Towards Reducing Gender Bias in South Asian Language Translations

by

Nazmul Hasan Wanjan
Student ID: 18201133
Basit Hussain
Student ID: 21141064
Malika Muradi
Student ID: 21241057

School of Computer Science and Engineering
Brac University
March 2023

Abstract

While we have come very far in reducing gender bias in the society by taking steps towards that for a long time, works on reducing such bias from Natural Language Processing tasks are fairly new. In this paper, we have tried to add to the existing works on that topic by looking into the gender bias that is observed during Machine Translation, one of the most notable applications of Natural Language Processing, and have tried to reduce that with a few pre existing methods. The purpose of this paper therefore is to help the current machine translation systems reach a better state with more accurate prediction of the subject's gender with lesser dependency on other factors (mostly occupation) that are often unrelated.

Keywords: Gender Bias, Machine Translation, Natural Language Processing, Bias

Table of Contents

Chapter 1

Introduction

Machine Learning fairness is a relatively new area which studies how to reduce favoring any specific segment of the population in machine learning models that may be the consequence of data or model inaccuracies. The end goal is to ensure equality for every segment of the population when they interact with the systems developed using machine learning. Older systems suffer from the problems of wrongly identifying segments of the population either due to the algorithmic or the dataset biases that are presented to it during the time of training. These biases, either coming willingly or unwillingly, hamper the modeling process and make it less generalized and usable for the whole population it is designed for. Additionally, it runs the risk of creating conflict among individuals in the real world as these models are often designed with the goal of being used in the real world systems.

Of the topics that concern machine learning fairness, bias reduction is a notable one. It is bias that leads us to stereotyping demographic groups, suppressing some and even leading the individuals within those demographic groups to believe certain aspects about themselves which are often misleading and wrong. Bias is an umbrella term to define the pre existing beliefs or shortcuts that may give a certain segment of the population some sort of edge over others. This has many types, one being gender bias (which is closely related to the topic of discussion of this paper), that comes from the beliefs about certain genders that exist in the society. For instance, historically, the majority of the existing cultures have been patriarchal. This means that the chief earning member is the father and he is also the head decision maker. Due to that, women are viewed as the segment who are weak and not suited to do physically and mentally challenging jobs. These pre-existing and dated thoughts, although useful in some situations in the past, are getting more and more inaccurate in the modern world where women, as well as other segments of the population, are joining the workforce for these demanding jobs.

In this work, we have tried to peer into the gender bias that may arise from the faulty and incomplete data and algorithms used for training the machine translation models, which is one of the most popular areas of Natural Language Processing. We specifically look into three of the South Asian languages: Bengali, Dari, and Urdu, and measure the gender bias that may occur when translating to and from English. Then we try to reduce the bias using different existing methodologies and compare the results to check which one performs the best for this particular task.

It has to be noted, however, that although we have used only three languages for this

work, it is nowhere near the number of languages used in the Southern part of the Asian subcontinent. Therefore, while this generalization may work for some related languages in the same family, it may not be suitable for all.

Chapter 2

Literature Review

Previous works done by the researchers have found significant bias in the existing machine translation systems. One work by mentions how the faulty word embeddings used for machine translation causes bias in the system and tries to use debiasing approaches to have a better model. Reviews on the bias reduction method are useful to get the whole picture of the current landscape, scope and state-of-the-art methods of the process. For instance, analyzes the gender bias in NLP systems across multiple languages by using Wikipedia Corpora across languages and extends Professional and Corpora Level Gender Bias Metrics. summarizes the current conceptualizations of bias, the previous works aimed at assessing the gender bias in Machine translation and mitigation strategies so far by looking at different existing benchmarks. With the view to shed further light on the phenomenon of Machine Bias and raising awareness about updating the current automatic translation tools using debiasing techniques, collects a comprehensive list of job positions from the BLS dataset provided by the U.S. Bureau of Labor Statistics. They then build sentences using a general pattern to backtranslate using Google Translate API. The authors found a strong tendency of the system towards male default which leads to inaccurate representation of the real world statistics, with a more pronounced inaccuracy in case of STEM jobs. (Cite) works towards developing a new method to measure gender bias in 9 gendered languages including English. They perform this by extending (cite)'s method that utilizes word embedding systems like Word2Vec to measure gender bias

Chapter 3

Workflow

Gathering the job position data from different job sites in 3 countries where the selected languages are spoken (Bn, Da, and Ur) If the dataset is small, augment with the BLS dataset using backtranslation Adding texts with the occupation to make full sentences a) without context, b) with context Check the no. of instances the predictions match Compare with the real world stats. Apply different bias reduction techniques on the generated sentences Check the results for each using different benchmarks