Evaluating Gender Bias in Machine Translation. (Reproduction)

Antoine RODRIGUEZ, Lilia HARIRECHE, Douaa BENHADDOUCHE
Machine Learning for Data Science, University of Paris, Paris, FRANCE
rodriguezantoine5@gmail.com, lilia.harireche@gmail.com, douaa2314@hotmail.fr

Abstract

Gender bias is one of the largest ethical issue surrounding Machine translation (MT), because it reduces significantly the translation quality, specially when the target language has grammatical gender. This work aims to evaluate a coreference resolution challenge set of English sentences that contains non-stéréotypical gender roles on eight diffrent languages.The evaluation method is based on morphological analysis and performed on industrial MT systems which show an important gender biased transtation errors for all the target languages.

Keywords: MT: Machine Translation, Gender bias, coreference resolution

1. Introduction

Stereotypes in their diffrent kind such as : lexical semantics (“man is to computer programmer as woman is to homemaker: Zhao et al., 2017), and natural language inference (associating women with gossiping and men with guitars; Bolukbasi et al., 2016), lead to a gender bias when the trained models do not consider roles and tasks relevant in the traslation. The Figure 1 below illustrate gender bias in machine translation from English to Spanish. In English source sentence the gender of nurse is unknown but the coreference ”her” identifies the doctor as female. Contrarily, the Spanish translation sentence uses morphological features for gender: “el doctor” (male), versus “la enfermera” (female). The alignement between source and target sentences reveals that a stereotypical assignment of gender roles changed the meaning of the translated sentence by changing the doctor’s gender. Our experiement compares and analysis the performances and the tolerance of this issue in the most a popular MT systems as Google Translate or Microsoft Translator, that suffers from biases, e.g., translating nurses as females and programmers as males, regardless of context. Generating a chosen translation was the solution proposed by Google to remedy social bias.

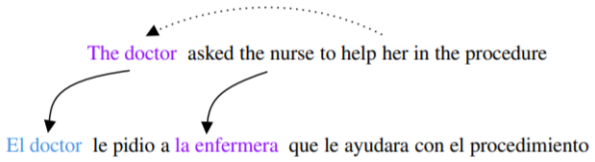

Figure 1: Gender bias exemple in MT from English to Spanish.

The challenge set is made up of 3888 English sentences was created by combining two datasets of two previous works: the Winogender (Rudinger et al., 2018), and the WinoBias (Zhao et al., 2018). Each sentence in these datasets describes a scenario with human entities, identified by their role (e.g., “the doctor” and “the nurse” in Figure 1), and a pronoun (“her” in the example), which needs to be correctly referenced to the right entity. Using this dataset, we evaluated the gender-bias problem on the four industrial MT for eight diverse target languages by computing measures of alignement and morphological analysis. The rest of the paper present our implementation of the evaluation method procedure and the obtained results on the diffetent MT systems.

2. Conception and Implementation

The translations were performed on four MT systems: Google Translate, Microsoft Translate, Amazon Translate and SYSTRAN. To estimate the gender-bias of an MT model we have followed the following steps:

A) Since we could not recover API keys to be able to do the translations between english and eight target languages with MT systems, we opted for the use of *translation_data* file, also we have used: *original_english_data* and *aligns_data* which is described below.

B) In order to align between the source and target translations, we have used *fast_align* proposed in (**?**) that produces outputs in the $i - j$, where a pair $i - j$ indicates that the ith word of the left language is aligned to the jth word of the right sentence, so we extracted the different files from the *aligns_data* folder based on a C++ implementation.

C) Then, we have extracted the target-side entitys gender using simple heuristics over language specific morphological analysis, which we perform by using implementation tools for each target language explained below. The implementation tools are provided in a *languages* folder in the code.

As explained before, to get a faithful translation, many languages associate between biological and grammar gender; which provides an easy gender identification using the morphological markers.
According to their properties, our eight target languages belong to four different families:

1) **Romance languages:** this family nouns’s gender agreement depends on the determiner, like in Spanish, French, and Italian, in our experiements we used the spaCy morphological analysis support proposed in

(Honnibal and Montani, 2017).

2) **Slavic languages (Cyrillic alphabet):** as Russian and Ukrainian, these two languages are quit similar since their words come from the same roots, to identify genders we used the morpho- logical analyzer developed by Korobov (2015).

3) **Semitic languages:** like Hebrew and Arabic, each with a unique alphabet. in Arabic it's easy to identifiy the female gender indicated via the ta marbuta character at the end of the word. For Hebrew, the female gender can be also identified by specific markers, for predictions we used the analyzer developed by Adler and Elhadad (2006),

4) **Germanic languages:** in german, we identify genders through the different determiners and with the female marker "in". For predictions, we used the morphological analyzer developed by Altinok (2018).

The implementations of the different morphological analysers are provided in a folder nammed Languages in our project code.

3. Results and Discussion

The process steps in the previous part allowed us to extract the translated genders, according to the MT model, which we have evaluated against the gold genders provided by the original english dataset using *evaluate_bias function* described in the code. After doing the evaluation, we decided to calculate three metrics for the four MT system: firstly, we have calculated the *Accuracy* which is the percentage of instances in which the translation preserved the gender of the entity from the original english sentence. Second, Δ_G that denotes the difference in performance (F1 score) between male and female. Third, Δ_S measures the difference in performance (F1 score) between stereotypical and non-stereotypical gender role assignments, as defined by Zhao et al. (2018). According to the results we obtained, We find that most tested systems across eight tested languages perform quite poorly on the accuracy, although the latter is hidh in Google Translate. Also, Δ_S metric shows that all tested systems have a significant and consistently better performance when presented with pro-stereotypical assignments, while their performance deteriorates when translating anti stereotypical roles. Finally Δ_G metric shows that all MT systems, except Microsoft Translator on German, perform better on male roles.

The table below shows us the different results obtained.

	Google			**Aws**		
	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S
FR	**62.6**	6.5	23.1	54.6	17.3	17.9
ES	48.4	20.7	14.1	**54.3**	12.6	14.8
IT	41.9	31.2	20.2	43.6	25.5	18.7
AR	45.9	42.5	11.5	**48.2**	37.8	13.5
HE	**50.8**	11.5	32.5	48	13.7	41.4
RU	**37.7**	36.8	18.2	39.7	34.7	15.1
UK	38.4	43.6	12.9	-	-	-
DE	59.4	12.6	10.4	-	-	-

	Bing			**Systran**		
	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S
FR	44.6	36.9	19.4	43.9	44.4	7.9
ES	40.8	33	24.2	40.9	44.3	14.1
IT	**41.9**	36.7	14.8	41.7	46.8	6.9
AR	45	47.1	10.8	45.6	49.2	2.1
HE	44	21.9	28.1	43.2	27.7	19.7
RU	36.8	42	16.3	37.3	44.1	16.7
UK	**41.3**	46.9	12	28.9	22.3	12.4
DE	74.1	0	8.7	8.6	34.5	7.3

Table 1: Summary table of the evaluation results.

4. Conclusion

In order to reproduce the analysis of gender bias in machine translation (MT), we have based on the article *Evaluating Gender Bias in Machine Translation* realized by G.Stanovsky, A.Smith, and L.Zettlemoyer. For this, we used the four MT system to translate the sentences in english into the eight target languages. We made an evaluation between the gold genders and those predicted based on the calculation of three metrics: *Accuracy,* Δ_G, Δ_S. The url to our reproduction repository is gender-bias-repo.

5. Acknowledgements

For the realization of this work, we had decided to divide the tasks between the writing of the article and the understanding of the parts of code, thus launching all the evaluations. Antoine Rodriguez worked on code. Douaa Benhaddouche wrote the article. Lilia Harireche worked on both article and code.

6. Bibliographical References

Meni Adler and Michael Elhadad. DEMorphy, German Language Morphological Analyzer. 2006.

Duygu Altinok. An Unsupervised Morpheme-Based HMM for Hebrew Morphological Disambiguation. *ArXiv*, 2018.

Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. Adaptive Neural Networks for Efficient Inference. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017.

Chris Dyer, Victor Chahuneau, and Noah 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*, Atlanta, Georgia, 2013. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

Mikhail Korobov. Morphological Analyzer and Generator for Russian and Ukrainian Languages. 2015.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*,

Valencia, Spain, 2017. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *NAACL (short)*, 2018.