Gender Bias in Artificial Intelligence: Exploring the impacts of stereotypes on English-Vietnamese Machine Translation

Diep Vu
Kristina Striegnitz, Advisor

ABSTRACT:

Artificial Intelligence (AI) is increasingly influencing people's opinion and behavior in daily life. Gender bias in AI, especially in Machine Translation has been a growing concern since the over-representation of men in the design of these technologies might gradually undermine decades of progress toward gender equality.

In my project, the main goal is to investigate gender bias in language models for translating from **English** to a language with gender neutral pronouns (**Vietnamese**). I am developing a collection of test sentences to probe a translation model for gender bias. First, I will use this test set to evaluate Google Translate.

Then, I will use a state-of-the-art Neural Network approach to train from a multi-language model (**Helsinki-NLP**) on a standard translation dataset from English-Vietnamese, and I will evaluate this model using my test sentences. Next, I will use a bias mitigation technique where I balance out the gendered words in the training dataset by swapping them and retrain and evaluate again.

The results of this project aim to demonstrate the need to augment current statistical translation tools with debiasing techniques. There is also the need to look further into using a bigger dataset with fewer stereotypes, which can be hard to achieve since language dataset always reflect its country's social context.

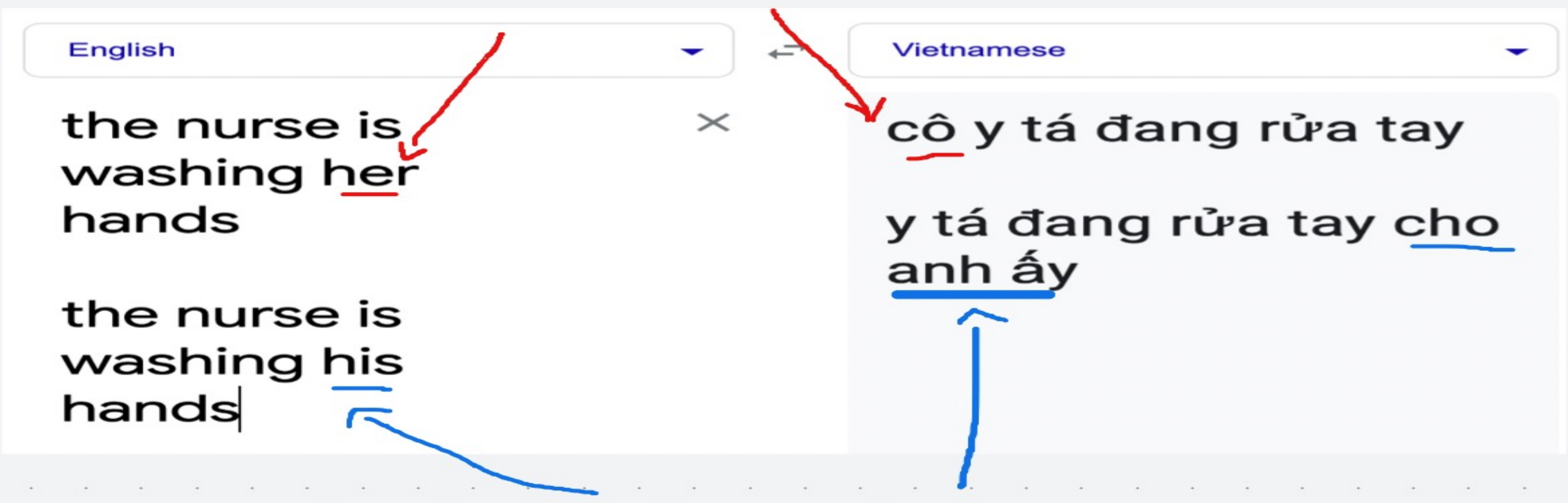
BACKGROUND:

- Machine Translation (MT) is the task of automatically converting text in source language to text in target language.
- State-of-the-art neral network model (**seq2seq** and **Transformer**) are trained on large collection of text from the web news, books, etc.
- Text in the source language might contain bias, e.g.
  - Web news: only **9%** of the occupation "**technician**", "**accountant**", "**doctor**" associate with **female**
  - Bureau Labor Statistics (BLS): with more than 5 times (~**40% female**) [1]

=> The system learns biases

RESULT:

prefix **"cô"** for female pronoun



the nurse is washing her hands

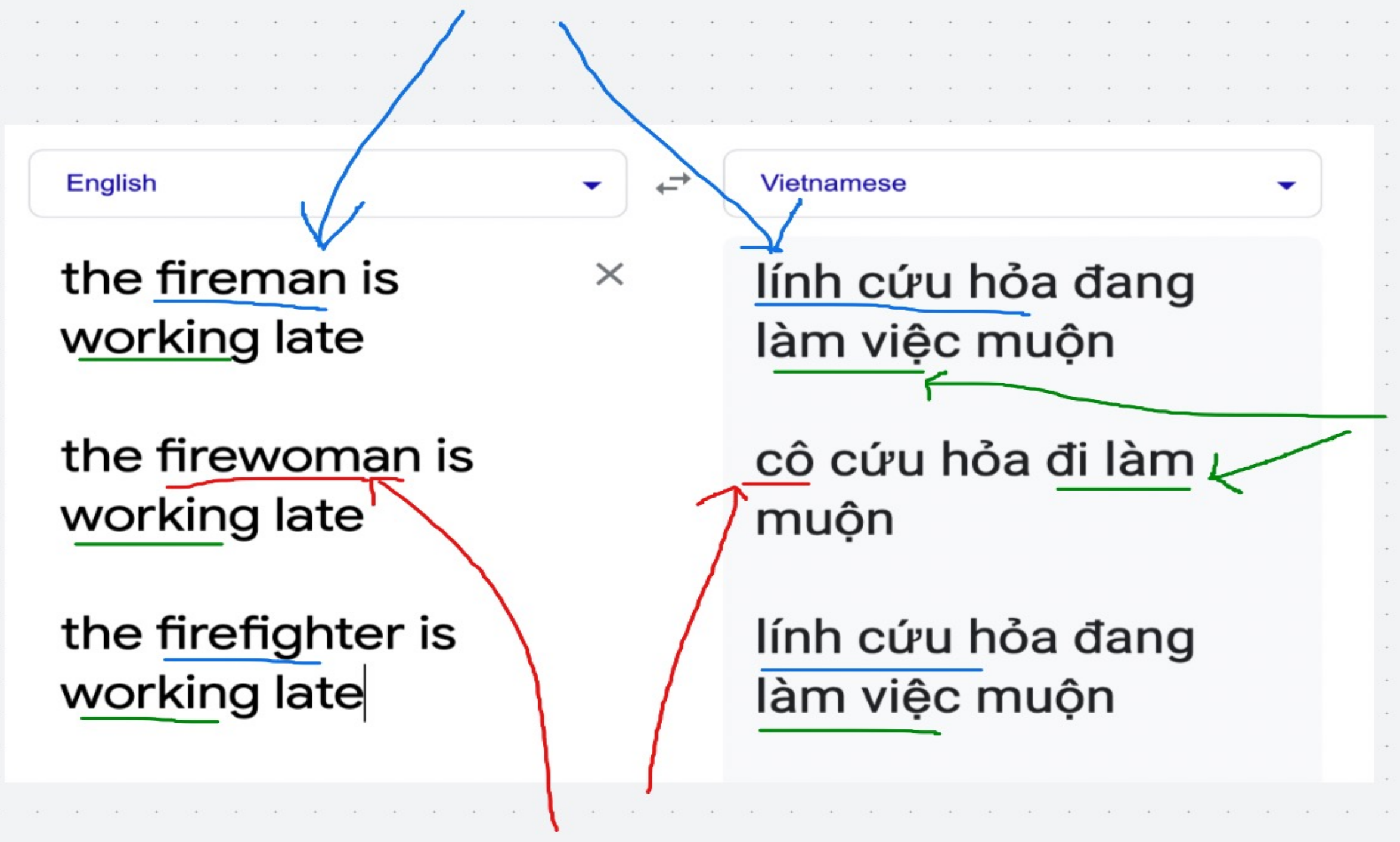
the nurse is washing his hands

cô y tá đang rửa tay

y tá đang rửa tay cho anh ấy

the translation gets wrong meaning: **"cho anh ấy"** means washing for other male's hands, not the nurse's hands

the system generates same translation for **fireman** and **firefighter**, by default someone working against the fire is **male**



the fireman is working late

the firewoman is working late

the firefighter is working late

lính cứu hỏa đang làm việc muộn

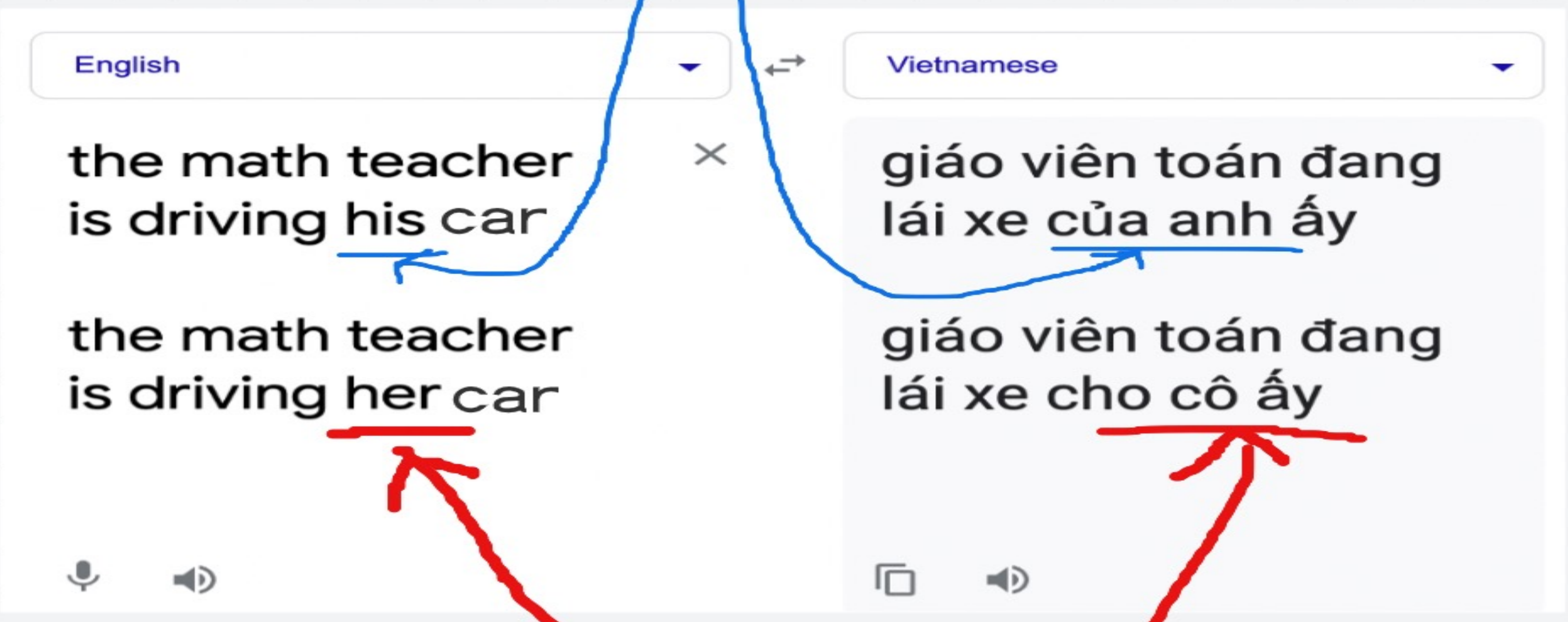
cô cứu hỏa đi làm muộn

lính cứu hỏa đang làm việc muộn

the system gets **wrong** translation and **negative** meaning for the sentence with **female** version, **"đi làm muộn"** means **"late for work"**

prefix "cô" for female pronoun

the system gets right translation: **"của anh ấy"** means the math teacher's car who is a **male**



the math teacher is driving his car

the math teacher is driving her car

giáo viên toán đang lái xe của anh ấy

giáo viên toán đang lái xe cho cô ấy

the translation gets wrong meaning: **"cho cô ấy"** means driving with other female's car, not the math teacher's car

- Results show that the system automatically puts the prefix **"nữ"** or **"cô"** for the **female** translation version and the **default/gender neutral version** is the same as **male**.
- These translations sometimes lose the original meaning, sound unnatural, **negative** towards **female** and tend to assign **male** pronouns with **gender neutral** texts.

TESTING FOR BIASES IN MT SYSTEM

- Overall I wrote **6 testing templates** and plugged in **60** different names for **$OCCUPATION** from the BLS to generate total **6*60*2 = 720 testing sentences** for **female** and **male** versions. Here are the templates:

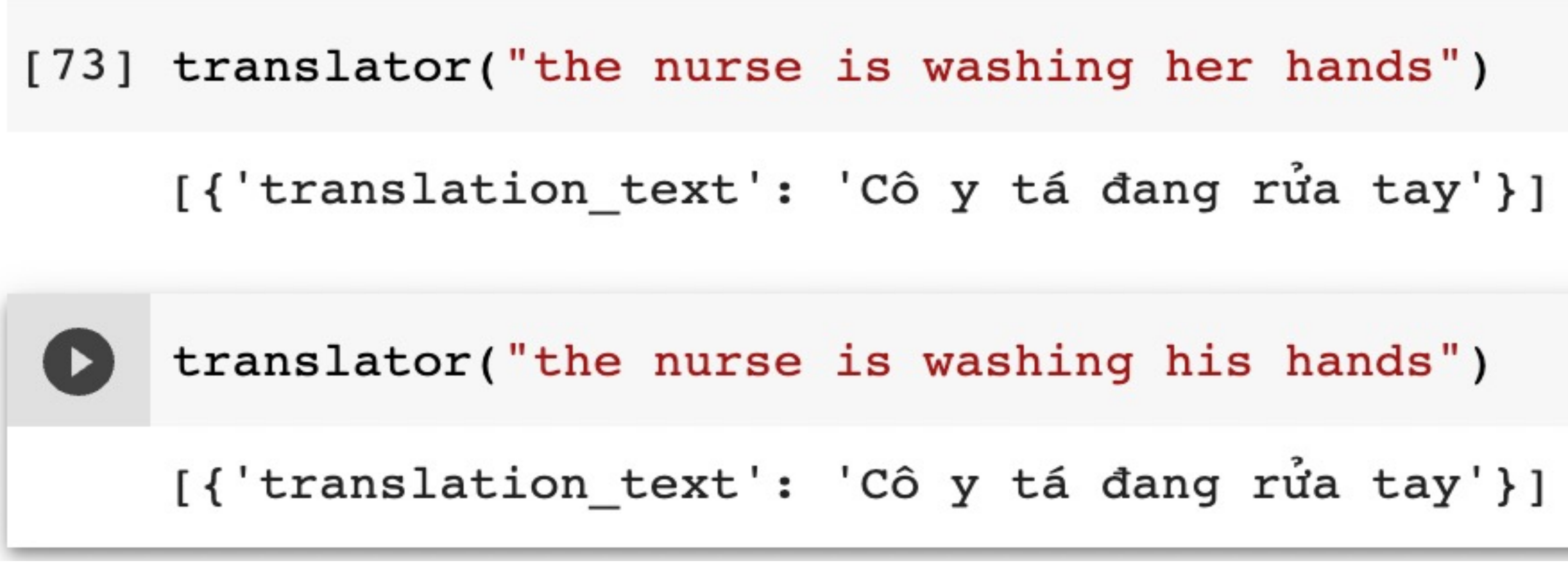
- the **$OCCUPATION** is washing **$PRONOUN** hands.
  the **$OCCUPATION** is **$CHARACTERISTICS.**
  the **$OCCUPATION** is looking at **$PRONOUN** in the mirror.
  the **$OCCUPATION** is working late.
  the **$OCCUPATION** is putting on **$PRONOUN** gown.
  the **$OCCUPATION** is driving **$PRONOUN** car.

References

[1] Zhao, Jieyu, et al. "Gender bias in coreference resolution: Evaluation and debiasing methods." *arXiv preprint arXiv:1804.06876* (2018).

CURRENT/FUTURE WORK:

- Use the Neural Network approach to train a multi-language model (**Helsinki-NLP**) and the model fails to detect the gender differences by putting **female** prefix **"cô"** or **male** version sentence.



```
[73] translator("the nurse is washing her hands")

     [{'translation_text': 'Cô y tá đang rửa tay'}]

     translator("the nurse is washing his hands")

     [{'translation_text': 'Cô y tá đang rửa tay'}]
```

- Apply the **debiasing** technique by **balancing out** the **gendered pronouns** in the training dataset, get bigger dataset with fewer stereotypes (big challenge!)
- Re-evaluate the model compared to Google Translate