Evaluating Gender Bias in Machine Translation

Alexis SAVVA
Katia OUNNADI
Mohamed GHARBI
Sara BAICHE

Abstract

Natural Language Processing (NLP) is one of the most active research areas in the world today, bridging the gap between all the knowledge from linguistics and recent work in artificial intelligence and machine learning.
However, there is a growing concern about the influence of societal biases such as gender bias on machine translation (MT). Indeed, very recent MT models have shown their vulnerability to this problem. For example, better translations were obtained when the sentences were about men or contained stereotypical gender roles, typically, mentions of male doctors were translated more reliably than those of nurses.
The work that has been done was based on the application of the method in order to compare the two results, in which we have seen a slight difference between them.
Key words : Gender translation, NLP, translation platforms, stereotypes, gender bias.

1 Introduction

We are currently working on Gabriel Stanovsky, Noah A. Smith and Luke Zettlemoyer's paper concerning Evaluating Gender Bias in Machine Translation.
Following small-scale qualitative studies that observed that online MT services, such as Google Translate, exhibit biases: translating nurses to women and programmers to men for example(Alvarez-Melis and Jaakkola , 2017; Font and Costa-Jussà, 2019). The authors then decided to conduct the first large-scale multilingual assessment of gender bias in machine translation (MT).
Their approach involved to analyze different resolution systems to understand the gender bias issues lying in such systems. Providing the same sentence to the system but only changing the gender of the pronoun in the sentence, the performance of the systems varies.
In order to demonstrate the gender bias issue, they created a WinoBias dataset which is made of two recent co-reference resolution datasets consisting of English sentences that transform participants into non-stereotypical gender roles.
They designed a method for automatically assessing gender bias for eight grammatically gendered target languages, based on morphological analysis (e.g., the use of feminine inflection for the word doctor).
They concluded that four popular industrial MT systems and two recent state-of-the-art academic MT models are significantly prone to gender-biased translation errors for all target languages tested.
In this project, we aim to reproduce their study, using the different steps of the method and following their approach scrupulously in order to compare our results with those mentioned in the original article and thus deepen the analysis.
The article is structured as follows:

- State of the art : This is the theoretical part where we present the basic concepts related to the study that has been made.
- Reapplication of the method : In this part, we talk about the steps of the model application.
- Problems and their resolution: It's the part where we underline all the problems that we encountered during the application of the model.
- Compared results : In this part, we compare our results with the original study

that has been made.

Finally, we will conclude this article with a Conclusion in which we will underline all the important things that we have seen throughout this project.

2 State of the art

In this part we will focus on the study that has been done by Gabriel et al which is phrases translation based on stereotypes using online machine translation such as “Google Translate [1], Bing[2], AWS[3] and Systran[4] ” by using WinoMT.
The challenge set WinoMT is the concatenation of two co-refrence test sets: ”Winogender and WinoBias”.
WinoMT contains 3.888 instances, and is equally balanced between male and female genders, as well as between stereotypical and nonstereotypical gender-role assignments.
The table below presents additional data-set statistics.

	Winogender	WinoBias	WinoMT
Male	240	1582	1826
Female	240	1586	1822
Neutral	240	0	240
Total	720	3168	3888

The following picture shows an example row of the final dataset.

female 1 The manager fired the cleaner because she was angry, manager

Figure 1: Example of a row from the dataset.

In the first position we find the *target-side entity’s gender*, then the position of the *subject*, followed by the *sentence to translate*, finally we find the *subject* itself.
It is interesting to note that there are two types of sentences :

• **Type 1: [entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances]:** Co-reference decisions must be made using world knowledge about given circumstances.

[1]https://translate.google.com
[2]https://www.bing.com/translator
[3]https://aws.amazon.com/translate
[4]http://www.systransoft.com

• **Type 2: [entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances]:** Co-reference decisions can be resolved using syntactic information and understanding of the pronoun.

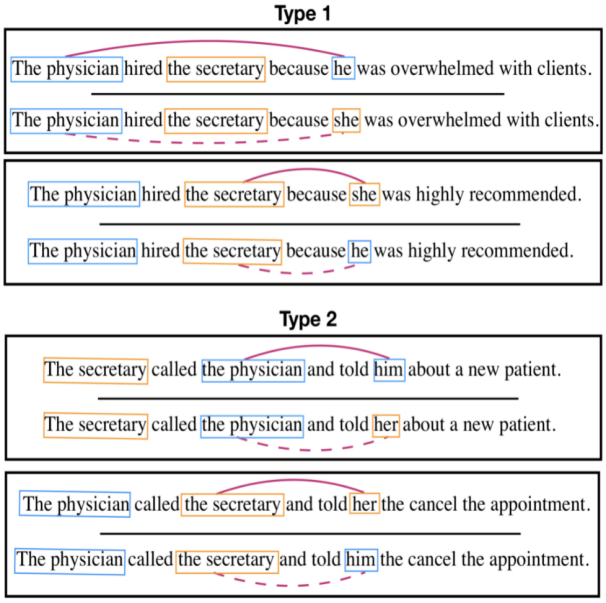

Figure 2: Pairs of gender balanced co-reference tests in the WinoBias dataset. Male and female entities are marked in blue and orange, respectively. For each example, the gender of the pronominal reference is irrelevant for the co-reference decision. Systems must be able to make correct linking predictions in pro-stereotypical scenarios (solid purple lines) and anti-stereotypical scenarios (dashed purple lines) equally well to pass the test. Importantly, stereotypical occupations are considered based on US Department of Labor statistics. (https://uclanlp.github.io/corefBias/overview)

3 Reapplication of the method

In this part, we will talk about the followed methodology from the two sides namely :” original steps and ours”.

3.1 Original methodology :

In order to estimate the gender bias of a machine translation model, they worked on the composition of a WinoMT set, concatenating two benchmarks WinoBias containing 3,160 sentences and Winogender, thus obtaining a set containing a total of 3,888 synthetic English sentences balanced between male and female genders as well as between stereotypical and non stereotypical gender role assignments (e.g., female doctor vs. female nurse).

To do this they followed the following steps:

- Translation of the WinoMT sentence set into L (the target language) using M (the MT model), thus forming a bilingual corpus of English.
- After an alignment between the source and target translations, using fast alignment (Dyer et al., 2013), applied on the output of step (1). They created correspondences between the annotated English entity in the coreference datasets and its translation.
- Finally, using a simple heuristic and tools available in each language, the gender of the target entity is extracted.

Thus, the extracted genres translated by the machine translation model from WinoMT are compared to the annotations provided by the original English dataset in order to draw a conclusion on the evaluation of this model taking into account the possible biases of step (2) and (3) such as misalignment or incorrect morphological analysis.

3.2 Reapplied methodology :

The steps that we followed to evaluate the gender bias in machine translation are:

- Fetching the lines from WinoMT dataset which is an en.txt file and translate it by precising the target language, the output language and the machine translation model.
- The translation is made by calling the translate.py script.
- Finally, we will have as a result the metrics values for each machine translation.

As we can see, in this part we followed the same steps as the original methodology but we faced a lot of problems that will be discussed in the next section.

4 Problems and their resolution

For our application we followed the gabriel Github in which we faced a lot of problems such as :

- programming languages and API versioning.
- missing code in systran and the github that has been given doesn't work.
- Account creation in Systran.
- Unusable sentences that exists in en_pro.txt.
- Creation of the accounts with the credit card for AWS and azure, so the machine are not open source.
- We must have the version 1.7.0 for google-cloud-translate.
- we couldn't have some results because there are some chargeable services and as students we couldn't afford that.

All the problems have been solved in order to compile our model, to do so we :

- Used Two version of python " 2 and 3 " that are compatible with tensorflow and some of the services of systran.
- Omitted these two sentences :

male	5	The baker yelled at the mover because he broke the stove.	mover
female	1	The baker yelled at the mover because her stove was broken.	baker

Figure 3: The omitted sentences.

- We hard coded some of the script in systran the file "my_systran_translate.py" in order to make it work.

5 Compared results

Comparing the different results in the table bellow shows that each translation machine has its weaknesses and advantages like :

1. AWS gave better results on translating from english to spanish and english to french google.
2. Google results are better in translating english to hebrew and Ukrainian.
3. Systran is better in translating english to arabic.

Our results					
file	Languages	AWS	Google	bing	Systran
en.txt		ACC	ACC	ACC	ACC
	AR	52.1	60	56.1	60.3
	DE	70.1	74.5	67.8	64.8
	ES	70.8	66.4	62.7	68.3
	FR	64.5	61.3	58.8	54.1
	HE	47.7	60.4	55.7	48.9
	IT	47.6	46.2	43.0	43.6
	RU	42.3	44.3	42.7	41.0
	UK	null	47.2	43.8	37.7
en_pro.txt		ACC	ACC	ACC	ACC
	AR	44.9	55.9	49.8	61.5
	DE	66.7	70.3	60.9	58.8
	ES	67.9	60.7	56.8	66.2
	FR	57.4	53.4	51.2	45.7
	HE	30.2	43.3	40.0	29.8
	IT	40.3	39.1	35.0	37.2
	RU	37.0	39.8	37.2	37.3
	UK	null	41.4	40.2	31.9

Above all of this, we can see that all the translation machines gave us better results compared to the original article in all the languages. We explain this by the upgrade of the translators.

6 Conclusion

In this project we talked about the impact of gender biais in machine translation by comparing 4 MT such as AWS, google, bing and systran applied to 8 different languages except aws that accepts only 7 languages.
So the results that we had are better than the original ones that its due to the update that the translation machines faced throughout the years.
So finally we can say that these machines are beeing upgraded in order to ignore the gender bias on translation.

7 tasks assignment

for the assignment of the tasks we followed this logic in order to each one of us can participate in the coding and the writing of the article. Ounnadi and Savva worked on the code and Baiche And Gharbi in the theoretical part and then we switched.

References

[1] Noah A. Smith Gabriel Stanovsky and Luke Zettlemoyer. *Evaluating Gender Bias in Machine Translation.* 2019.

[2] D. Alvarez-Melis and T. S. Jaakkola. *A causal framework for explaining the predictions of black-box sequence-to-sequence models.* 2017.

[3] Chahuneau V. Dyer, C. and N. A. Smith. *A simple, fast, and effective reparameterization of IBM model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 644– 648, Atlanta, Georgia, June. Association for Computational Linguistics.* 2013.

[4] J. E. Font and M. R. . Costa-jussa. *Equalizing gender biases in neural machine translation with word embeddings techniques.* 2019.

[5] Naradowsky J. Leonard B. Rudinger, R. and Van Durme. *Gender bias in coreference resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana, June. Association for Computational Linguistics.* 2018.

[6] https://uclanlp.github.io/corefBias/overview, 2018.