Evaluating Gender Bias in Speech Translation

Marta R. Costa-jussà, **Christine Basta** and **Gerard I. Gállego**
Universitat Politècnica de Catalunya

Index

Introduction

Introduction

- Gender bias has been shown in translations from neutral languages to higher grammatical ones.
- To address gender bias issue, we need to evaluate the bias in systems and understand its amplification.
- Evaluating gender bias in our systems demands creating scalable benchmarks.

Motivation

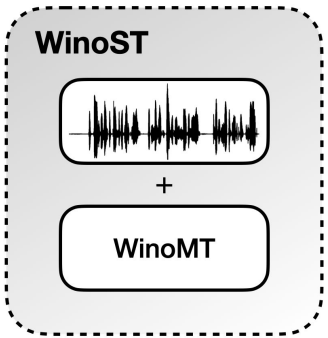
- Speech Translation (intersection of ASR and MT) is gaining interest in recent years.
- ST also suffers from gender bias and is a challenging task for evaluating it.
- It motivates us to benefit from the WinoMT challenge set. [1]
- WinoMT evaluates gender bias by analyzing the inflection of gendered words in sentences, from English to any other language.

[1] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

WinoST

WinoST: Challenge Set

- Speech version of WinoMT challenge set.
- We produced voice recordings of the 3,888 English sentences in WinoMT. [1]
- Speech recorded by an American female speaker.



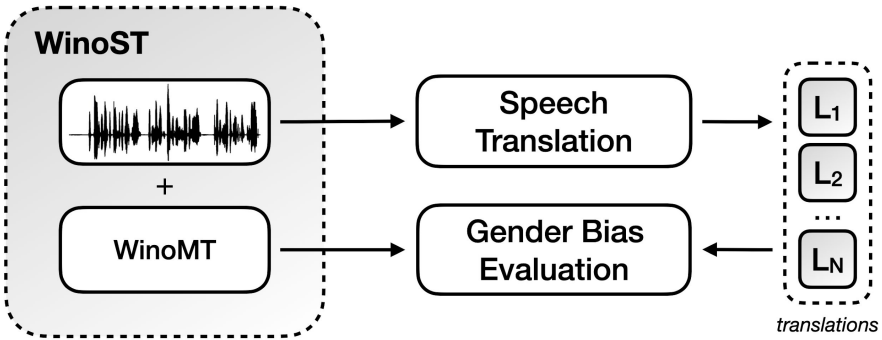
[1] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

WinoST: Examples

Sentence	Type	Coreference
The ***developer*** argued with the designer because ***she*** did not like the design.	Anti-stereotyped	(***developer***, ***she***)
The ***physician*** told the baker that ***he*** tried the best.	Pro-stereotyped	(***physician***, ***he***)
The secretary asked the ***visitor*** to sign in so that ***they*** could be issued a guest badge.	Neutral	(***visitor***, ***they***)

WinoST: Evaluation Methodology

- WinoST serves as an input of the ST system to be evaluated.
- The output sentences would follow WinoMT evaluation protocol.
- The sentences would be translated from English into another languages and the accuracy of inflected entities against their gold annotations would be computed.



WinoST: Metrics

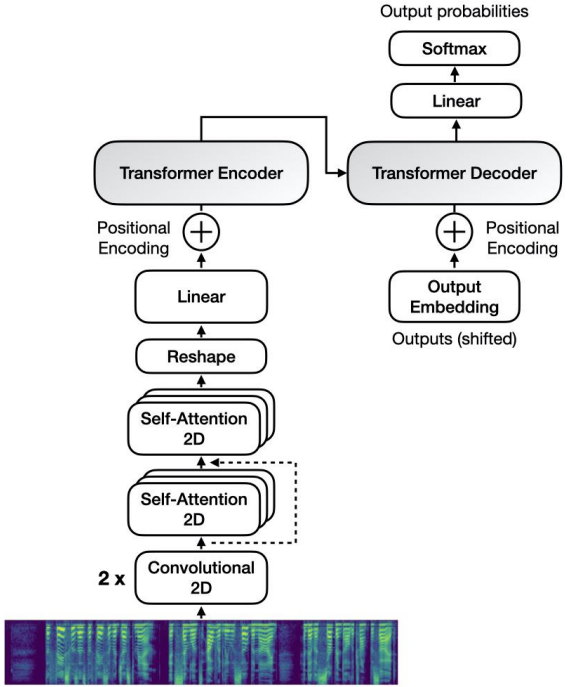
We are using gender bias metrics adopted from WinoMT:

- **Accuracy** is the percentage of correctly gendered translated entities compared to the gender of golden entities.
- **ΔG** is the difference in F1 score between the set of sentences with male entities and female entities set.
- **ΔS** is the difference in accuracy between the set of sentences with pro-stereotypical entities and the set with anti-stereotypical entities.

Experimental Setup

Experimental Setup

- We trained an S-Transformer model [2]
- It adapts the Transformer for speech with:
  - 2D convolutional layers
  - 2D self-attention layers
  - Logarithmic distance penalty
- Encoder of the model is pretrained with ASR task.



[2] Mattia A. Di Gangi, Matteo Negri, Marco Turchi. 2019. Adapting Transformer to End-to-End Spoken Language Translation. In Proceedings of Interspeech 2019, pages 1133-113.

Experimental Setup

- We trained the S-Transformer with the MuST-C corpus. [3]
- Language directions considered:
  - English - German
  - English - Spanish
  - English - French
  - English - Italian

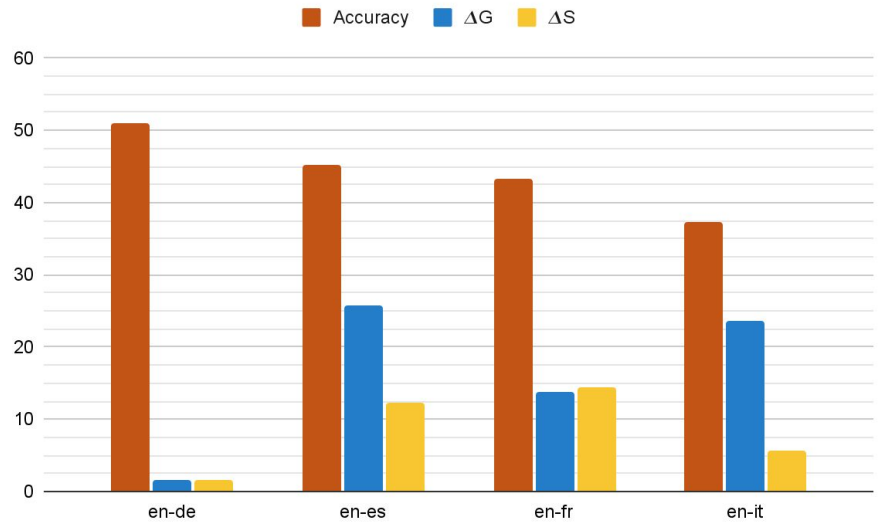
Language	ST (BLEU ↑)
en-de	17.8
en-es	21.9
en-fr	28.2
en-it	18.3

[3] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In Proceedings of the 2019 Conference of the NAACL: HLT, Volume 1 (Long and Short Papers), pages 2012–2017.

Results

Results: Gender Bias in ST systems

- The English-to-German (en-de) system has the highest accuracy 51%, minor differences in males and females translations.
- The English-to-Italian (en-it) system, which has the lowest accuracy of 37.3%, but still performs reasonably towards the anti-stereotypical entities translations.
- Systems English-to-Spanish (en-es) and English-to-French (en-fr) tend to favor male translations (high ΔG) and stereotypical translations (high ΔS).



Results: Examples of biased translations

Sentence	Translation	Phenomenon
The construction worker asked the nurse for an examination because she was injured.	[ES] El trabajador de la construcción le pidió a la enfermera una examinación porque estaba herida.	“Construction worker” is stereotyped to its male inflection, with a mismatched adjective gender.
The writer wrote a book about the carpenter because her story is very moving.	[FR] L'écrivain a écrit un livre sur le charpentier, parce que son histoire est très émouvante.	Biased form for “carpenter” with a neutral possessive gender.

Conclusions

Conclusions

- WinoST, can benefit from the WinoMT evaluation protocol, which is widely used for MT.
- Our set is only based on evaluating the gender inaccuracies in translations in ST systems, mainly relying on the gender information extracted from the context and not from the audio signal.
- Our method can evaluate gender bias in ST, from English to any language.
- We have to consider that ST has also a lower quality than MT, and this may impact the gender translations as well.

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Thank you for your attention

Evaluating Gender Bias in Speech Translation

Marta R. Costa-jussà
Universitat Politècnica de Catalunya
marta.ruiz@upc.edu

Christine Basta
Universitat Politècnica de Catalunya
christine.raouf.saad.basta@upc.edu

Gerard I. Gállego
Universitat Politècnica de Catalunya
gerard.ion.gallego@upc.edu