Natural Language Processing

The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing

Course

DIS25a Natural Language Processing [SS 2021]

Attendees:

Constantin Krah
Andreas Kruff
Joshua Thos
Anh Huy Matthias Tran

Agenda

1. Problem Case
2. Problem Case Examples
3. Gender Bias in NMT Systems
4. **Gender Bias in Training Data**
5. Used Dataset
6. Common Approaches to reduce biased training data
   a. Evaluation Metrics
   b. Result interpretation
7. Removing gender bias by domain adaption
   a. Result interpretation
8. Conclusion
9. Ethical consideration
10. EsuPol Use cases
11. Sources

Problem Case

- Gender bias is a common problem in linguistic data
  - *Stereotyping created by certain linguistic elements like gendered pronoun (he/him/his, she/her/hers)*
  - *Automated systems trained on the basis of biased data show an even higher bias after the training process*

- Pronouns like “they” or “them” also describe non-binary individuals of unknown gender
- In general NMT (Neural machine translation) are optimized for masculine defaults
- General problems in translations from or into richly gender-inflected languages like german

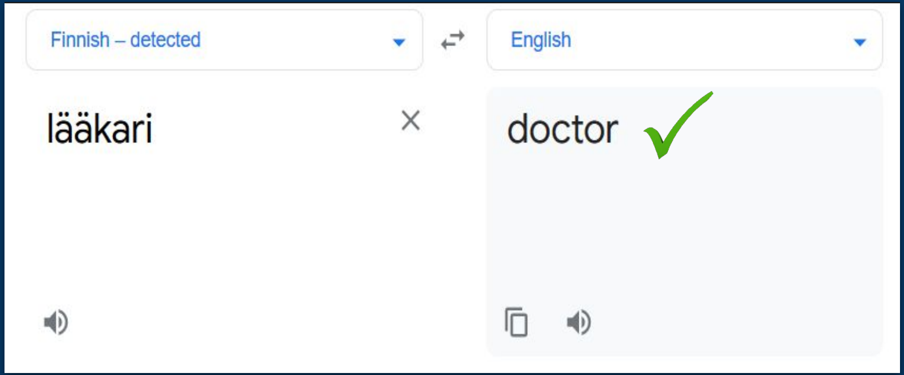
Problem Case Examples

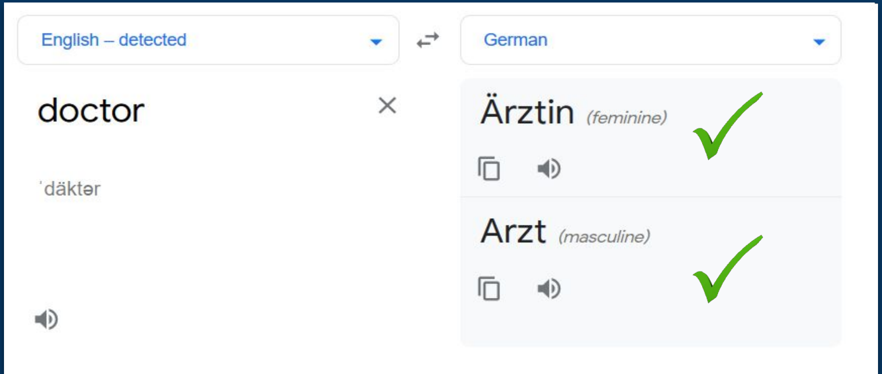
- *Every translation mechanism requires a "source language" and a target language (a counterpart). It is important that the semantic equivalence is maintained (A = B)*
  - English: Now, however, he is to go before the courts once more.
  - German: Nun ist es aber so, daß er wieder angeklagt werden soll.

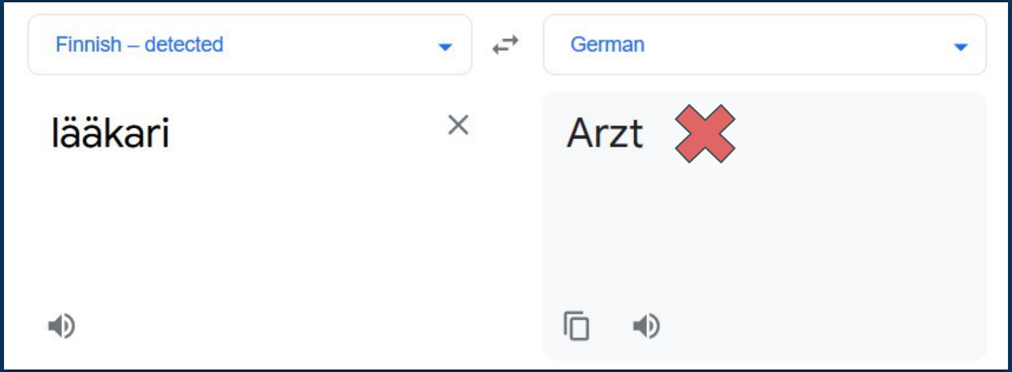
- Translation from Finnish into English had prompted a gendering of the pronoun in the English translations (gender neutral pronoun HÄN means he and she)

Finnish	English
Hän on lääkäri	She is a doctor
Hän on sairaanhoitaja	He is a nurse

Gender Bias in NMT Systems







Gender Bias in Training Data

- data displaying distinctive skewings is not biased, if the population represents this skewing in reality

Doctor 1305 (55.2 p.m.)		Architect 887 (37.52 p.m.)		Engineer 986 (41.71 p.m.)		Physicist 159 (6.73 p.m.)	
Arzt male	*Ärztin* female	*Architekt* male	*Architektin* female	*Ingenieur* male	*Ingenieurin* female	*Physiker* male	*Physikerin* female
853 (36.08 p.m.)	35 (1.84 p.m.)	669 (28.3 p.m.)	19 (0.8 p.m.)	451 (19.08 p.m.)	6 (0.25 p.m.)	131 (5.54 p.m.)	3 (0.13 p.m.)

Politician 722 (30.54 p.m.)		Nurse 106 (4.48 p.m.)		Teacher 1640 (69.37 p.m.)		Seller 469 (19.84 p.m.)	
Politiker male	*Politikerin female*	*(Kranken-) Pfleger male*	*(Kranken-) Schwester , Pflegerin female*	*Lehrer male*	*Lehrerin female*	*Verkäufer male*	*Verkäuferin female*
544 (23.01 p.m.)	27 (1.14 p.m.)	7 (0.29 p.m.)	64 (2.7 p.m.)	1,125 (48.81 p.m.)	147 (6.38 p.m.)	436 (18.92 p.m.)	19 (0.82 p.m.)

Used dataset, problems and processing

- 17.2 M sentence pairs (german/english)
- data ranges from webpages to translations of the bible.
- English part clearly contains imbalances with regard to gender specific pronouns, nouns and adjectives
  - especially refer to family relations and professions

Table 1 Gender analysis of English subset (raw counts; p.m. = per million)

Pronouns	*He*	38,990 (1648.56 p.m.)	Nouns/adjectives	*Man*	9181 (388.19 p.m.)
	She	11,068 (467.97 p.m.)		*Woman*	6500 (274.95 p.m.)

Common Approaches to reduce biased training data

- **Downsampling**

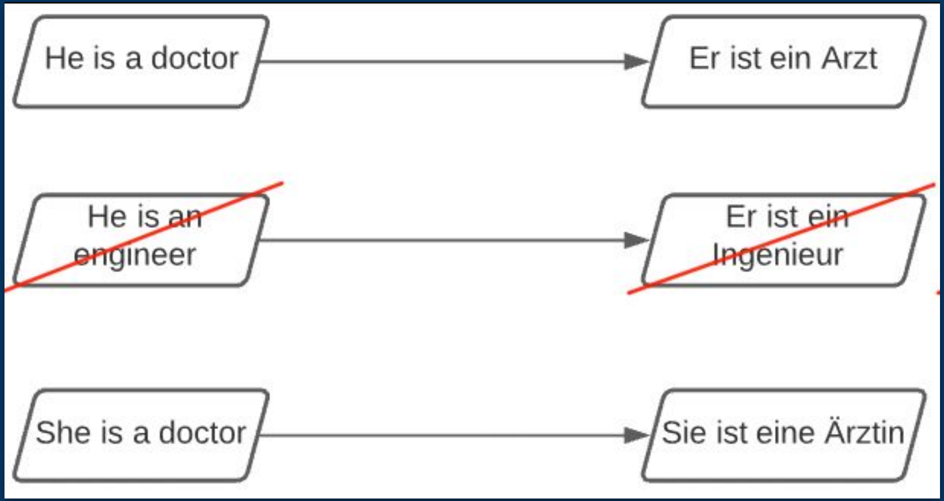
- **Upsampling**

- **Counterfactual augmentation**

Common Approaches to reduce biased training data

Downsampling

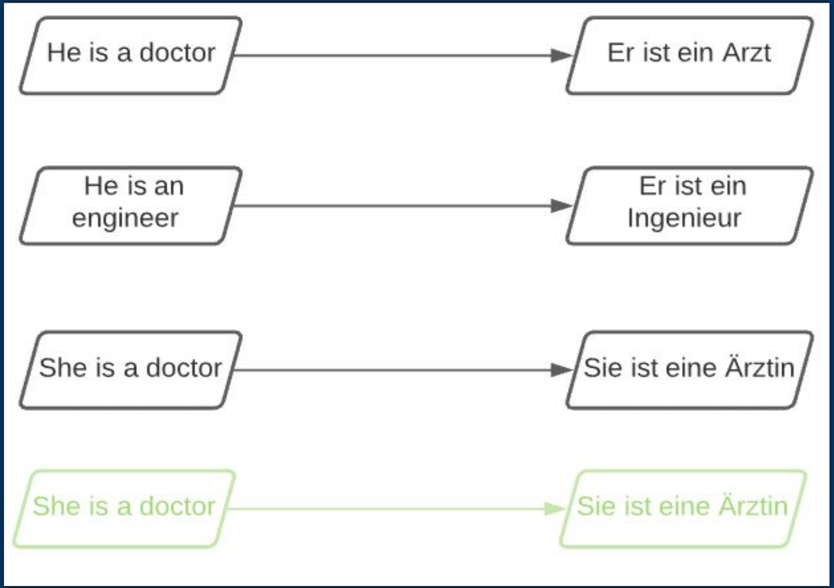
- Remove data until the ratio of gendered terms is balanced for both languages



Common Approaches to reduce biased training data

Upsampling

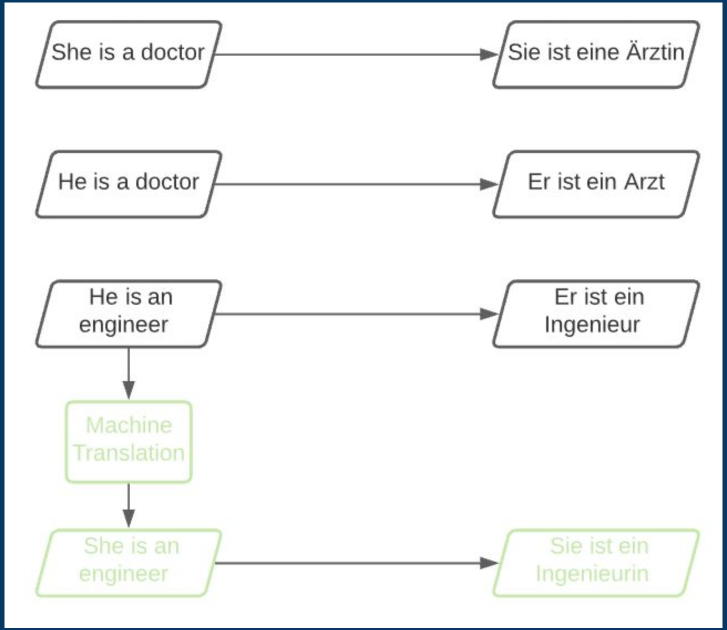
- Add duplicated data until the ratio of gendered terms is balanced for both languages



Common Approaches to reduce biased training data

Counterfactual augmentation

- Automatically introduce counterfactual sentences that include the under-represented gendered terms



Common Approaches to reduce biased training data

- If downsampling: add sentence pair to the final dataset only if it is "gender-balanced"
  common case, if the English side has the same number of male and female entities (50/50)

- If upsampling: include all gendered sentence pairs in the final dataset
  *measure its overall **gender skew** as the total number of **male entitis** in all English sentences **minus** the t**otal number of female entities***

  - Continue to iterate through non balanced gendered sentence pairs adding them to the final dataset again if they reduce the absolute overall skew. (Stop when overall skew 0)

Evaluation metrics

BLEU	Semantic translation quality
Accuracy	The percentage of translations that assign the correct gender to the entity
M:F	Male to Female ratio of sentences
ΔG	Accuracy of translation hypotheses for male and female gender
ΔS	Tendency to stereotype professions

Evaluation of Debias before Training

Table 5 Test set BLEU and WinoMT accuracy, masculine/feminine performance score difference ΔG and pro/anti stereotypical performance score difference ΔS, for the baseline system and systems trained on four different gender-based training sets

System	# Sentence pairs in training data	BLEU	Acc	M:F	ΔG	ΔS
Baseline	17.2 M	42.7	60.1	3.4	18.6	13.4
Downsamped	15.5 M	38.2	47.9	7.1	39.8	8.0
Upsampled	18.1 M	40.4	62.0	3.0	14.6	17.5
Counterfactual	18.6 M	41.1	59.1	3.4	19.0	9.0

Evaluation: Debias

Table 5 Test set BLEU and WinoMT accuracy, masculine/feminine performance score difference ΔG and pro/anti stereotypical performance score difference ΔS, for the baseline system and systems trained on four different gender-based training sets

System	# Sentence pairs in training data	BLEU	Acc	M:F	ΔG	ΔS
Baseline	17.2 M	42.7	60.1	3.4	18.6	13.4
Downsamped	15.5 M	38.2	47.9	7.1	39.8	8.0
Upsampled	18.1 M	40.4	62.0	3.0	14.6	17.5
Counterfactual	18.6 M	41.1	59.1	3.4	19.0	9.0

Interpretation:

- **Accuracy** should be high, **ΔG** and **ΔS** should be close to 0
- High positive **ΔG** tends to give more accurate translations for male subjects
- High positive **ΔS** indicates a tendency to stereotype male and female subjects

Evaluation: Debias

Table 5 Test set BLEU and WinoMT accuracy, masculine/feminine performance score difference ΔG and pro/anti stereotypical performance score difference ΔS, for the baseline system and systems trained on four different gender-based training sets

System	# Sentence pairs in training data	BLEU	Acc	M:F	ΔG	ΔS
Baseline	17.2 M	42.7	60.1	3.4	18.6	13.4
Downsamped	15.5 M	38.2	47.9	7.1	39.8	8.0
Upsampled	18.1 M	40.4	62.0	3.0	14.6	17.5
Counterfactual	18.6 M	41.1	59.1	3.4	19.0	9.0

Interpretation:

- **Accuracy** should be high, **ΔG** and **ΔS** should be close to 0
- High positive **ΔG** tends to give more accurate translations for male subjects
- High positive **ΔS** indicates a tendency to stereotype male and female subjects
- **M:F** strongly correlates with **ΔG**
- **ΔS** can be significantly skewed for systems with very high/low M:F

Evaluation: Debias

Table 5 Test set BLEU and WinoMT accuracy, masculine/feminine performance score difference ΔG and pro/anti stereotypical performance score difference ΔS, for the baseline system and systems trained on four different gender-based training sets

System	# Sentence pairs in training data	BLEU	Acc	M:F	ΔG	ΔS
Baseline	17.2 M	42.7	60.1	3.4	18.6	13.4
Downsamped	15.5 M	38.2	47.9	7.1	39.8	8.0
Upsampled	18.1 M	40.4	62.0	3.0	14.6	17.5
Counterfactual	18.6 M	41.1	59.1	3.4	19.0	9.0

Interpretation:

- **Accuracy** should be high, **ΔG** and **ΔS** should be close to 0
- High positive **ΔG** tends to give more accurate translations for male subjects
- High positive **ΔS** indicates a tendency to stereotype male and female subjects

Removing gender bias by domain adaptation

- Debias a fully trained NMT after the fact by domain adaptation to a tiny gender balanced set
- Domain Adaptation:
  a. ability to apply an algorithm trained in one or more “source domains” to a different (but related) “target domain”
  b. Source and target domain all have the same feature space (but different distributions)

Removing gender bias by domain adaptation

1. Debias a fully trained NMT after the fact by domain adaptation to a tiny gender balanced set
2. creating a **tiny data** set in the form:
3. **with overlap**: “The [PROFESSION] finished [his/her] work”

```
the bailiff finished his work .|Der Gerichtsvollzieher beendete seine Arbeit .
the bailiff finished her work .|Die Gerichtsvollzieherin beendete ihre Arbeit .
the ceo finished his work .|Der Geschäftsführer beendete seine Arbeit .
the ceo finished her work .|Die Geschäftsführerin beendete ihre Arbeit .
the investigator finished his work .|Der Ermittler beendete seine Arbeit .
the investigator finished her work .|Die Ermittlerin beendete ihre Arbeit .
the advisor finished his work .|Der Berater beendete seine Arbeit .
the advisor finished her work .|Die Beraterin beendete ihre Arbeit .
the hunter finished his work .|Der Jäger beendete seine Arbeit .
the hunter finished her work .|Die Jägerin beendete ihre Arbeit .
```

Removing gender bias by domain adaptation

1. Debias a fully trained NMT after the fact by domain adaptation to a tiny gender balanced set
2. creating a **tiny data** set in the form:
3. **with overlap**: “The [PROFESSION] finished [his/her] work”

```
the bailiff finished his work .|Der Gerichtsvollzieher beendete seine Arbeit .
the bailiff finished her work .|Die Gerichtsvollzieherin beendete ihre Arbeit .
the ceo finished his work .|Der Geschäftsführer beendete seine Arbeit .
the ceo finished her work .|Die Geschäftsführerin beendete ihre Arbeit .
the investigator finished his work .|Der Ermittler beendete seine Arbeit .
the investigator finished her work .|Die Ermittlerin beendete ihre Arbeit .
the advisor finished his work .|Der Berater beendete seine Arbeit .
the advisor finished her work .|Die Beraterin beendete ihre Arbeit .
the hunter finished his work .|Der Jäger beendete seine Arbeit .
the hunter finished her work .|Die Jägerin beendete ihre Arbeit .
```

Removing gender bias by domain adaptation

1. Debias a fully trained NMT after the fact by domain adaptation to a tiny gender balanced set
2. creating a **tiny data** set in the form:
3. **without overlap**: “The [ADJECTIVE][man/woman] OR [PROFESSION] finished [his/her] work”

```
the bailiff finished his work .|Der Gerichtsvollzieher beendete seine Arbeit .
```

```
the actor finished her work .|Die Schauspielerin beendete ihre Arbeit .
the actor finished his work .|Der Schauspieler beendete seine Arbeit .
the actuary finished her work .|Die Aktuarin beendete ihre Arbeit .
the actuary finished his work .|Der Aktuar beendete seine Arbeit .
the agent finished her work .|Die Agentin beendete ihre Arbeit .
the agent finished his work .|Der Agent beendete seine Arbeit .
the aggressive man finished his work .|Der aggressive Mann beendete seine Arbeit .
the aggressive woman finished her work .|Die aggressive Frau beendete ihre Arbeit .
the aide finished her work .|Die Adjutantin beendete ihre Arbeit .
the aide finished his work .|Der Adjutant beendete seine Arbeit .
```

WinoMT

- The WinoMT Framework evaluates if a gender prediction is right by setting the pronoun and the grammatical subject into relation.

[The doctor]$_S$ asked [the nurse]$_O$ to help [her]$_P$ in the procedure

- WinoMT contains 3,888 concatenated sentences
  - Equally balanced between male and female genders, as well as between stereotypical and nonstereotypical gender-role assignments
- The labeled subjects can be compared to the so called ‘gold-label’ gendered target, a evaluation set with correctly setted gender.

Result interpretation: Domain Adaptation

	BLEU	Acc	M:F	ΔG	ΔS
Baseline	42.7	60.1	3.4	18.6	13.4
Tiny (No English profession overlap)	*40.6*	*71.2*	*1.7*	*3.9*	*10.6*
Tiny	40.8	78.3	1.3	– 0.7	6.5
Tiny (EWC)	42.2	74.2	1.6	2.2	8.4

Conclusion

- Removing gender bias from MT training data before training is ineffective
  - The approach of debiasing the training data couldn’t outperform the original NMT system
  - [**big losses**] in **translation quality** for [**small to no**] reduction in **gender bias** ✖
- Fine-tuning approach with **domain adaptation** improved the gender-related metrics
  - [**small losses**] in **translation quality** for [**big reduction**] in **gender bias** ✓
- The Fine-tuning approach is far less computationally intensive

Ethical Consideration

Should autonomous intelligent systems be knowingly designed and constructed to reflect the various biased overtly manifest in human societies?

Ethical Consideration

OR

Should they be purposefully designed and constructed to be less biased than those societies currently are?

Ethical Consideration

AND

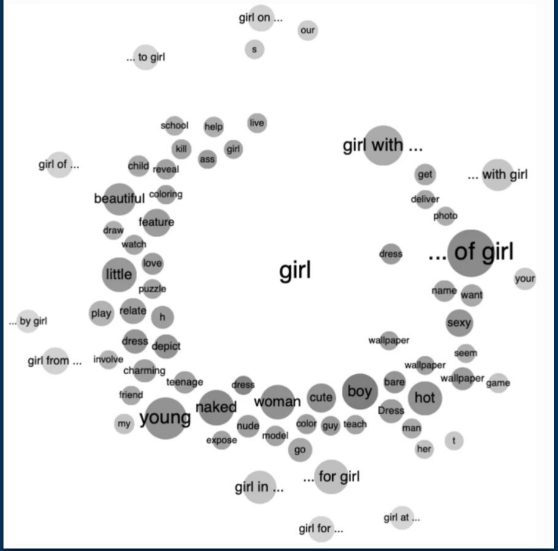
Should we expect such systems to be more ethical than the communities from which their training data was obtained?

ESUPOL Usecase

European Political Sphere

1. Identify male and female politicians used in the query
   a. Generalization of Names to Pronouns
2. Create bubble Clusters segmented by
   a. Gender
   b. Domain
3. Analyze Genderbias per Cluster
   a. Analyze Influence of Search term Suggestions on overall sentiment
   b. Possible Removal of offensive terms
   c. Sentiment Analysis from NLP Lessons

Fig. 1



ESUPOL USECASE

Possible Tools:

- **NLTK** — for POS Tags Analysis of Names using name lists
- **Sketchengine or Word Cloud from NLTK** — for Visual representation of Clusters
- **Spacy** Libraries — for Sentiment Analysis
    - Textblob
    - Sentiws
- **WinoMT** Framework — for Translation Evaluation
- Various other Translation Tools

Sources

The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing

https://link.springer.com/article/10.1007%2Fs10676-021-09583-1

https://github.com/DCSaunders/gender-debias/blob/master/data/handcrafted-nooverlap/handcrafted-nooverlap-ende

https://github.com/DCSaunders/gender-debias/blob/master/data/handcrafted/handcrafted.ende

https://en.wikipedia.org/wiki/Domain_adaptation