FONDAZIONE BRUNO KESSLER

HLT MACHINE TRANSLATION

UNIVERSITÀ DI TRENTO

Gender in Danger: Bias and Automatic Translation

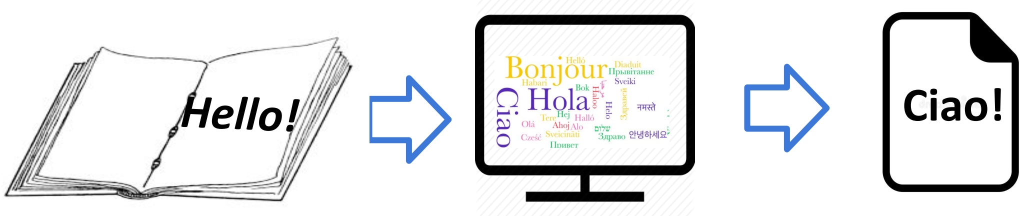
Translating Europe Workshop - TEW 2021, December 3rd

Beatrice Savoldi

beatrice.savoldi@unitn.it

MACHINE TRANSLATION

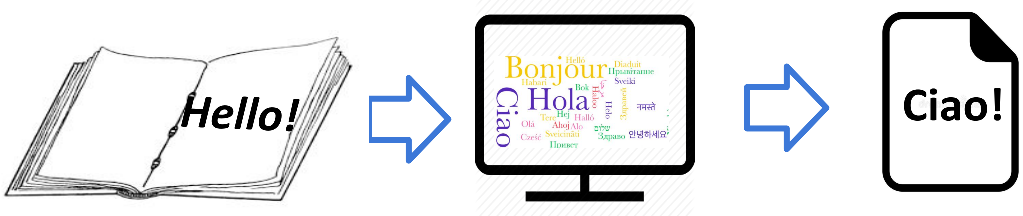
- Machine Translation (MT) popularity
  - **Neural Paradigm**: data-driven approach
  - Increasingly fluent and adequate translations
  - Improvements on syntax, lexicon, morphology (Bentivogli et al, 2016)



MACHINE TRANSLATION

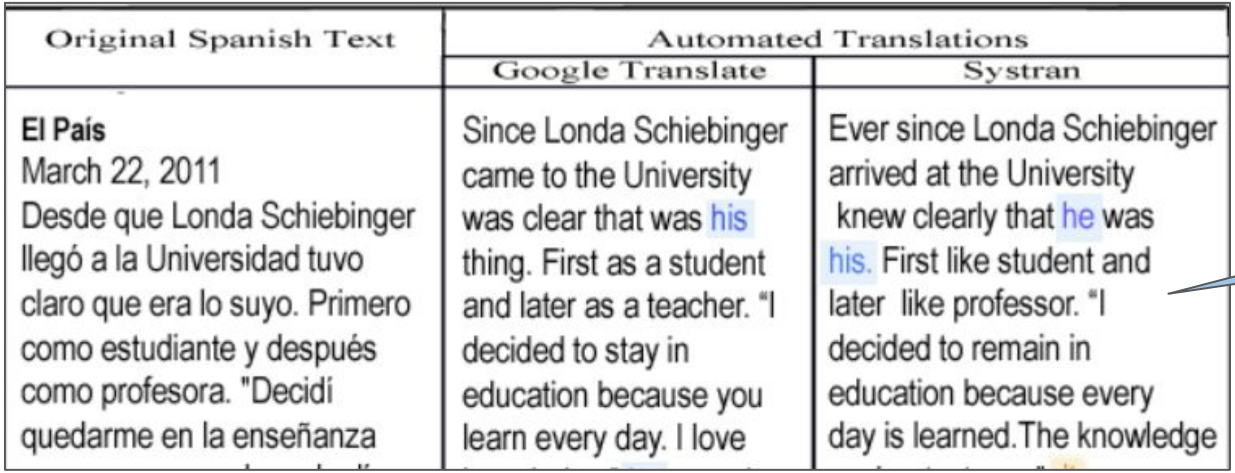
- Machine Translation (MT) popularity
  - **Neural Paradigm**: data-driven approach
  - Increasingly fluent and adequate translations
  - Improvements on syntax, lexicon, morphology (Bentivogli et al, 2016)

→ **but gender translation is an issue**



Gender Bias in MT

Analyses and real-world use prove that MT shows ***biased behaviours*** with respect to gender, leading to different types of ***harms***:



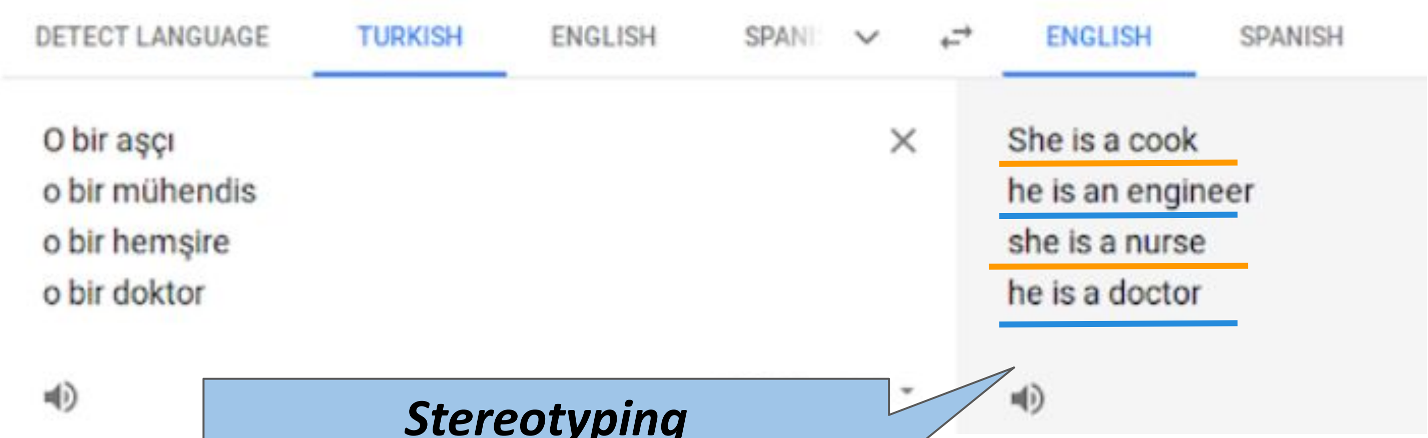
Original Spanish Text	Automated Translations	
	Google Translate	Systran
El País March 22, 2011 Desde que Londa Schiebinger llegó a la Universidad tuvo claro que era lo suyo. Primero como estudiante y después como profesora. "Decidí quedarme en la enseñanza	Since Londa Schiebinger came to the University 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

Under-representation

> "masculine skew"

GENDER BIAS IN MT

Analyses and real-world use prove that MT shows ***biased behaviours*** with respect to gender, leading to different types of ***harms***:



O bir aşçı
o bir mühendis
o bir hemşire
o bir doktor

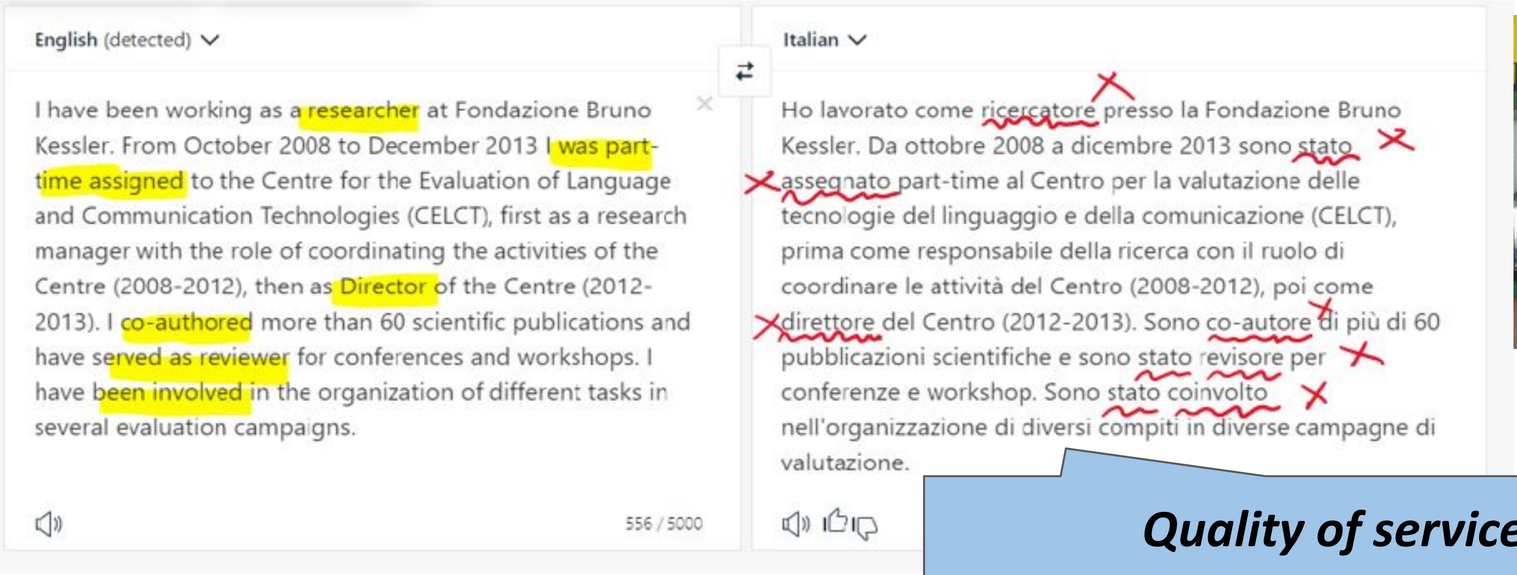
She is a cook
he is an engineer
she is a nurse
he is a doctor

Stereotyping

> stereotypical associations e.g. *pretty engineer as* feminine

GENDER BIAS IN MT

Analyses and real-world use prove that MT shows ***biased behaviours*** with respect to gender, leading to different types of ***harms***:



I have been working as a researcher at Fondazione Bruno Kessler. From October 2008 to December 2013 I was part-time assigned to the Centre for the Evaluation of Language and Communication Technologies (CELCT), first as a research manager with the role of coordinating the activities of the Centre (2008-2012), then as Director of the Centre (2012-2013). I co-authored more than 60 scientific publications and have served as reviewer for conferences and workshops. I have been involved in the organization of different tasks in several evaluation campaigns.

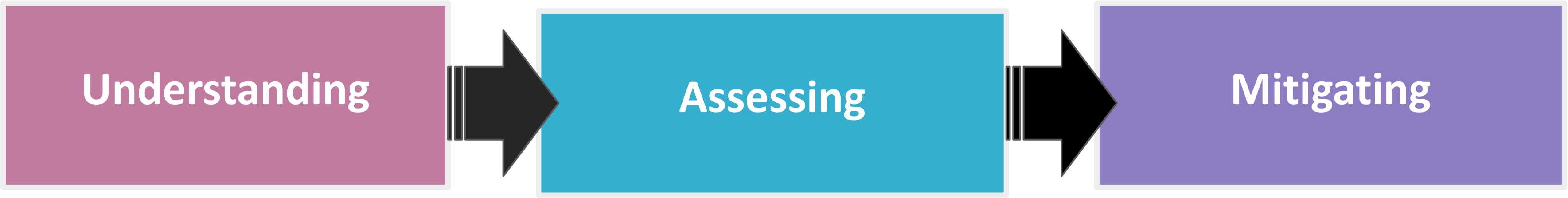
Ho lavorato come ricercatore presso la Fondazione Bruno Kessler. Da ottobre 2008 a dicembre 2013 sono stato assegnato part-time al Centro per la valutazione delle tecnologie del linguaggio e della comunicazione (CELCT), prima come responsabile della ricerca con il ruolo di coordinare le attività del Centro (2008-2012), poi come direttore del Centro (2012-2013). Sono co-autore di più di 60 pubblicazioni scientifiche e sono stato revisore per conferenze e workshop. Sono stato coinvolto nell'organizzazione di diversi compiti in diverse campagne di valutazione.

[Luisa Bentivogli, short bio](#)

Quality of service

> worse performance for women

Gender Bias in MT: Outline



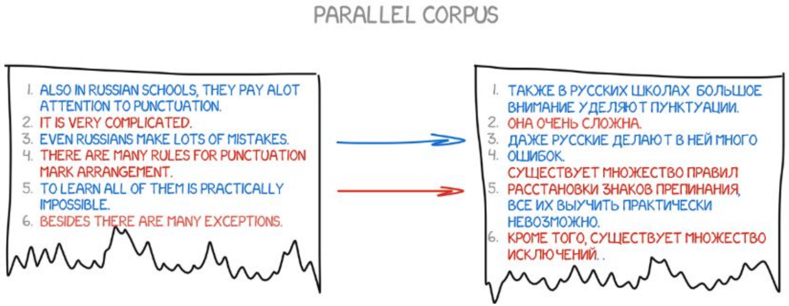
Understanding

Assessing

Mitigating

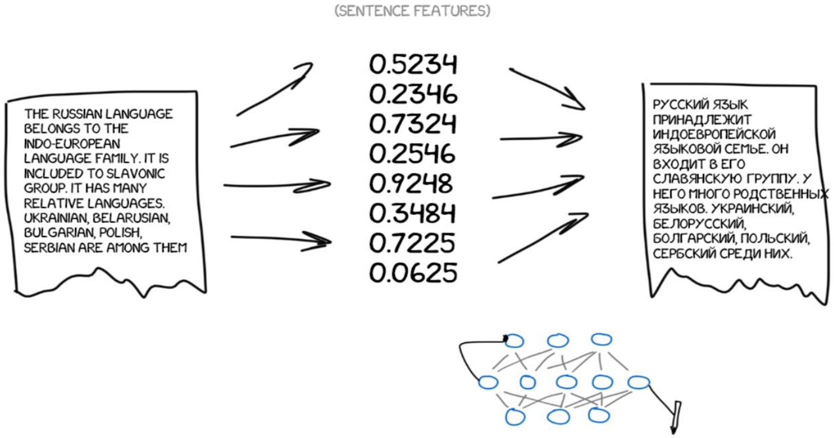
WHAT ARE THE SOURCES OF BIAS?

MT models are fed with (lots) of *parallel data* and learn patterns across languages from such **training data**



1. ALSO IN RUSSIAN SCHOOLS, THEY PAY ALOT ATTENTION TO PUNCTUATION.
2. IT IS VERY COMPLICATED.
3. EVEN RUSSIANS MAKE LOTS OF MISTAKES.
4. THERE ARE MANY RULES FOR PUNCTUATION MARK ARRANGEMENT.
5. TO LEARN ALL OF THEM IS PRACTICALLY IMPOSSIBLE.
6. BESIDES THERE ARE MANY EXCEPTIONS.

1. ТАКЖЕ В РУССКИХ ШКОЛАХ БОЛЬШОЕ ВНИМАНИЕ УДЕЛЯЮТ ПУНКТУАЦИИ.
2. ОНА ОЧЕНЬ СЛОЖНА.
3. ДАЖЕ РУССКИЕ ДЕЛАЮТ В НЕЙ МНОГО ОШИБОК.
4. СУЩЕСТВУЕТ МНОЖЕСТВО ПРАВИЛ РАССТАНОВКИ ЗНАКОВ ПРЕПИНАНИЯ.
5. ВСЕ ИХ ВЫУЧИТЬ ПРАКТИЧЕСКИ НЕВОЗМОЖНО.
6. КРОМЕ ТОГО, СУЩЕСТВУЕТ МНОЖЕСТВО ИСКЛЮЧЕНИЙ. .



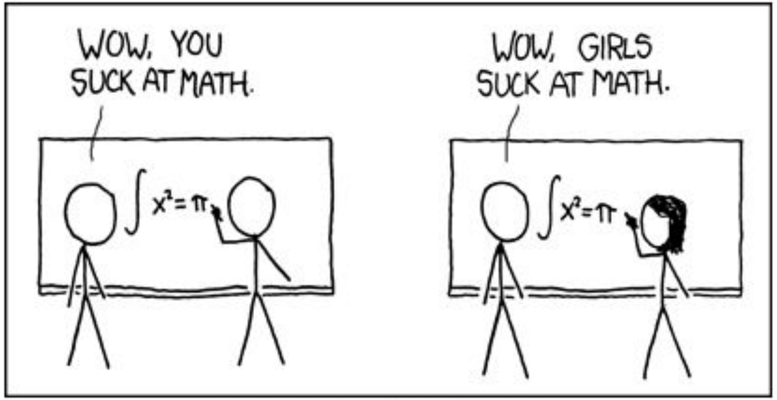
THE RUSSIAN LANGUAGE BELONGS TO THE INDO-EUROPEAN LANGUAGE FAMILY. IT IS INCLUDED TO SLAVONIC GROUP. IT HAS MANY RELATIVE LANGUAGES. UKRAINIAN, BELARUSIAN, BULGARIAN, POLISH, SERBIAN ARE AMONG THEM

РУССКИЙ ЯЗЫК ПРИНАДЛЕЖИТ ИНДОЕВРОПЕЙСКОЙ ЯЗЫКОВОЙ СЕМЬЕ. ОН ВХОДИТ В ЕГО СЛАВЯНСКУЮ ГРУППУ. У НЕГО МНОГО РОДСТВЕННЫХ ЯЗЫКОВ. УКРАИНСКИЙ, БЕЛОРУССКИЙ, БОЛГАРСКИЙ, ПОЛЬСКИЙ, СЕРБСКИЙ СРЕДИ НИХ.

Image credit: Vasily Zubarev

WHAT ARE THE SOURCES OF BIAS?

- … systems’ ability in learning patterns turns into weakness as training data can encode gender disparities



Title text: It's pi plus C, of course.

xkcd.com

WHAT ARE THE SOURCES OF BIAS?

- … systems' ability in learning patterns turns into weakness as training data can encode gender disparities

BUT

> ***Training data bias* as an overloaded term** (Suresh and Guttard, 2019)

> Different sources of bias (Friedman & Nissenbaum, 1996)

What are the Sources of Bias?

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

  - Europarl Corpus (Kohen, 2005)
    - 30% sentences uttered by women (Vanmassenhove et al., 2018)
    - 40% highest peak of Women in the EU Parliament (Women infographics)

→ glass ceiling that has hampered women's access to political positions

WHAT ARE THE SOURCES OF BIAS?

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

  - ❖ Europarl Corpus (Kohen, 2005)
  - ❖ Social Connotations and Language use
    - ■ explicit female markings for doctor (female, woman or lady doctor) (Romaine, 2001)

→ qualitative asymmetries regarding how linguistic expressions are connoted, deployed and perceived

What are the Sources of Bias?

- **Technical bias**: due to technical constraints and decisions
  - Data curation/data annotation
    - how are data processed and annotated? (Wagner et al., 2016)
  - Models design
    - *algorithmic bias* that leads under-represented feminine forms to further decrease in an MT output (Vanmassenhove et al., 2020,2021)
  - Evaluation procedure
    - gender asymmetries in test data reward biased predictions (Sun et al., 2019)
    - inadequate choice of evaluation metrics (e.g. aggregate measures can hide subgroup underperformance) (Mitchell et al., 2018)

ASSESSING GENDER BIAS

….Traditional metrics and Generic Test sets are unsuitable

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

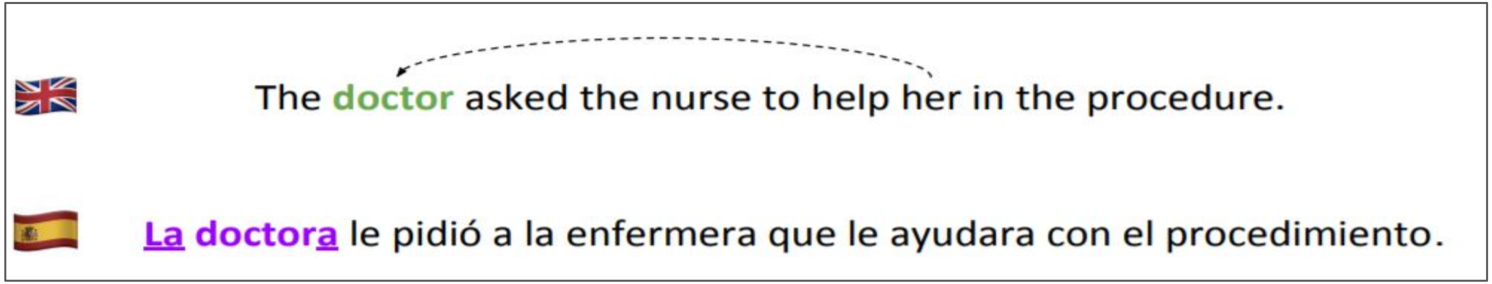
⟶ isolate gender as a variable

⟶ 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, easy to overfit



🇬🇧 The doctor asked the nurse to help her in the procedure.

🇪🇸 La doctora le pidió a la enfermera que le ayudara con el procedimiento.

WinoMT (Stanovsky *et al.*, 2019)

GBET BENCHMARKS

- **Challenge datasets**
  (Prates *et al.*, 2018; Cho *et al.*, 2019; Escudé Font & Costa-jussà, 2019; Stanovsky *et al.*, 2019)
- **Natural datasets**
  (Habash *et al.*, 2019; Bentivogli et al., 2020)

→ selected and annotated gender instances from conversational language

→ more authentic conditions but.. treat all gendered words equally

Src	She'd get together 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)**

GBET BENCHMARKS

- Challenge datasets
  (Prates *et al.*, 2018; Cho *et al.*, 2019; Escudé Font & Costa-jussà, 2019; Stanovsky *et al.*, 2019)
- Natural datasets
  (Habash *et al.*, 2019; Bentivogli et al., 2020)

>> Benchmarks are formalizations and respond to different conceptualization of bias (Barocas et al., 2019)

>> Relevant to monitor system’s behaviour and mitigating strategies

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 a 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

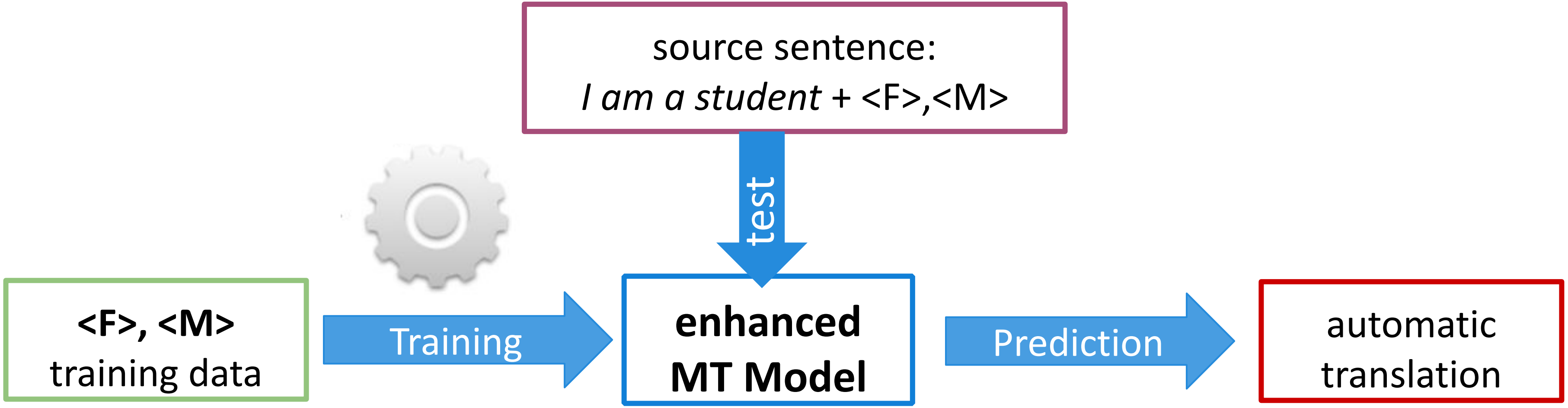
- **Based on counterfactual data augmentation** (CDA) (Saunders & Byrne, 2020)
  - CDA: creation of synthetic sentences with balanced F/M representation
  - MT model is fine-tuned on such a parallel set

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., 2018)
  - 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)

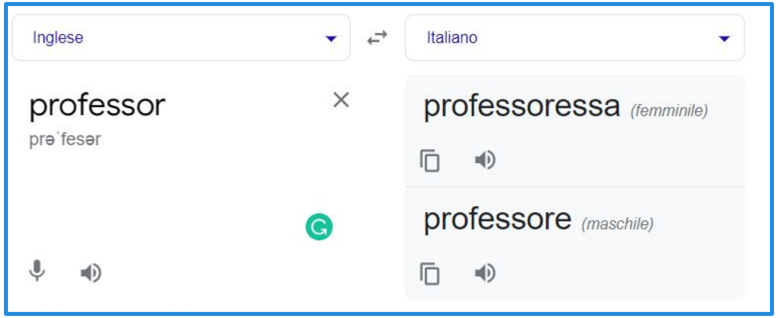
MITIGATING APPROACHES

- **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 always produces both forms from an MT output
    - the user chooses the appropriate form

MITIGATING APPROACHES

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

→ double output implemented by **Google Translate**



professor
prəˈfesər

professoressa *(femminile)*

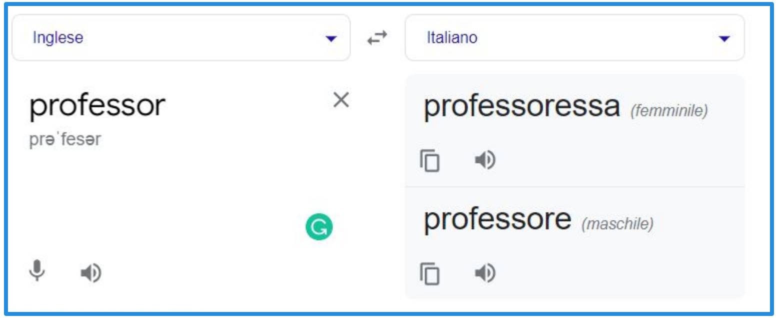
professore *(maschile)*

… only available for certain languages

MITIGATING APPROACHES

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

→ double output implemented by **Google Translate**

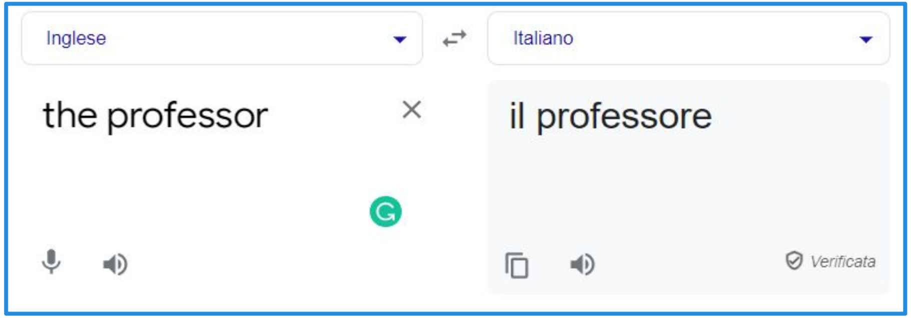


Inglese ⇄ Italiano

professor
prəˈfesər

professoressa *(femminile)*

professore *(maschile)*



Inglese ⇄ Italiano

the professor

il professore

… only available for certain languages, mostly for single words

Mitigating Approaches

- No conclusive state-of-the-art method for mitigating bias
  - Response to specific aspects of the problem with *modular solutions*

- **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?*

(1) GENDER-NEUTRAL LANGUAGE

Except for one work in MT (Saunders et al., 2020), work on gender bias has focused on binary masculine/feminine dichotomy

- Indirect Non-binary Language: overcomes gender specifications
    - using e.g. *humankind vs.* mankind; *service* vs. waiter and waitress
    - endorsed for many official documents (Papadimoulis, 2018)
    - a challenging goal for grammatical gender languages

TO CONCLUDE: *where to?*

(2) HUMAN-IN-THE-LOOP

Language technologies are built **for people…**
⟶ but to date evaluations on gender bias in MT are restricted to lab tests

- Studies relying on participatory design and HCI approaches(Liebling et al., 2020, Cercas Curry et al., 2020)
- Consider different MT users… Translators included (Ragni & Vieira, 2020)

TO CONCLUDE: *where to?*

(2) 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

To Conclude: *where to?*

(2) 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

Thanks for listening!

@fbk_mt @BeatriceSavoldi

beatrice.savoldi@unitn.it