Evaluating the Underlying Gender Bias in Contextualized Word Embeddings

Christine Basta **Marta R. Costa-jussà** **Noe Casas**
Universitat Politècnica de Catalunya
{christine.raouf.saad.basta,marta.ruiz,noe.casas}@upc.edu

Abstract

Gender bias is highly impacting natural language processing applications. Word embeddings have clearly been proven both to keep and amplify gender biases that are present in current data sources. Recently, contextualized word embeddings have enhanced previous word embedding techniques by computing word vector representations dependent on the sentence they appear in.

In this paper, we study the impact of this conceptual change in the word embedding computation in relation with gender bias. Our analysis includes different measures previously applied in the literature to standard word embeddings. Our findings suggest that contextualized word embeddings are less biased than standard ones even when the latter are debiased.

1 Introduction

Social biases in machine learning, in general and in natural language processing (NLP) applications in particular, are raising the alarm of the scientific community. Examples of these biases are evidences such that face recognition systems or speech recognition systems works better for white men than for ethnic minorities (Buolamwini and Gebru, 2018). Examples in the area of NLP are the case of machine translation that systems tend to ignore the coreference information in benefit of an stereotype (Font and Costa-jussà, 2019) or sentiment analysis where higher sentiment intensity prediction is biased for a particular gender (Kiritchenko and Mohammad, 2018).

In this work we focus on the particular NLP area of word embeddings (Mikolov et al., 2010), which represent words in a numerical vector space. Word embeddings representation spaces are known to present geometrical phenomena mimicking relations and analogies between words (e.g. *man* is to *woman* as *king* is to *queen*). Following this property of finding relations or analogies, one popular example of gender bias is the word association between *man* to *computer programmer* as *woman* to *homemaker* (Bolukbasi et al., 2016). Pre-trained word embeddings are used in many NLP downstream tasks, such as natural language inference (NLI), machine translation (MT) or question answering (QA). Recent progress in word embedding techniques has been achieved with contextualized word embeddings (Peters et al., 2018) which provide different vector representations for the same word in different contexts.

While gender bias has been studied, detected and partially addressed for standard word embeddings techniques (Bolukbasi et al., 2016; Zhao et al., 2018a; Gonen and Goldberg, 2019), it is not the case for the latest techniques of contextualized word embeddings. Only just recently, Zhao et al. (2019) present a first analysis on the topic based on the proposed methods in Bolukbasi et al. (2016). In this paper, we further analyse the presence of gender biases in contextualized word embeddings by means of the proposed methods in Gonen and Goldberg (2019). For this, in section 2 we provide an overview of the relevant work on which we build our analysis; in section 3 we state the specific request questions addressed in this work, while in section 4 we describe the experimental framework proposed to address them and in section 5 we present the obtained and discuss the results; finally, in section 6 we draw the conclusions of our work and propose some further research.

2 Background

In this section we describe the relevant NLP techniques used along the paper, including word embeddings, their debiased version and contextualized word representations.