Gender bias in Neural Machine Translation

Argentina Anna Rescigno
Eva Vanmassenhove
Johanna Monti
Andy Way

6th October 2020

This work has been supported by the Dublin City University Faculty of Engineering & Computing and the University of Naples "L'Orientale" Department of Literary, Linguistic and Comparative Studies under the Erasmus+ Traineeship project number 2019-1-IT02-KA103-061753

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

Presentation Outline

- **Introduction**
  - A Note on Terminology
  - A Quick Problem Sketch
- **Experimental setup**
  - Compilation of Datasets
  - Description of the MT systems
- **Results & Analysis**
- **Three main points:**
  - Why does this kind of bias matter
  - What is its impact and on whom
  - Why we need to correct this bias
- **Conclusions and Future Work**

Engaging Content
Engaging People

Introduction

Introduction: a note on terminology

Natural Gender
“Gender based on the ***sex*** *or, for neuter, the lack of sex of the referent of a noun, as English girl (feminine) is referred to by the feminine pronoun she, boy (masculine) by the masculine pronoun he, and table (neuter) by the neuter pronoun it.”* *Collins Dictionary* 2018, HarperCollins, London, viewed September 2020 http://www.collinsdictionary.com

Introduction: a note on terminology

Natural Gender	Grammatical Gender
*“Gender based on the **sex** or, for neuter, the lack of sex of the referent of a noun, as English girl (feminine) is referred to by the feminine pronoun she, boy (masculine) by the masculine pronoun he, and table (neuter) by the neuter pronoun it.”* *Collins Dictionary* 2018, HarperCollins, London, viewed September 2020 http://www.collinsdictionary.com	*“Gender based on arbitrary assignment, without regard to the referent of a noun, as in French ‘le livre’ (masculine), “the book,” and German ‘das Mädchen’ (neuter), “the girl.”* *Collins Dictionary* 2018, HarperCollins, London, viewed September 2020 http://www.collinsdictionary.com

Introduction: a note on terminology

Natural Gender	Grammatical Gender	Social Gender
“Gender based on the ***sex*** *or, for neuter, the lack of sex of the referent of a noun, as English girl (feminine) is referred to by the feminine pronoun she, boy (masculine) by the masculine pronoun he, and table (neuter) by the neuter pronoun it.”*	*“Gender based on arbitrary assignment, without regard to the referent of a noun, as in French ‘le livre’ (masculine), “the book,” and German ‘das Mädchen’ (neuter), “the girl.”*	- *Embedded in the lexicon of many languages* - *Systematic structural bias.* - *Masculine forms the default for generic use.*
Collins Dictionary 2018, HarperCollins, London, viewed September 2020 http://www.collinsdictionary.com	*Collins Dictionary* 2018, HarperCollins, London, viewed September 2020 http://www.collinsdictionary.com	

Introduction: a note on terminology

Romance Languages (e.g. ES, FR, IT)

- animate/persons/animals

↓

grammatical gender = natural gender

- inanimate objects

↓

grammatical gender = arbitrary

Introduction: a note on terminology

Romance Languages (e.g. ES, FR, IT)	English
● animate/persons/animals ↓ grammatical gender = natural gender ● inanimate objects ↓ grammatical gender = arbitrary	● grammatical gender is not inflectional ● ***pronominal gender*** → gender expressed through the pronouns = natural gender ● ***gender-neutralization*** of the language

HELLO
my pronouns are
theirs
zim
their
her
ze
she
his
xe
they
them
xim
hers
xey
him
he

Introduction: a quick problem sketch

A simple example:

I am happy!

🇮🇹 Io sono contento! Io sono contenta! [Natural Gender] [Grammatical Gender]

I am happy!

🇫🇷 Je suis heureux! Je suis heureuse! [Natural Gender] [Grammatical Gender]

Introduction: a quick problem sketch

		Subject gender	Predicative nominative gender	Agreement?
English	Mark is an efficient nurse.	M	covered	/
Italian	Mark è un'infermiera efficiente.	M	F	X
French	Mark est une infirmière efficace.	M	F	X
Spanish	Mark es una enfermera eficiente.	M	F	X

Nov 2019

- ➢ **Lack of diversity** → preference for masculine & gender-bias exemptions
- ➢ **Agreement errors**

Engaging Content
Engaging People

Experimental Setup

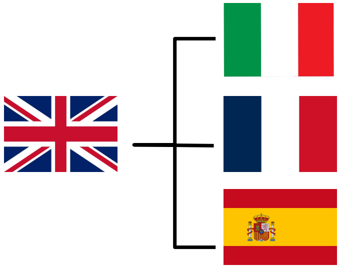
Compilation of Datasets

Gender bias in MT

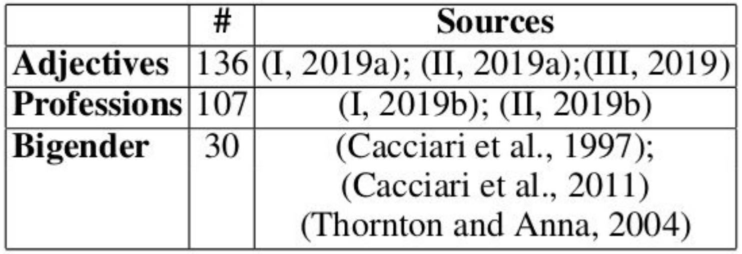
Google Translate

DeepL DeepL Translator

bing Translator Bing Microsoft Translator



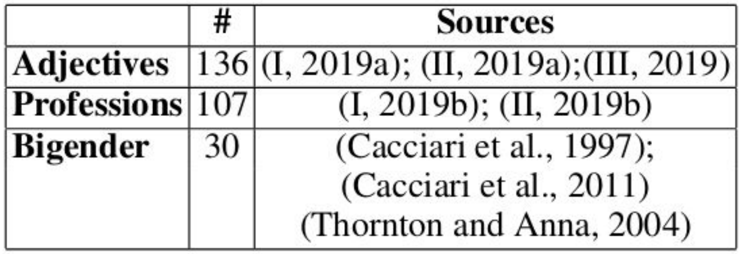
- personality adjectives
- profession nouns
- bigender nouns (in Italian)
    - minimal sentence “I am a(n)...”
    - sentence with a referring adjective



	#	Sources
Adjectives	136	(I, 2019a); (II, 2019a);(III, 2019)
Professions	107	(I, 2019b); (II, 2019b)
Bigender	30	(Cacciari et al., 1997); (Cacciari et al., 2011) (Thornton and Anna, 2004)

Table 1: Overview of adjectives, profession and bigender nouns along with the sources from which they were retrieved

Compilation of Datasets



	#	Sources
Adjectives	136	(I, 2019a); (II, 2019a);(III, 2019)
Professions	107	(I, 2019b); (II, 2019b)
Bigender	30	(Cacciari et al., 1997); (Cacciari et al., 2011) (Thornton and Anna, 2004)

Table 1: Overview of adjectives, profession and bigender nouns along with the sources from which they were retrieved

English	Italian		French		Spanish	
I am an assistant.	Sono un assistente.	M	Je suis un assistant.	M	Soy asistente.	*
I am a beautiful assistant.	**Sono una bellissima assistente.**	**F**	**Je suis une belle assistante.**	**F**	**Soy una bella asistente.**	**F**
I am an efficient assistant.	Sono un assistente efficiente.	M	Je suis un assistant efficace.	M	Soy un asistente eficiente.	M
I am a translator.	Sono un traduttore.	M	Je suis un traducteur.	M	Soy un traductor.	M
I am a beautiful translator.	**Sono una bellissima traduttrice.**	**F**	**Je suis une belle traductrice.**	**F**	**Soy una bella traductora.**	**F**
I am an efficient translator.	Sono un traduttore efficiente.	M	Je suis un traducteur efficace.	M	Soy un traductor eficiente.	M

Description of MT systems

Google Translate

- 2003
- statistical MT system
- 2016 → neural MT system
- 2018 → double alternatives on word level

Description of MT systems

Google Translate

DeepL Translator

- 2017
- convolutional neural networks
- Linguee database (dictionary)
- nine languages supported
- provides not morphological alternatives
- serves also as glossary

Description of MT systems

Google Translate

DeepL Translator

Bing Microsoft Translator

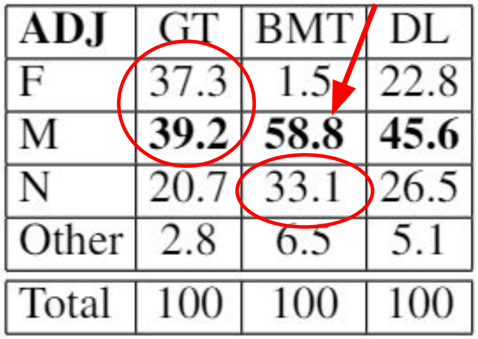
- originally a statistical MT system
- switched to a neural system
- does not provides alternatives but
- provides examples of usage

Engaging Content
Engaging People

Results & Analysis

Results & Analysis

❑ ADJECTIVES

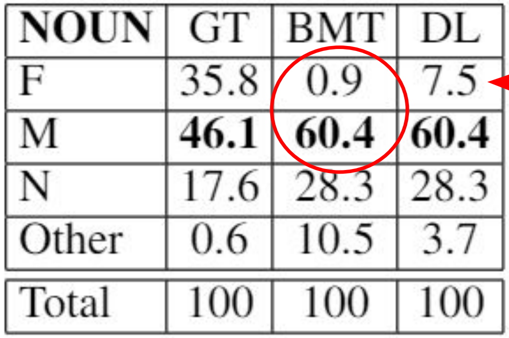


ADJ	GT	BMT	DL
F	37.3	1.5	22.8
M	**39.2**	**58.8**	**45.6**
N	20.7	33.1	26.5
Other	2.8	6.5	5.1
Total	100	100	100

Table 2: Results in % for male (M), female (F) and neutral (N) adjectives generated for EN → IT for GT, BMT and DL. The "Other" label includes all results obtained that do not correspond to the "adjective" category

Results & Analysis

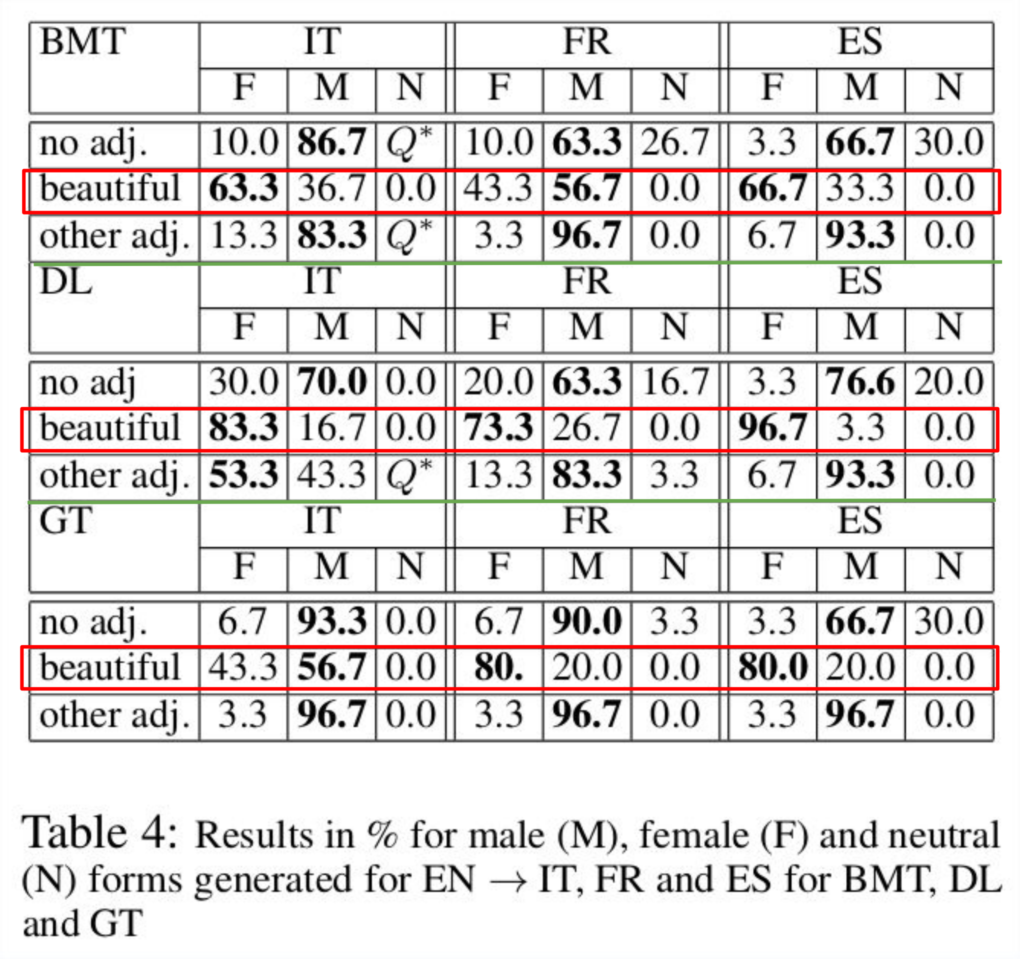
❑ NOUNS



NOUN	GT	BMT	DL
F	35.8	0.9	7.5
M	**46.1**	**60.4**	**60.4**
N	17.6	28.3	28.3
Other	0.6	10.5	3.7
Total	100	100	100

Table 3: Results in % for male (M), female (F) and neutral (N) nouns generated for EN → IT for GT, BMT and DL. The "Other" label includes all results obtained that do not correspond to the "noun" category

Results & Analysis



BMT	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	10.0	**86.7**	Q^*	10.0	**63.3**	26.7	3.3	**66.7**	30.0
beautiful	**63.3**	36.7	0.0	43.3	**56.7**	0.0	**66.7**	33.3	0.0
other adj.	13.3	**83.3**	Q^*	3.3	**96.7**	0.0	6.7	**93.3**	0.0
DL	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj	30.0	**70.0**	0.0	20.0	**63.3**	16.7	3.3	**76.6**	20.0
beautiful	**83.3**	16.7	0.0	**73.3**	26.7	0.0	**96.7**	3.3	0.0
other adj.	**53.3**	43.3	Q^*	13.3	**83.3**	3.3	6.7	**93.3**	0.0
GT	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	6.7	**93.3**	0.0	6.7	**90.0**	3.3	3.3	**66.7**	30.0
beautiful	43.3	**56.7**	0.0	**80.**	20.0	0.0	**80.0**	20.0	0.0
other adj.	3.3	**96.7**	0.0	3.3	**96.7**	0.0	3.3	**96.7**	0.0

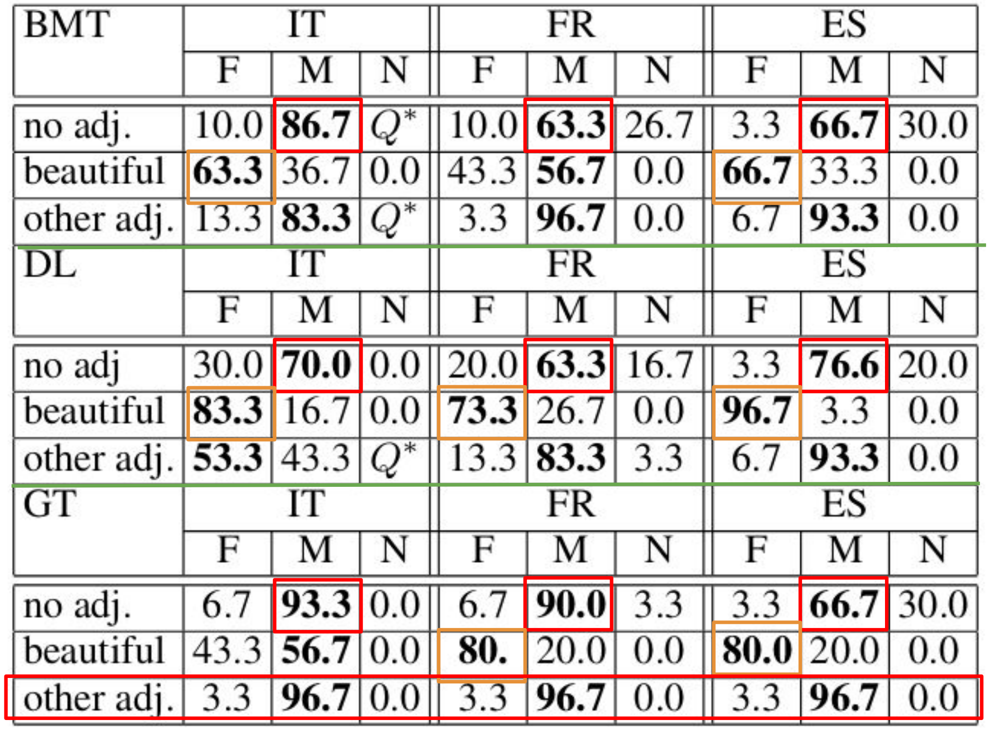
Table 4: Results in % for male (M), female (F) and neutral (N) forms generated for EN → IT, FR and ES for BMT, DL and GT

- *beautiful*

other adjectives:

- *efficient*
- *intelligent*
- *sad*
- *famous*

Results & Analysis



BMT	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	10.0	**86.7**	Q^*	10.0	**63.3**	26.7	3.3	**66.7**	30.0
beautiful	**63.3**	36.7	0.0	43.3	**56.7**	0.0	**66.7**	33.3	0.0
other adj.	13.3	**83.3**	Q^*	3.3	**96.7**	0.0	6.7	**93.3**	0.0
DL	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj	30.0	**70.0**	0.0	20.0	**63.3**	16.7	3.3	**76.6**	20.0
beautiful	**83.3**	16.7	0.0	**73.3**	26.7	0.0	**96.7**	3.3	0.0
other adj.	**53.3**	43.3	Q^*	13.3	**83.3**	3.3	6.7	**93.3**	0.0
GT	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	6.7	**93.3**	0.0	6.7	**90.0**	3.3	3.3	**66.7**	30.0
beautiful	43.3	**56.7**	0.0	**80.**	20.0	0.0	**80.0**	20.0	0.0
other adj.	3.3	**96.7**	0.0	3.3	**96.7**	0.0	3.3	**96.7**	0.0

Table 4: Results in % for male (M), female (F) and neutral (N) forms generated for EN → IT, FR and ES for BMT, DL and GT

- *beautiful*

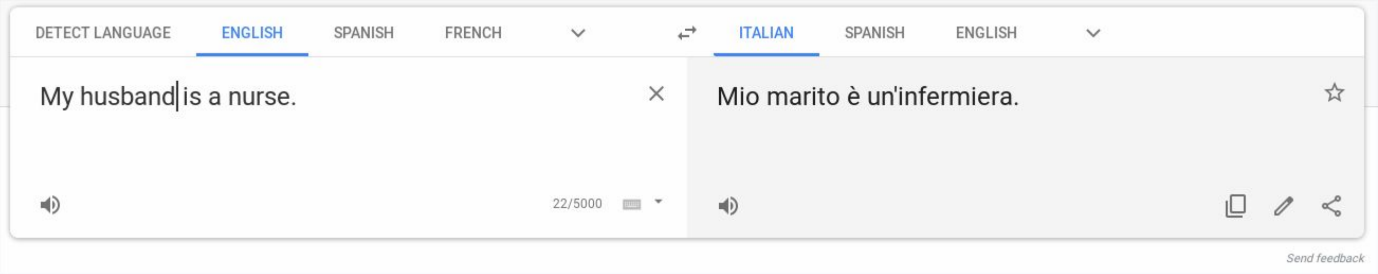
other adjectives:

- *efficient*
- *intelligent*
- *sad*
- *famous*

Engaging Content
Engaging People

iMpacT

- **From a linguistic point of view:**
  - Avoiding basic gender agreement mistakes

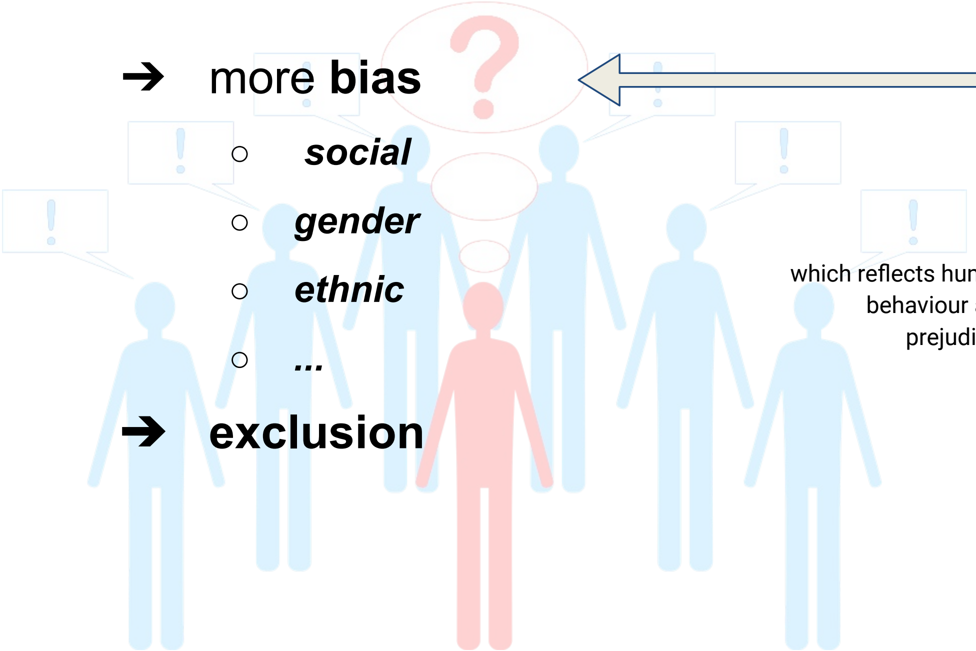


My husband is a nurse.

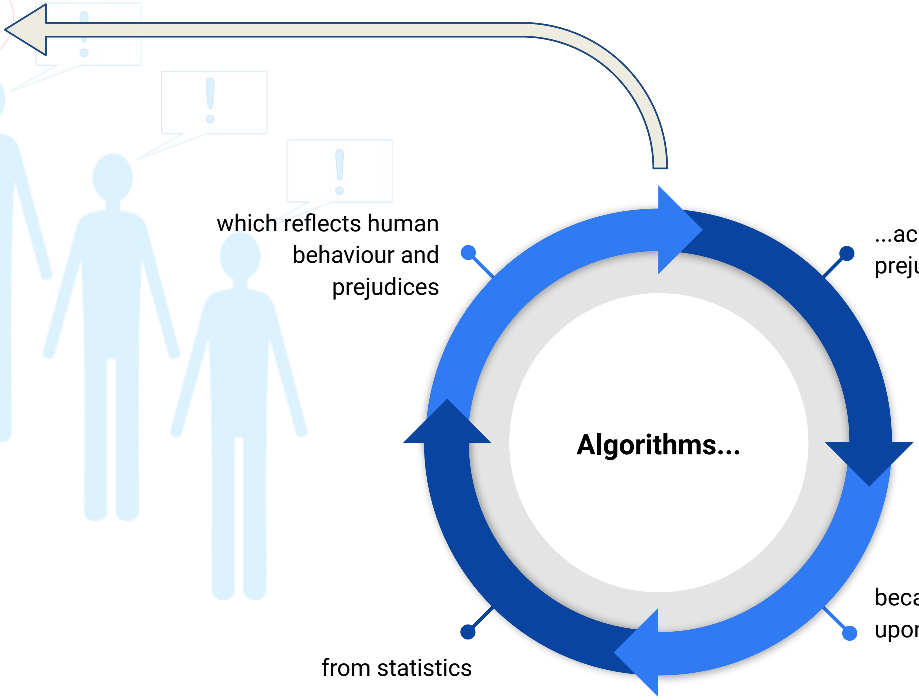
Mio marito è un'infermiera.

- **From a technological point of view:**
  - Solving these issues is not trivial (see attempts Google)
  - Black box of NLP (we have no/little control over the actual output that are being generated)
- **From a societal/ethical point of view:**
  - Identifying biases in current state-of-the-art systems is important so they don't end up getting mistaken for ‘objective’ translations
  - if an MT system is being used without human in the loop: real-world consequences

Break the cycle



- ➔ more **bias**
  - ○ ***social***
  - ○ ***gender***
  - ○ ***ethnic***
  - ○ ***…***
- ➔ **exclusion**



Algorithms...

...acquire social prejudices

because they are built upon training data

from statistics

which reflects human behaviour and prejudices

Engaging Content
Engaging People

Conclusion and Future Work

Conclusion and Future Work

Conclusion:

- Remove gender bias in training data
- Train algorithms to address the problem
- Stop using masculine “neutral” in machine learning texts
- Evaluation of gender phenomena is challenging

Future Work:

- Extend to other language pairs (different languages → different gender phenomena)
- Larger evaluation of more diverse set of words
- Create language specific challenge sets to evaluate how biased is an MT system
- Train our own MT system to verify whether machine bias influences the output of the translation

ADAPT

Engaging Content
Engaging People

Thank you for your attention!

This work has been supported by the Dublin City University Faculty of Engineering & Computing under the Daniel O'Hare Research Scholarship scheme and by the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund, by the European Commission as part of the FALCON project (contract number 610879)

DCU

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

sfi Science Foundation Ireland

References

Bond, E., 2020. Cambridge Researchers Tackle Neural Machine Translation'S Gender Bias | Slator. [online] Slator. Available at: <https://slator.com/machine-translation/cambridge-researchers-tackle-neural-machine-translations-gender-bias/> [Accessed 29 September 2020].

Caliskan, A., Bryson, J. and Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), pp.183-186.

Devlin, H., 2017. AI Programs Exhibit Racial And Gender Biases, Research Reveals. [online] the Guardian. Available at: <https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals> [Accessed 29 September 2020].

Monti, J., 2017. Questioni di genere in traduzione automatica. In: A. de Meo, L. di Pace, A. Manco and J. Monti, ed., Al femminile. Scritti linguistici in onore di Cristina Vallini. Firenze: Cesati, pp.411-431.

Prates, M., Avelar, P. and Lamb, L., 2019. Assessing gender bias in machine translation: a case study with Google Translate. Neural Computing and Applications, 32(10), pp.1-19.

Saunders, D. and Byrne, B., 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. *arXiv: 2004.04498v3*

Vanmassenhove, E., Shterionov, D. and Way, A., 2019. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. *arXiv: 1906.12068*

Zou, J. and Schiebinger, L., 2018. AI can be sexist and racist — it's time to make it fair. Nature, 559(7714), pp.324-326.

Engaging Content
Engaging People

Contact info

Argentina A. Rescigno: argentina.res@gmail.com

Eva Vanmassenhove: vanmassenhove.eva@gmail.com

Johanna Monti: johmonti@gmail.com

Andy Way: andy.way@adaptcentre.ie