrow F(w_i^\intercal \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$
- Solution with $F(x) = e^x$
  - $w_i^\intercal \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$
  - $\rightarrow w_i^\intercal \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$

# WEIGHTED LEAST SQUARES LOSS

$$J = \sum_{i,j}^{V} f(X_{ij})(w_i^\intercal \tilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik}))^2$$




*Function $f$*

# GENDER NEUTRAL WORD EMBEDDINGS:

- Reserve $k$ dimensions of Featurespace $\mathbb{R}^d$ for gender information:
  - $w^{(g)} \in \mathbb{R}^k$: gender component
  - $w^{(a)} \in \mathbb{R}^{d-k}$: neutral component
  - Embedding vector becomes $[w^{(a)}; w^{(g)}]$
  - Calculate gender direction $v_g \in \mathbb{R}^{d-k}$
  - Define vocabulary subsets:
  - $\Omega_m$=(male), $\Omega_f$=(female), $\Omega_n$=(neutral)

# LOSS CALCULATION:

$$J = J_G + \lambda_d J_D + \lambda_e J_e$$

# LOSS CALCULATION:

$$J = J_G + \lambda_d J_D + \lambda_e J_e$$

- $J_G = \sum_{i,j}^{V} f(X_{ij})(w_i^\intercal \tilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik}))^2$

# LOSS CALCULATION:

$$J = J_G + \lambda_d J_D + \lambda_e J_e$$

- $J_G = \sum_{i,j}^{V} f(X_{ij})(w_i^\intercal \tilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik}))^2$
  - $J_D^{L1} = -\left\|\sum_{w \in \Omega_M} w^{(g)} - \sum_{w \in \Omega_F} w^{w(g)}\right\|_1$

# LOSS CALCULATION:

$$J = J_G + \lambda_d J_D + \lambda_e J_e$$

- $J_G = \sum_{i,j}^{V} f(X_{ij})(w_i^\top \tilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik}))^2$
  - $J_D^{L1} = -\left\| \sum_{w \in \Omega_M} w^{(g)} - \sum_{w \in \Omega_F} w^{w(g)} \right\|_1$
  - $J_D^{L2} = \sum_{w \in \Omega_M} \left\| \beta_1 e - w^{(g)} \right\|_2^2 + \sum_{w \in \Omega_F} \left\| \beta_2 e - w^{(g)} \right\|_2^2$

# LOSS CALCULATION:

$$J = J_G + \lambda_d J_D + \lambda_e J_e$$

- $J_G = \sum_{i,j}^{V} f(X_{ij})(w_i^\top \tilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik}))^2$
  - $J_D^{L1} = -\left\|\sum_{w\in\Omega_M} w^{(g)} - \sum_{w\in\Omega_F} w^{w(g)}\right\|_1$
  - $J_D^{L2} = \sum_{w\in\Omega_M} \left\|\beta_1 e - w^{(g)}\right\|_2^2 + \sum_{w\in\Omega_F} \left\|\beta_2 e - w^{(g)}\right\|_2^2$
- $J_E = \sum_{w\in\Omega_N} (v_g^\top w^{(a)})^2$

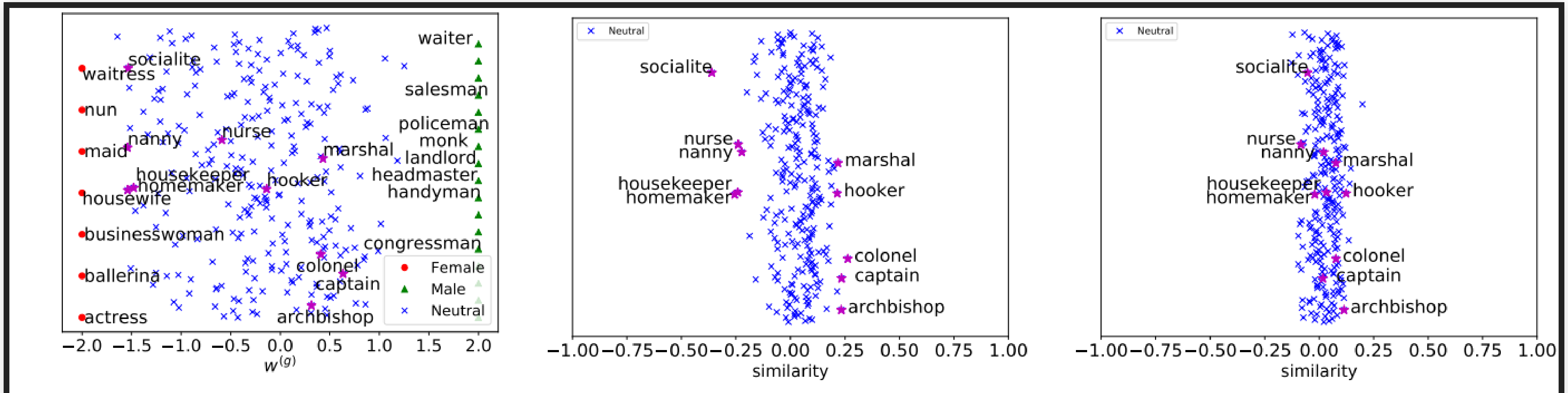
# LOSS CALCULATION:

$$J = J_G + \lambda_d J_D + \lambda_e J_e$$

- $J_G = \sum_{i,j}^{V} f(X_{ij})(w_i^\top \tilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik}))^2$
  - $J_D^{L1} = -\left\| \sum_{w \in \Omega_M} w^{(g)} - \sum_{w \in \Omega_F} w^{w(g)} \right\|_1$
  - $J_D^{L2} = \sum_{w \in \Omega_M} \left\| \beta_1 e - w^{(g)} \right\|_2^2 + \sum_{w \in \Omega_F} \left\| \beta_2 e - w^{(g)} \right\|_2^2$
- $J_E = \sum_{w \in \Omega_N} (v_g^\top w^{(a)})^2$

# ANALYSIS: SIMILARITY TO GENDER-NEUTRAL PROFESSIONS

- (a): Plotting $w^{(g)}$ on random axis
- (b), (c): Plotting $\dfrac{w^{(a)\top} v_g}{\left\|w^{(a)}\right\| \left\|v_g\right\|}$ for GN-GloVe & GloVe




(a) $w^{(g)}$ dimension for all the professions

(b) Gender-neutral profession words projected to gender direction in GloVe

(c) Gender-neutral profession words projected to gender direction in GN-GloVe

# ANALYSIS: RELATIONAL TASK

- Create relational pairs:
  - definiational *actor* - *actress*
  - steoreotypical: *nurse* - *doctor*
  - gender-unrelated: *cup* - *lid*
- Test against *she* - *he* via cosine similarity

| Dataset | Embeddings | Definition | Stereotype | None |
|---|---|---|---|---|
| SemBias | GloVe | 80.2 | 10.9 | 8.9 |
| | Hard-Glove | 84.1 | 6.4 | 9.5 |
| | GN-GloVe | 97.7 | 1.4 | 0.9 |
| SemBias (subset) | GloVe | 57.5 | 20 | 22.5 |
| | Hard-Glove | 25 | 27.5 | 47.5 |
| | GN-GloVe | 75 | 15 | 10 |

# ANALYSIS: SIMILARITY TASK

- Analogy (Accuracy): Google analogy dataset
  - *infrequent infrequently immediate immediately*
  - *Athens Greece Bern Switzerland*
- Similiarity (Ranking): WS353-ALL
  - *stock egg* 1.81
  - *fertility egg* 6.69

| Embeddings | Analogy | | Similarity | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Google | MSR | WS353-ALL | RG-65 | MTurk-287 | MTurk-771 | RW | MEN-TR-3k |
| GloVe | **70.8** | **45.8** | 62.0 | 75.3 | 64.8 | 64.9 | 37.3 | 72.2 |
| Hard-GloVe | **70.8** | **45.8** | 61.2 | 74.8 | 64.4 | 64.8 | 37.3 | 72.2 |
| GN-GloVe-L1 | 68.9 | 43.7 | **62.8** | 74.1 | 66.2 | **66.2** | **40.0** | **74.5** |
| GN-GloVe-L2 | 68.8 | 43.6 | 62.5 | **76.4** | **66.8** | 65.6 | 39.3 | 74.4 |

# ANALYSIS: COREFERENCE RESOLUTION

- Ontonotes 5.0
- Winobias
  - pro/-anti-stereotype: *The CEO raised the salary of the receptionist because **he/she** is generous.*
- Best results when only $w^{(a)}$ is used

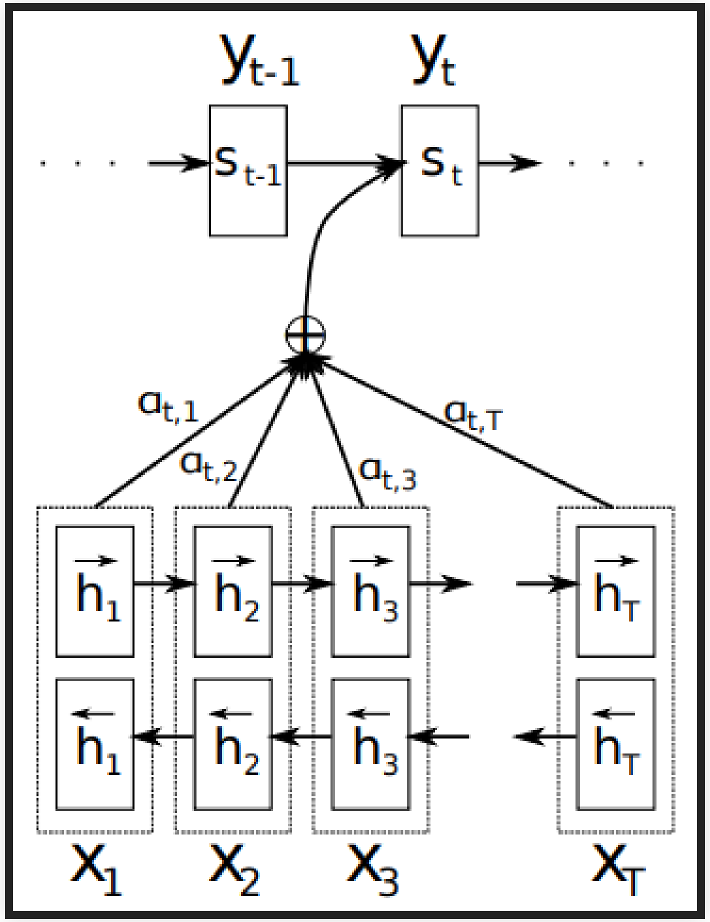
| Embeddings | OntoNotes-test | PRO | ANTI | Avg | Diff |
|---|---|---|---|---|---|
| GloVe | 66.5 | 76.2 | 46.0 | 61.1 | 30.2 |
| Hard-Glove | 66.2 | 70.6 | 54.9 | 62.8 | 15.7 |
| GN-GloVe | 66.2 | 72.4 | 51.9 | 62.2 | 20.5 |
| GN-GloVe($w_a$) | 65.9 | 70.0 | 53.9 | 62.0 | 16.1 |

# EQUALIZING GENDER BIAS IN NEURAL MACHINE TRANSLATION WITH WORD EMBEDDINGS TECHNIQUES (FONT ET AL. 2019)

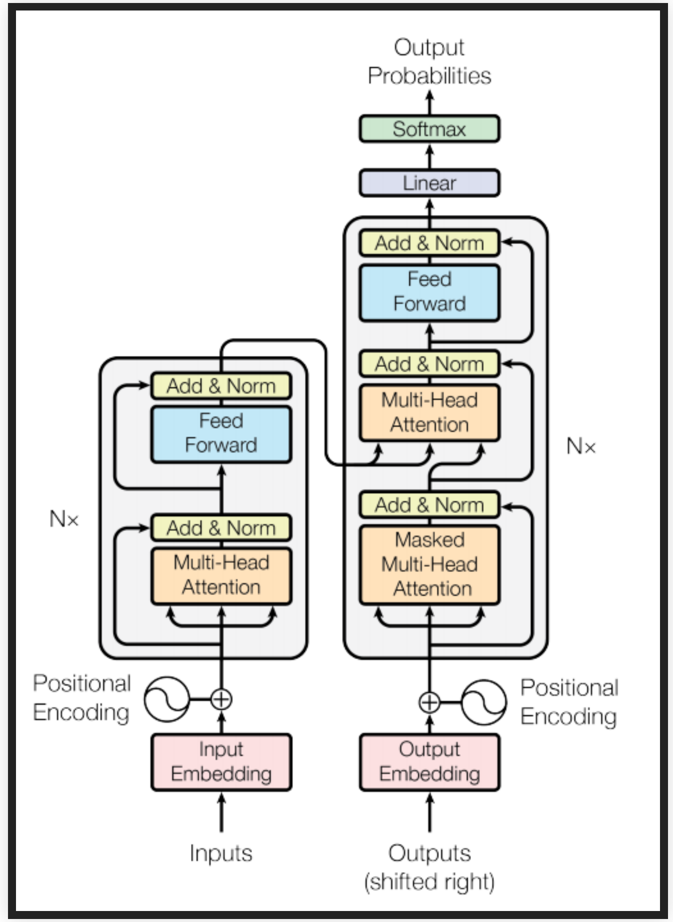
- Problem: Gender bias in Machine Translation
- Possible Embeddings:
  - GloVe
  - Hard-GloVe (post-processed)
  - GN-Glove (learned)
- Putting it all together:
  -

*Experiment series defined by (Font et al. 2019)*

# MINI-WALKTHROUGH: TRANSFORMER




*Attention (Bahdanau et al. 2016)*




*Transformer (Vaswani et al. 2017)*

# DATA

- Training data: English $\longrightarrow$ Spanish (16. Mio pairs)
- Test set: *newstest2013* (3.000 pairs)

- Bias assessment (*Occupations*) set:

> ***I've known {her, him, <proper noun>} for a long time, my <u>friend</u> works as {a, an} <occupation>.***

# RESULTS

- GN-GloVe shows a higher accuracy when predicting technical professions (criminal investigator, heating mechanic, refrigeration mechanic)

*Percentage of “friend” being translated as “amiga” or “amigo” in test sentences with female-male pronouns and proper names for the Occupations test. (Font et al. 2019)*

# CONCLUSION

- **Wrap-up**
  - (Gender-)Bias is prominint in NLP - also in MT.
  - Embeddings can be debiased during Learning (GN-GloVe)
  - All tested embedding-types solve the pronoun resolution (even MT-trained ones)
  - Hard-GloVe excels in solving proper name resolution

- **Wrap-up**
  - (Gender-)Bias is prominint in NLP - also in MT.
  - Embeddings can be debiased during Learning (GN-GloVe)
  - All tested embedding-types solve the pronoun resolution (even MT-trained ones)
  - Hard-GloVe excels in solving proper name resolution
- **Critique**
  - Analysis, why is GN-GloVe actually worse with proper names?
  - Why not applying the Losses of GN-GloVe to the machine translation task?
  - If transformer-trained Embeddings are already so strong, is debiasing still necessary?

# REFERENCES

Bahdanau, D., Cho, K., & Bengio, Y. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*. Available at: http://arxiv.org/abs/1409.0473 [Accessed December 10, 2019].

Font, J.E., & Costa-jussà, M.R. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv:1901.03116 [cs]*. Available at: http://arxiv.org/abs/1901.03116 [Accessed November 20, 2019].

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci USA* 115(16): E3635–E3644. Available at: http://arxiv.org/abs/1711.08412 [Accessed December 1, 2019].

Pennington, J., Socher, R., & Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics Available at: https://www.aclweb.org/anthology/D14-1162 [Accessed November 27, 2019].

Prates, M.O.R., Avelar, P.H.C., & Lamb, L. 2019. Assessing gender bias in machine translation – a case study with google translate. *arXiv:1809.02208 [cs]*. Available at: http://arxiv.org/abs/1809.02208 [Accessed November 26, 2019].

Vaswani, A. et al. 2017. Attention is All you Need. In I. Guyon et al. (eds) *Advances in Neural Information Processing Systems 30*, 5998–6008. Curran Associates, Inc. Available at: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf [Accessed December 8, 2018].

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. 2018. Learning gender-neutral word embeddings. *arXiv:1809.01496 [cs, stat]*. Available at: http://arxiv.org/abs/1809.01496 [Accessed November 26, 2019].