Good but not always fair
tackling gender bias in automatic translation

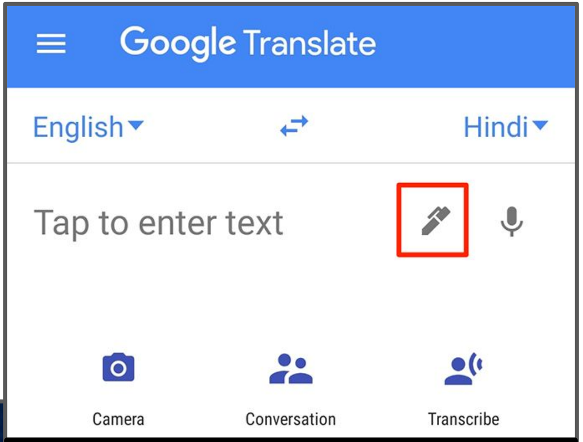
Luisa Bentivogli
MT Unit @ FBK - Italy

Translating and the Computer - TC44 **Luxembourg, 24-25 November 2022**

LANGUAGE TECHNOLOGY IS UBIQUITOUS

Chatbot

Vocal assistant



Machine Translation

Spell checker
&
auto-completion

LANGUAGE TECHNOLOGY IS NOT NEUTRAL

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By JAMES VINCENT
Mar 24, 2016, 11:43 AM GMT+1 |

Rhett Jones
10/10/18 10:32AM • Filed to: ALGORITHMS
79 3

Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes with the Word 'Women's'

LANGUAGE TECHNOLOGY IS NOT NEUTRAL

Q for @LinkedIn: I'm a French translator. When French clients search for "translator" on your site, they search for the default masculine term. Your algorithm does not return translators who use the feminine form, so I am left out of the search results. Help?

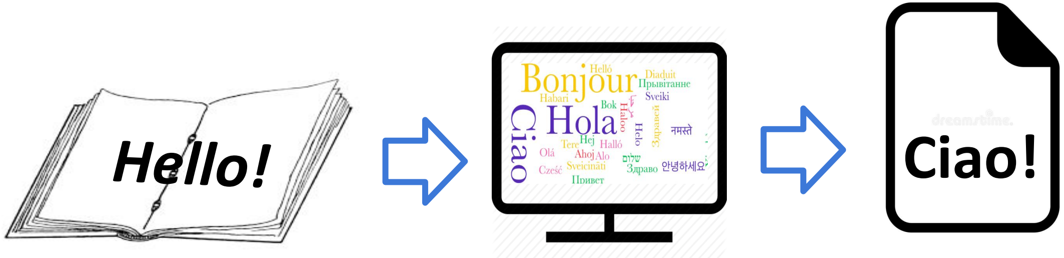
@queerterpreter

10:00 PM · Apr 23, 2020 · Twitter Web App

MACHINE TRANSLATION

- MT popularity: Neural Paradigm
    - Increasingly fluent and adequate translations
    - Improvements on syntax, lexicon, morphology

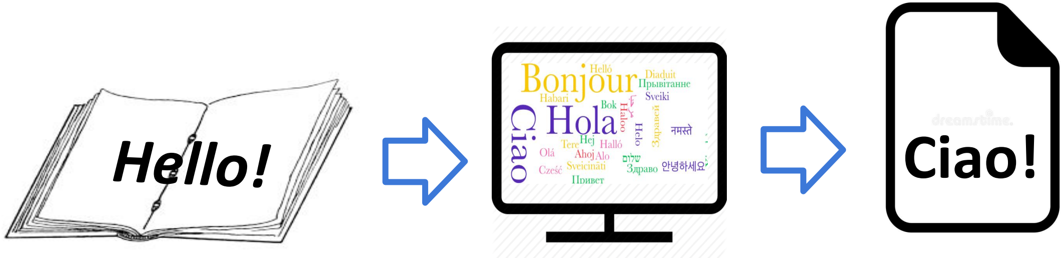
(Bentivogli et al, 2016)



MACHINE TRANSLATION

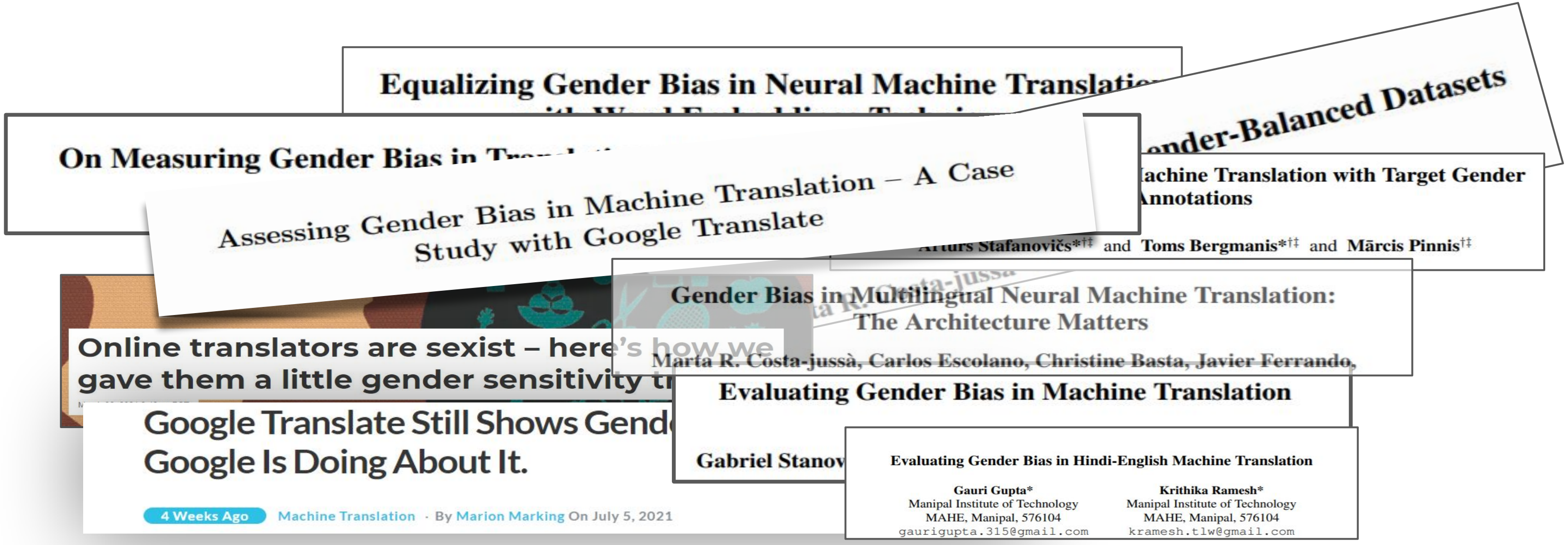
- MT popularity: Neural Paradigm
    - Increasingly fluent and adequate translations
    - Improvements on syntax, lexicon, morphology

→ **but gender translation is an issue**



GENDER BIAS IN MT

- a rapidly emergent field that lacks cohesion

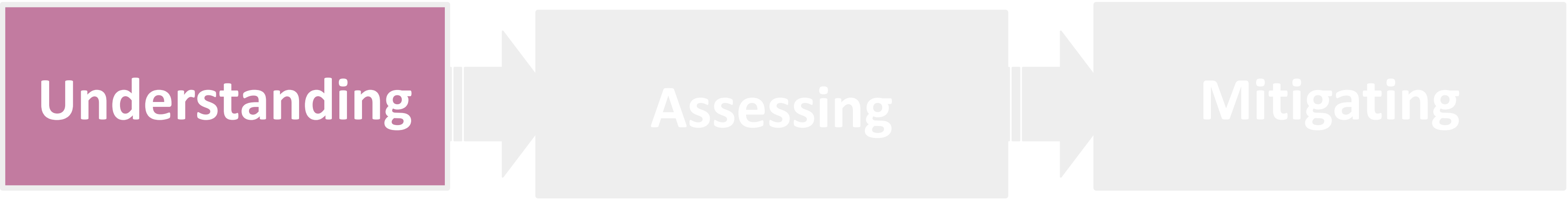


GENDER BIAS IN MT

- a rapidly emergent field that lacks cohesion
  → review within a unified framework

GENDER BIAS IN MT

- a rapidly emergent field that lacks cohesion
  → review within a unified framework



WHAT IS BIAS?

The word “bias” has multiple meanings (Campolo et al., 2017)

- **Statistics**:
  - divergence from an expected value, neutral meaning
- **Cognitive science**:
  - outcome of psychological heuristics, i.e. mental shortcuts that can be critical to support prompt reactions

WHAT IS BIAS?

- **Normative sense:** judgement based on preconceived notions or prejudices vs. impartial evaluation of facts

“Computer systems that *systematically* and *unfairly* discriminate against certain individuals or groups of individuals in favor of others”

(Friedman and Nissenbaum, 1996)

WHAT IS BIAS?

- Bias investigation is not only a scientific and technical endeavour but also an ethical one
- Normative process → *What* is deemed as an harmful behavior, *how* and to *whom*? (Blodgett et al., 2020)

WHICH BIASED BEHAVIOURS IN MT?

TYPES OF HARMS (Crawford, 2017)

Representational harms	diminishing the representation of social groups and their identity, which, in turn, affects attitudes and beliefs
Allocational harms	uneven distribution of resources allocated by a system

WHICH BIASED BEHAVIOURS IN MT?

TYPES OF HARMS (Crawford, 2017)

Representational harms	*Under-representation*
	Stereotyping
Allocational harms	*Quality of service*

HARM: UNDER-REPRESENTATION

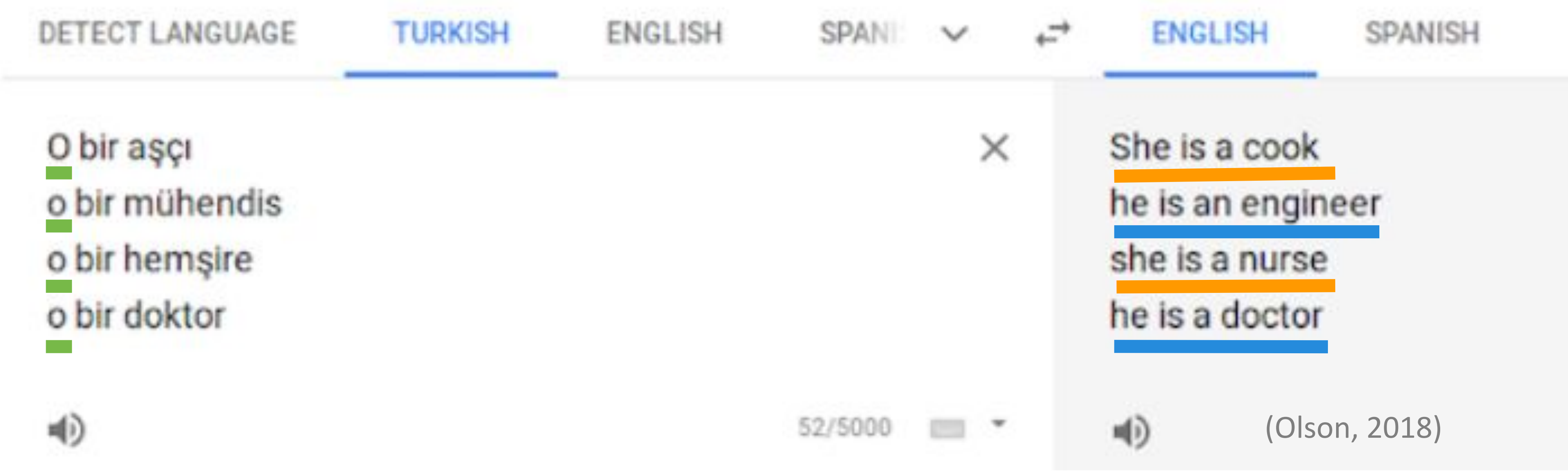
- reduction of visibility through language


Original Spanish Text	Automated Translations	
	Google Translate	Systran
El País March 22, 2011 Desde que Londa Schiebinger llegó a la Universidad tuvo claro que era lo suyo. Primero como estudiante y después como profesora. "Decidí quedarme en la enseñanza	Since Londa Schiebinger came to the University was clear that was his thing. First as a student and later as a teacher. "I decided to stay in education because you learn every day. I love	Ever since Londa Schiebinger arrived at the University knew clearly that he was his. First like student and later like professor. "I decided to remain in education because every day is learned.The knowledge

(Schiebinger, 2014)

HARM: STEREOTYPING

- propagating negative generalizations of a social group



(Olson, 2018)

HARM: QUALITY OF SERVICE

- disparity in the quality of the offered service

My short bio

THE ROOTS OF GENDER BIAS

Where does the problem come from?

- (Some) concurring factors…
    - cross-linguistic, sociolinguistic
    - societal
    - technical
- … corroborating one with another

GENDER ACROSS LANGUAGES

The linguistic structures used to refer to the extra-linguistic reality of gender vary across languages (Stahlberg et., 2007):

1. Genderless languages
2. Notional gender languages
3. Grammatical gender languages

GENDER ACROSS LANGUAGES

The linguistic structures used to refer to the extra-linguistic reality of gender vary across languages (Stahlberg et., 2007):

1. **Genderless languages**
   - gender repertoire at its minimum
   - kinship terms and address

   e.g. *brother/sister* → in Finnish *sisko/veli*

GENDER ACROSS LANGUAGES

The linguistic structures used to refer to the extra-linguistic reality of gender vary across languages (Stahlberg et., 2007):

2. **Notional Gender Languages**
   - Pronominal gender (he/she)
   - Lexical gender (boy/girl)
   - Some residual derivation (actor/actress)
   - Compounds (chairman/chairwoman)

GENDER ACROSS LANGUAGES

The linguistic structures used to refer to the extra-linguistic reality of gender vary across languages (Stahlberg et., 2007):

3. **Grammatical Gender Languages**
   - the gender identity of a referent is overtly expressed on numerous POS (nouns, adjective, determiners,verbs…)
   - complex morphosyntactic system of agreement *e.g.*

     (ES) *El es un buen amigo* vs. *Ella es una buena amiga*

GENDER ACROSS LANGUAGES

- Translating into **grammatical gender languages**

En: «a good friend»

It: «una buona amica» *(Fem.)*

It : «un_ buon_ amico» *(Masc.)*

- One-to-many problem

SOCIAL GENDER CONNOTATIONS

How linguistic expressions are connoted, deployed and perceived

- Semantic derogation
  e.g. couturier (fashion designer) vs. couturière (seamstress)
  governor vs. governess (Schultz, 1975)

R La Repubblica

Sanremo, Beatrice Venezi: "Direttore, non direttrice". E i social si spaccano sulla scelta

Sanremo, Beatrice Venezi: "Direttore, non direttrice". E i social si spaccano sulla scelta. La musicista sul palco dell'Ariston ha chiesto ad ...

6 mar 2021

SOCIAL GENDER AND TRANSLATION

"Same **cook** I suppose, Maxim?"

French: la même cuisinière
Italian: lo stesso cuoco
Spanish: el mismo cocinero
Portuguese: a mesma cozinheira
German: dieselbe Köchin

- social connotations of gender influence translation choices

  (Wandruszka 1969, cited in Nissen, 2002: 32)

  → translation adapted according to translators' societal expectations

WHAT ARE THE SOURCES OF BIAS?

- *Training data bias* as an overloaded term (Suresh & Guttard, 2019)

Categorizing sources of bias (Friedman & Nissenbaum, 1996):

- **Pre-existing bias:** rooted in practices, institutions, attitudes
- **Technical bias:** due to technical decisions
- **Emergent bias:** arise in interaction with users

WHAT ARE THE SOURCES OF BIAS?

Pre-existing bias: rooted in practices, institutions, attitudes

- Europarl Corpus (Kohen, 2005)
    - 30% sentences uttered by women (Vanmassenhove, 2018)

→ *historical bias that hampered women's access to political positions*

- Social Connotations and Language use
    - explicit female markings for doctor (*female, woman, lady* doctor) (Romaine, 2001)

→ *qualitative asymmetries: how language is deployed and perceived*

WHAT ARE THE SOURCES OF BIAS?

Technical bias: due to technical constraints and decisions

WHAT ARE THE SOURCES OF BIAS?

Technical bias: due to technical constraints and decisions

- Data creation/curation/ annotation
  - qualitative and quantitative misrepresentation of certain demographics
  - annotations which do not reflect the information in data

WHAT ARE THE SOURCES OF BIAS?

Technical bias: due to technical constraints and decisions

- Data curation/da
  - qualitative and
    certain demogra
  - annotations which do not reflect the information in data

Gender inference based on e.g., voice, pictures, proper names

WHAT ARE THE SOURCES OF BIAS?

Technical bias: due to technical constraints and decisions

- Models design
  - *algorithmic bias* that leads under-represented feminine forms to further decrease in an MT output (Vanmassenhove et al., 2020)
  - chosen components can amplify bias (e.g. word segmentation)

WHAT ARE THE SOURCES OF BIAS?

Technical bias: due to technical constraints and decisions

- Evaluation procedures
  - gender asymmetries in test data reward biased predictions
  - aggregate measures can hide subgroup underperformance

WHAT ARE THE SOURCES OF BIAS?

Emergent bias: a system is used in a different context than the one it was designed for, result of changing values

- MT systems that are not intentionally envisioned for a diverse range of users will not generalize for the feminine/non-binary segment of the population
    - in interaction with an MT system, women will likely be misgendered / linguistic style not preserved
      (Hovy et al., 2020)

GENDER BIAS IN MT

Understanding

Assessing

Mitigating

ASSESSING GENDER BIAS

Traditional metrics and Generic Test sets are unsuitable

>>> **Gender Bias Evaluation Test Sets** (GBETs) (Sun et al,. 2019)

→ isolate gender as a variable

→ MT GBETS: **challenge** or **natural** datasets

GBET BENCHMARKS

- **Challenge datasets**

(Prates *et al.*, 2018; Cho *et al.*, 2019; Escudé Font & Costa-jussà, 2019; Stanovsky *et al.*, 2019)

⟶ synthetic *ad-hoc* sentences focusing on (occupational) stereotypes
⟶ controlled environment but limited variety of phenomena

(En) I've known *her* for a long time, my *friend* works as an *accounting clerk*. (Es) *La* conozco desde hace mucho tiempo, mi *amiga* trabaja como *contable*.
(En) I've known *him* for a long time, my *friend* works as an *accounting clerk*. (Es) *Lo* conozco desde hace mucho tiempo, mi *amigo* trabaja como *contable*.

Occupation test **(Font & Costa-jussa, 2019)**

GBET Benchmarks

- **Natural datasets** (Habash *et al.*, 2019; Bentivogli et al., 2020)

→ selected/annotated gender instances from conversational data
→ more authentic conditions but treat all gendered words equally

Src	She came back to meet two of her **dearest friends, these older** women...
Ref-IT	Tornava per incontrare un paio **delle sue** più **care amiche**, **queste** signore **anziane**

***MuST-SHE* (Bentivogli *et al.*, 2020)**

GENDER BIAS IN MT

Understanding → Assessing → Mitigating

MITIGATING APPROACHES

Different strategies:

- Model debiasing on general-purpose MT models
  - architectural changes and dedicated training procedures

- Debiasing through external components
  - external dedicated components in the inference phase

MITIGATING APPROACHES: TRAINING TIME

Based on counterfactual data augmentation (CDA) (Saunders & Byrne, 2020)

- CDA: creation of synthetic sentences with balanced F/M representation

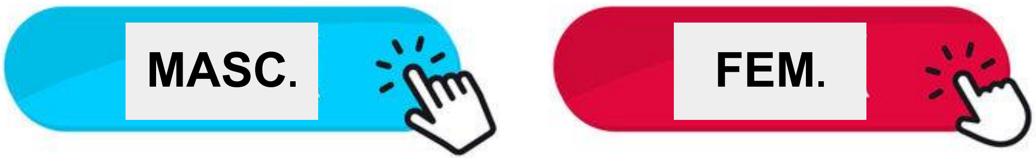
Src	The [PROFESSION] finished [his|her] work.
It-M Ref	[PROFESSION] ha finito il suo lavoro.
It-F Ref	[PROFESSION] ha finito il suo lavoro.

- MT model is fine-tuned on such parallel set

Mitigating Approaches: Inference Time

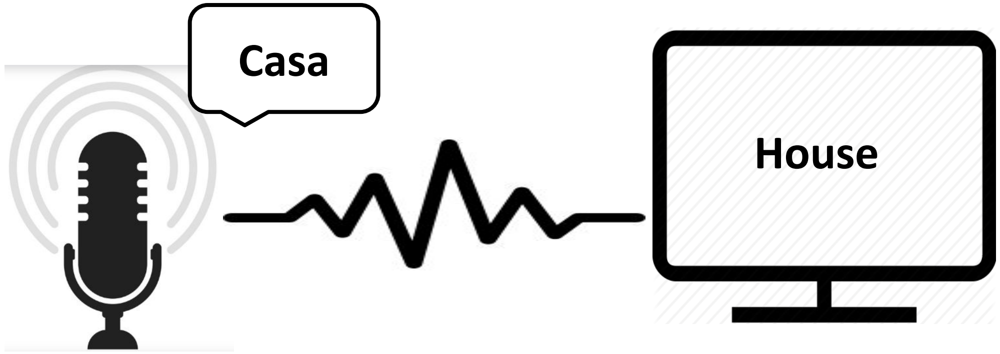
Gender Re-inflection (Habash et al., 2019; Alhafni et al., 2020)

- Scenario: 1-2 person references e.g., *I am/ you are a student*
- Post-processing component re-inflecting into MASC/FEM forms
    - the user chooses the appropriate form



MT@FBK RESEARCH

- **Speech Translation (ST)** is the task of translating audio speech in one language into text in another language



MuST-SHE

Test set for evaluating gender translation in MT & ST

- **Natural Spoken language:** TED Talks data
- **Aligned** (*audio-transcript-translation*) triplets
- **Multilingual:** En→It, En→Fr, En→Es
- **Common subset:** cross-lingual comparisons
- **Gender-sensitive design:** each segment contains at least one English gender-neutral word translated into the corresponding masculine/feminine target word(s)

(Bentivogli *et al.*, ACL 2020)

MuST-SHE

Categorization of gender phenomena

Category 1 | No gender info (apart from audio)

"I'm a good friend" uttered by a man/woman

Category 2 | Gender info in utterance content

"<u>he/she</u> is a good friend"

MuST-SHE

- Natural variety of (balanced) **Fem** and **Masc** phenomena
- Each target **gender-marked word** annotated with its <*swapped*> form

Src-en	she… **the first** Somali **senator**
Ref-es	… **la**<*el*> **primera**<*primero*> **senadora**<*senador*> somalí

MuST-SHE ENRICHMENT

>>> New annotation layers

- POS (word-level)

POS & CLASS

- Articles
- Pronoun
- Adj-det
- Adj-des
- Verb
- Noun

closed class

open class

Src-en	she… **the first** Somali **senator**
Ref-es	… **la** **<Art>** **primera** **<Adj-det>** **senadora** **<Noun>** somalí

MuST-SHE Enrichment

>>> New annotation layers

- POS (word-level)
- AGR (chain-level)

AGREEMENT

- Dependency among words
- Phrase level

e.g. *Noun+modifiers*

Src-en	*she*… **the first** Somali **senator**
Ref-es	… **[la primera senadora]** $_{\textbf{<AGR>}}$ somalí

RESEARCH QUESTIONS

Assessing

- How are ST systems affected by the problem of gender bias?
- Are ST systems exploiting audio information to translate gender?

(Bentivogli *et al.*, ACL 2020)

→ ST systems are biased

→ Beyond MT textual modality:

  - Direct ST leverages cue from audio input
  - Relying on audio signal alone can be problematic?

RESEARCH QUESTIONS

Assessing

- Investigation of algorithmic bias: can word segmentation hinder or favor (feminine) gender translation?

(Gaido,Savoldi *et al.*, ACL Findings 2021)

→ the segmentation method impacts models' ability to translate gender (analysis on 5 different methods)

- *BPE* leads to higher overall translation quality
- *Char* leads to higher gender translation accuracy

RESEARCH QUESTIONS

Assessing

- How are different part-of-speech impacted by gender bias?
- How do systems deal with gender agreement?

(Savoldi *et al.*, ACL 2022)

Extensive *manual* and automatic analysis:

→ POS are not equally biased → nouns the most impacted
→ Respecting agreement is not an issue in current systems
→ benchmarks fail to recognize neutral language in system output
→ higher generic performance do not grant advantage for gender

RESEARCH QUESTIONS

Assessing

- Dynamic perspective: does gender translation improve, worsen, or reach a plateau during training?
- How does gender bias relate to progress of generic performance?

(Savoldi *et al.*, GeBNLP 2022)

➔ Feminine gender learnt late over the course of training and does not reach plateau at the end of training

- ◆ Training stopped according to overall quality are not suitable to account for gender bias

RESEARCH QUESTIONS

Mitigating

- Test different debiasing strategies to improve gender translation related to the **speaker** in a scenario where it is **known**
- Avoid the usage of biometric features

Gender-aware ST:

→ notable improvement for feminine gender translation

→ is able to ignore audio features and rely on the provided speaker's gender information

TO CONCLUDE: what now?

- **Advancements** only reported in terms of *performance*
  - how do they reduce the addressed harm?

- **No conclusive state-of-the-art method for bias mitigation**
  - Response to specific aspects of the problem with *modular solutions*
    - Can they be integrated within the same MT system? How?

TO CONCLUDE: *what now?*

- **Gender bias in MT is a socio-technical problem**
  - engineering interventions alone are not a panacea
  - integration with long-term multidisciplinary commitment and practices

There is plenty of (interdisciplinary) ground to cover...

TO CONCLUDE: *where to?*

Interpretability
Algorithmic side

Beyond
Dichotomies

Human-in-the-loop

GENDER INCLUSIVE LANGUAGE

To date, gender bias mitigation in MT is focussing only on the masculine/feminine dichotomy

- Direct Non-Binary Language: increase the visibility of non-binary individuals
- Indirect Non-binary Language: overcomes gender specifications

GENDER INCLUSIVE LANGUAGE

- Direct Non-Binary Language: grassroots efforts
  - Innovative: neomorphemes (-ə), neopronouns (hen)
  - Creative: “emojiself” pronouns
  - Mostly still ungrammatical

>>> development (and acceptance) of such forms progresses at different paces across languages and cultures

GENDER INCLUSIVE LANGUAGE

- Indirect Non-Binary Language: top-down reccomendations
  - Neutral expressions: *humankind vs.* mankind
  - Endorsed for many official documents (Papadimoulis, 2018)
  - A challenging goal for grammatical gender languages

GENDER INCLUSIVE LANGUAGE

- Indirect Non-Binary Language: top-down reccomendations
  - Neutral expressions: *humankind vs.* mankind
  - Endorsed for many official documents (Papadimoulis, 2018)
  - A challenging goal for grammatical gender languages

GENDER NEUTRAL (MACHINE) TRANSLATION

- methods and benchmarks

HUMAN IN THE LOOP

Language technologies are built **by people…**

- Gender bias attested also for **rule-based MT** (Frank et al., 2004)
  - lack of feminine forms in dictionaries
  - lack of morphological rules for feminine

HUMAN IN THE LOOP

Language technologies are built **by people**…

reflect on the background, diversity and biases of people involved in the MT pipeline - annotators, translators, developers - and its implications on the models

HUMAN IN THE LOOP

Language technologies are built **for people…**

→ to date evaluations on gender bias in MT are restricted to lab tests

- Studies relying on participatory design and HCI approaches
- Consider different MT users, including translators (Ragni & Vieira, 2020)

INVOLVING TRANSLATORS

Productivity:

- Overall translation quality vs. gender translation accuracy
  - do suggestions from a de-biased MT really help translators?
  - is it easier/quicker to correct gender errors or other errors?

Ethics:

- MT errors pose serious risks, MT suggestions prime translators
  - Do translators working with biased MT propagate it? (post-edits become training data…)

The "Gender Bias" Group

Beatrice Savoldi

Marco Gaido

Matteo Negri

Luisa Bentivogli

FONDAZIONE
BRUNO KESSLER

The “Gender Bias” Group

MT
MACHINE
TRANSLATION

Beatrice Savoldi

Marco Gaido

Matteo Negri

Luisa Bentivogli

amazon | science

Amazon Research Awards

Dennis Fucci

Andrea Piergentili

Thanks for listening!

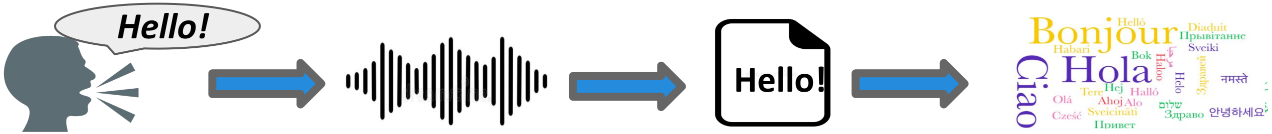
@fbk_mt

Additional slides

SPEECH TRANSLATION MODELS

CASCADE APPROACH

ASR component *MT component*

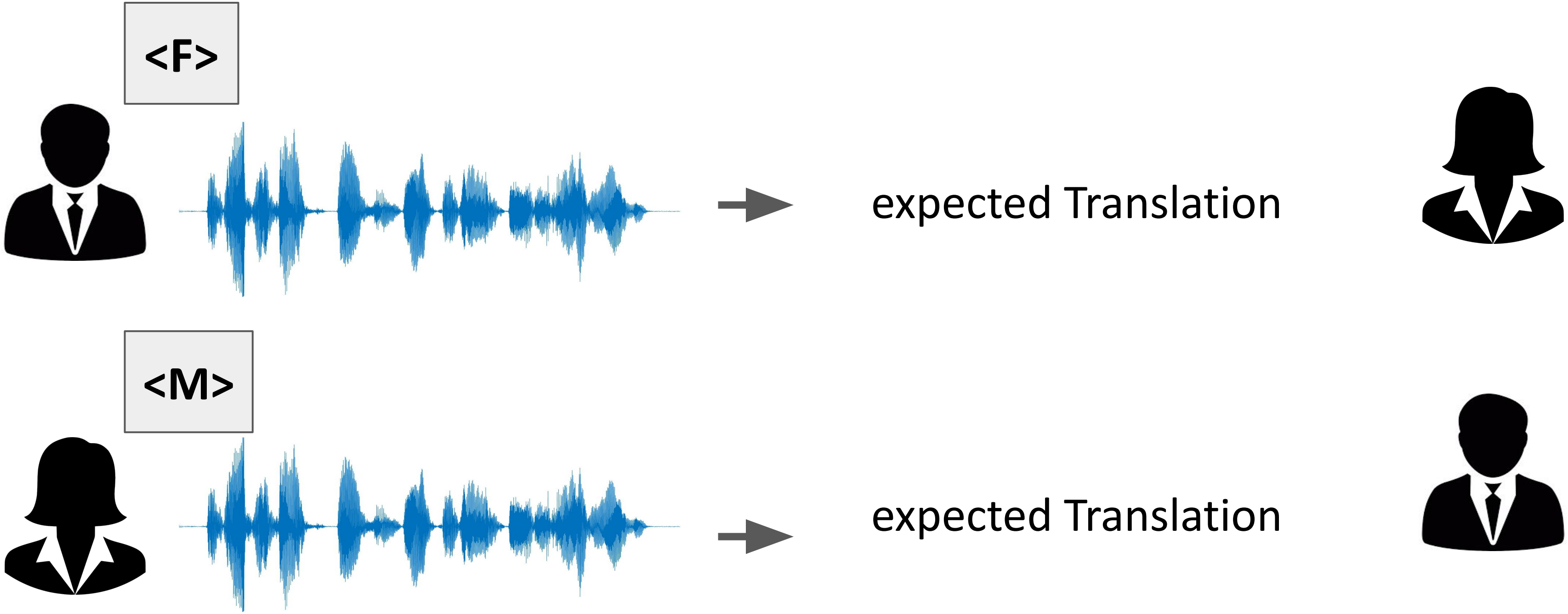


DIRECT APPROACH

Direct translation without intermediate representation

CONFLICTING VOCAL CHARACTERISTICS AND TAGS

- Evaluate on **MuST-SHE *Wrong-Ref***



MITIGATING APPROACHES

Different strategies:

1. Counterfactual data augmentation (CDA) - based (Saunders & Byrne, 2020)
2. Gender Tagging (Vanmassenhove et al., 2018; Stafanovičs et al., 2020)
3. Gender Re-Inflection (Habash et al., 2019; Alhafni et al., 2020)

>> Interventions accounting for ''technical bias''

Mitigating Approaches

- **Based on counterfactual data augmentation** (CDA) (Saunders & Byrne, 2020)
  - CDA: creation of synthetic sentences with balanced F/M representation
  - MT model is fine-tuned on such parallel set

Src	The [PROFESSION] finished [his|her] work.
It-M Ref	[PROFESSION] ha finito il suo lavoro.
It-F Ref	[PROFESSION] ha finito il suo lavoro.

MITIGATING APPROACHES

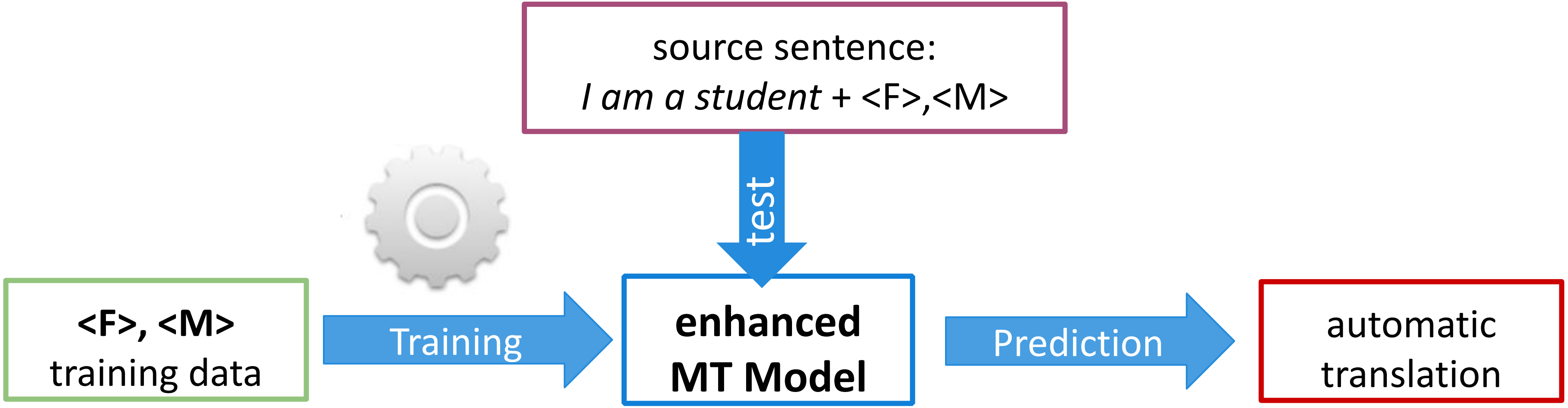
- **Based on counterfactual data augmentation** (CDA) (Saunders & Byrne, 2020)
  - CDA: creation of synthetic sentences with balanced F/M representation
  - MT model is fine-tuned on such parallel set

Src	The [PROFESSION] finished [his|her] work.
It-M Ref	[PROFESSION] ha finito il suo lavoro.
It-F Ref	[PROFESSION] ha finito il suo lavoro.

→ Helpful for stereotyping scenario with pre-defined list of lexicon, but does not cover under-representation on variable language data

MITIGATING APPROACHES

- **Gender Tagging** (Vanmassenhove et al., 2020)
  - Fed a <F>, <M> tag representing speaker's gender to each source sentence, both at training and inference time



Mitigating Approaches

- **Gender Tagging** (Vanmassenhove et al., 2020)
  - Fed a <F>, <M> tag representing speaker's gender to each source sentence, both at training and inference time

→ requires acquiring metadata and knowing speaker's gender in advance (not always feasible)

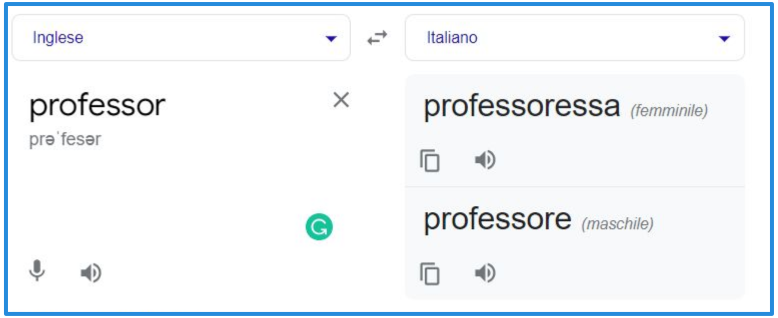
Mitigating Approaches

- **Gender Re-inflection** (Habash et al., 2019; Alhafni et al., 2020)
  - Scenario: 1-st person references to the speaker (e.g., *I am a student*)
  - Post-processing component re-inflecting into masculine/feminine forms
    - the component <u>always produces both forms</u> from an MT output
    - the <u>user chooses</u> the appropriate form

MITIGATING APPROACHES

- **Gender Re-inflection** (Habash et al., 2019; Alhafni et al., 2020)

→ double output implemented by **Google Translate**

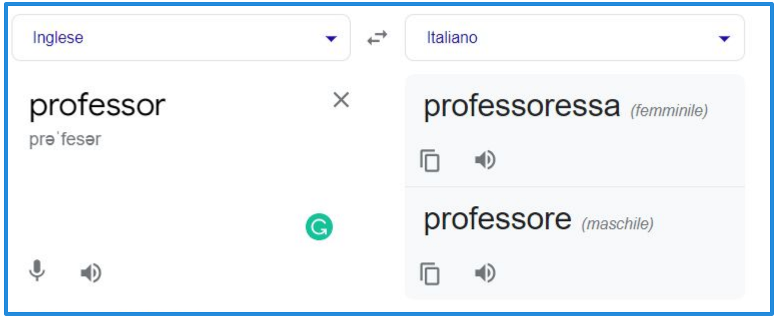


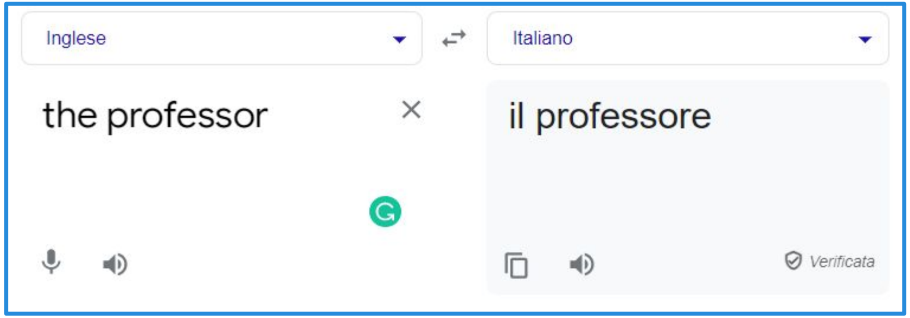
… only available for certain languages

MITIGATING APPROACHES

- **Gender Re-inflection** (Habash et al., 2019; Alhafni et al., 2020)

→ double output implemented by **Google Translate**





… only available for certain languages, mostly for single words

(1) NON-TEXTUAL MODALITIES

- Lack of studies on gender bias for e.g. **audiovisual translation** → different challenges and risks arise from not exclusively textual modalities
    - Audio-guided: ST represents a small niche (Costa-jussa' et al., 2020)
    - Image-guided: rely on images for gender disambiguation (Frank et al., 2018; Ive et al., 2019)

EN: A baseball player in a black shirt just tagged a player in a white shirt.