Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques

Joel Escudé Font and Marta R. Costa-jussà
Universitat Politècnica de Catalunya, 08034 Barcelona
joel.escude@estudiant.upc.edu,marta.ruiz@upc.edu

Abstract

Neural machine translation has significantly pushed forward the quality of the field. However, there are remaining big issues with the translations and one of them is fairness. Neural models are trained on large text corpora which contains biases and stereotypes. As a consequence, models inherit these social biases. Recent methods have shown results in reducing gender bias in other natural language processing applications such as word embeddings.We take advantage of the fact that word embeddings are used in neural machine translation to propose the first debiased machine translation system. Specifically, we propose, experiment and analyze the integration of two debiasing techniques over GloVe embeddings in the Transformer translation architecture. We evaluate our proposed system on a generic English-Spanish task, showing gains up to one BLEU point. As for the gender bias evaluation, we generate a test set of occupations and we show that our proposed system learns to equalize existing biases from the baseline system.

1 Introduction

Language is one of the most interesting and complex skills used in our daily life, and may even be taken for granted on our ability to communicate. However, the understanding of meanings between lines in natural languages is not straightforward for the logic rules of programming languages. Natural language processing (NLP) is a sub-field of artificial intelligence that focuses on making natural languages understandable to computers. Similarly, the translation between different natural languages is a task for machine translation (MT).

Neural machine translation (NMT) is a recent approach in MT which learns patterns between source and target language corpora to produce text translations using deep neural networks (Sutskever et al., 2014).

One downside of models trained with human generated corpora is that social biases present in the data are learned. This is shown when training word embeddings, a vector representation of words, in news sets with crowd-sourcing evaluation to quantify the presence of biases, such as gender bias, in those representation (Bolukbasi et al., 2016). This can affect downstream applications (Zhao et al., 2018a) and are at risk of being amplified (Zhao et al., 2017).

The objective of this work is to study the presence of gender bias in MT and give insight on the impact of debiasing in such systems. An example of this gender bias is the word ”friend” in the English sentence ”She works in a hospital, my friend is a nurse” would be correctly translated to ”amiga” (feminine of friend) in Spanish, while ”She works in a hospital, my friend is a doctor” would be incorrectly translated to ”amigo” (masculine of friend) in Spanish. We consider that this translation contains gender bias since it ignores the fact that, for both cases, ”friend” is a female and translates by focusing on the occupational stereotypes, i.e. translating doctor as male and nurse as female.

The main contribution of this study is providing progress on the recent detected problem which is gender bias in MT (Prates et al., 2018). The progress towards reducing gender bias in MT is made in two directions: first, we define a framework to experiment, detect and evaluate gender bias in MT for a particular task; second, we propose to use debiased word embeddings techniques in the MT system to reduce the detected bias. This is the first study in proposing debiasing techniques for MT.

The rest the paper is organized as follows. Section 2 reports material relevant to the background of the study. Section 3 presents previous work on the bias problem. Section 4 reports the methodology used for experimentation and section 5 details the experimental framework. The results and discussion are included in section 6 and section 7 presents the main conclusions and ideas for further work.

2 Background

This section presents the two most important models that are used in this paper. First, we describe what is the transformer model which is the state-of-the-art model in MT and second, we report a brief description of word embeddings and the corresponding techniques to debias them.

2.1 Transformer

The Transformer (Vaswani et al., 2017) is a neural network architecture purely based on self-attention mechanisms that show an improvement in performance on MT tasks over previous recurrent and convolutional models, also is more efficient in using computation resources and faster in training.

Neural networks start by representing individual words as vectors, word embeddings (more on this later), to process language as a vector space representation which can have a fixed or variable length. Words surrounding another word determine its meaning and how it is represented in this space, thus context influences in deciding the appropriate meaning for a given task using such representation.

The Transformer computes a reduced constant number of steps using a self-attention mechanism on each one. This mechanism models the relations between words independently of their position, thus improving the number of steps needed to determine a target word. An attention score is computed for all words in a sentence when comparing the contribution of each word to the next representation. An encoder reads an input sentence to generate a representation which is later used by a decoder to produce a sentence output word by word. New representations are generated at each step in parallel for all words. The decoder uses self-attention in the generated words and also uses the representations from the last words in the encoder to produce a single word each time.

2.2 Word embeddings

Word embeddings are vector representations of words. This representation is less sparse and more expressive, opposite to discrete atomic symbols and one-hot vectors. It is used in many NLP applications. Based on the hypothesis that words appearing in same contexts share semantic meaning, this continuous vector space representation gathers semantically similar words.

Arithmetic operations can be performed with these embeddings, in order to find analogies between pairs of nouns with the pattern ”A is to B what C is to D” (Mikolov et al., 2013). For nouns, such as countries and their respective capitals or for the conjugations of verbs.

While there are many techniques for extracting word embeddings, in this work we are using Global Vectors, or GloVe (Pennington et al., 2014). Glove is an unsupervised method for learning word embeddings. This count-based method, uses statistical information of word occurrences from a given corpus to train a vector space for which each vector is related to a word and their values describes their semantic relations.

2.3 Debiasing word embeddings

The presence of biases in word embeddings has aroused as a topic of discussion about fairness. More specifically, gender stereotypes are learned from human generated corpora as shown by (Bolukbasi et al., 2016). Several debiasing approaches have been proposed. Debiaswe is a post-process method for debiasing previously generated embeddings (Bolukbasi et al., 2016). GN-GloVe is a method for generating gender neutral embeddings (Zhao et al., 2018b). The main ideas behind these algorithms are described next.

Debiaswe (Bolukbasi et al., 2016) is a post-process method for debiasing word embeddings. It consists of two main parts: First the direction of the embeddings where the bias is present is identified. Second, the gender neutral words in this direction are neutralized to zero and also equalizes the sets by making the neutral word equidistant to the remaining ones in the set. The disadvantage of the first part of the process is that it can remove valuable information in the embed-