Mitigating Gender Bias Amplification in Distribution by Posterior Regularization

Shengyu Jia[♣]*, **Tao Meng**[♠]*, **Jieyu Zhao**[♠], **Kai-Wei Chang**[♠]
[♣] Tsinghua University
[♠] University of California, Los Angeles
jiasy16@mails.tsinghua.edu.cn,
{mengt18, jieyuzhao, kwchang}@ucla.edu

Abstract

Advanced machine learning techniques have boosted the performance of natural language processing. Nevertheless, recent studies, e.g., Zhao et al. (2017) show that these techniques inadvertently capture the societal bias hidden in the corpus and further amplify it. However, their analysis is conducted only on models' top predictions. In this paper, we investigate the gender bias amplification issue from the distribution perspective and demonstrate that the bias is amplified in the view of predicted probability distribution over labels. We further propose a bias mitigation approach based on posterior regularization. With little performance loss, our method can almost remove the bias amplification in the distribution. Our study sheds the light on understanding the bias amplification.

1 Introduction

Data-driven machine learning models have achieved high performance in various applications. Despite the impressive results, recent studies (e.g., Wang et al. (2019); Hendricks et al. (2018)) demonstrate that these models may carry societal biases exhibited in the dataset they trained on. In particular, Zhao et al. (2017) show that a model trained on a biased dataset may amplify the bias. For example, we can consider a task of labeling the activity and objects depicted in an image. The training set contains 30% more images with "woman cooking" than "man cooking". However, when evaluating the top predictions of a trained model, the disparity between males and females is amplified to around 70%. Based on this observation, Zhao et al. (2017) conduct a systematic study and propose to calibrate the top predictions of a learned model by injecting corpus-level constraints to ensure that the gender disparity is not amplified.

However, when analyzing the top predictions, the models are forced to make one decision. Therefore, even if the model assigns high scores to both labels of "woman cooking" and "man cooking", it has to pick one as the prediction. This process obviously has a risk to amplify the bias. However, to our surprise, we observe that gender bias is also amplified when analyzing the posterior distribution of the predictions. Since the model is trained with regularized maximal likelihood objective, the bias in distribution is a more fundamental perspective of analyzing the bias amplification issue.

In this paper, we conduct a systematic study to quantify the bias in the predicted distribution over labels. Our analysis demonstrates that when evaluating the distribution, though not as significant as when evaluating top predictions, the bias amplification exists. About half of activities show significant bias amplification in the posterior distribution, and on average, they amplify the bias by 3.2%.

We further propose a new bias mitigation technique based on posterior regularization because the approaches described in Zhao et al. (2017) can not be straightforwardly extended to calibrate bias amplification in distribution. With the proposed technique, we successfully remove the bias amplification in the posterior distribution while maintain the performance of the model. Besides, the bias amplification in the top predictions based on the calibrated distribution is also mitigated by around 30%. These results suggest that the bias amplification in top predictions comes from both the requirement of making hard predictions and the bias amplification in the posterior distribution of the model predictions. Our study advances the understanding of the bias amplification issue in natural language processing models. The code and data are available at https://github.com/uclanlp/reducingbias.

*Both authors contributed equally to this work and are listed in alphabetical order.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2936–2942
July 5 - 10, 2020.

2 Related Work

Algorithmic Bias Machine learning models are becoming more and more prevalent in the real world, and algorithmic bias will have a great societal impact (Tonry, 2010; Buolamwini and Gebru, 2018). Researchers have found societal bias in different applications such as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), machine translation (Stanovsky et al., 2019) and online advertisement (Sweeney, 2013). Without appropriate adjustments, the model can amplify the bias (Zhao et al., 2017). Different from the previous work, we aim at understanding the bias amplification from the posterior perspective instead of directly looking at the top predictions of the model.

Posterior Regularization The posterior regularization framework (Ganchev et al., 2010) is aiming to represent and enforce constraints on the posterior distribution. It has been shown effective to inject domain knowledge for NLP applications. For example, Ji et al. (2012); Gao et al. (2014) design constraints based on similarity to improve question answering and machine translation, respectively. Yang and Cardie (2014) propose constraints based on lexical patterns in sentiment analysis. Meng et al. (2019) apply corpus-level constraints to guide a dependency parser in the cross-lingual transfer setting. In this paper we leverage corpus-level constraints to calibrate the output distribution. Our study resembles to the confidence calibration (Guo et al., 2017; Naeini et al., 2015). However, the temperature turning and binning methods proposed in these papers cannot straightforwardly be extended to calibrate the bias amplification.

3 Background

We follow the settings in Zhao et al. (2017) to focus on the imSitu vSRL dataset (Yatskar et al., 2016), in which we are supposed to predict the activities and roles in given images and this can be regraded as a structure prediction task (see Fig. 1).

We apply the Conditional Random Field (CRF) model for the structure prediction task. We denote $\mathbf{y}$ as a joint prediction result for all instances, and $\mathbf{y}^i$ as a prediction result for instance i. We use $\mathbf{y}_v$ to denote the predicted activity, and $\mathbf{y}_r$ to denote the predicted role. An activity can have multiple roles and usually one of them conveys the gender information. For an instance i, the CRF model predicts the scores for every activity and role, and

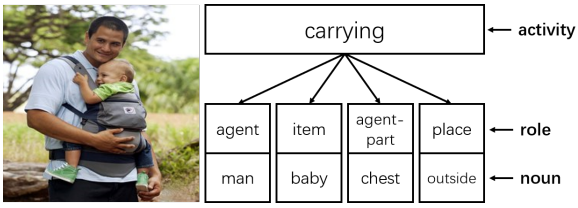


Figure 1: An instance from the imSitu dataset. Given an input image, the task it to identify the activity depicted in the image as well as the objects (noun) and their semantic role.

the score for a prediction is the summation of all these scores. Formally,

$$f_\theta(\mathbf{y}^i, i) = s_\theta(\mathbf{y}^i_v, i) + \sum\nolimits_{e \in \mathbf{y}^i_r} s_\theta(\mathbf{y}^i_v, e, i),$$

where $s_\theta(\mathbf{y}^i_v, i)$ and $s_\theta(\mathbf{y}^i_v, e, i)$ are the scores for activity $\mathbf{y}^i_v$ of instance i, and the score for role e of instance i with activity $\mathbf{y}^i_v$, respectively. We can infer the top structure for instance i by:

$$\arg\max_{\mathbf{y}^i \in \mathcal{Y}^i} f_\theta(\mathbf{y}^i, i),$$

where $\mathcal{Y}^i$ refers to all the possible assignments to the instance.

4 Bias Amplification Quantification and Corpus-level Constraints

Zhao et al. (2017) demonstrate bias amplification in the top prediction and present a bias mitigation technique by inference with corpus-level constraints. In the following, we extend their study to analyze the bias amplification in the posterior distribution by the CRF model and define the corresponding corpus-level constraints.

Formally, the probability of prediction $\mathbf{y}^i$ for instance i and the joint prediction $\mathbf{y}$ defined by CRF model with parameters θ are given by

$$\begin{aligned} p_\theta(\mathbf{y}^i, i) &\propto \exp(f_\theta(\mathbf{y}^i, i)), \\ p_\theta(\mathbf{y}) &= \prod\nolimits_i p_\theta(\mathbf{y}^i, i), \end{aligned} \tag{1}$$

since instances are mutually independent.

In this section, we will define how to quantify the bias and the bias amplification in the distribution, and introduce the corpus-level constraints towards restricting the bias in the distribution.

We focus on the gender bias on activities in the vSRL task. To quantify the gender bias given a particular activity v^*, Zhao et al. (2017) uses the percentage that v^* is predicted together with male agents among all prediction with genders. This

evaluation focuses on the top prediction. In the contrast, we define bias function $B(p, v^*, D)$ w.r.t distribution p and activity v^*, evaluating the bias toward male in dataset D based on the conditional probability $P(X|Y)$, where $event\ Y$: given an instance, its activity is predicted to be v^* and its role is predicted to have a gender; $event\ X$: this instance is predicted to have gender male. Formally,

$$\begin{aligned} &B(p, v^*, D) \\ =&\mathbb{P}_{i\sim D, \mathbf{y}\sim p}(\mathbf{y}_r^i \in M | \mathbf{y}_v^i = v^* \wedge \mathbf{y}_r^i \in M \cup W) \\ =&\frac{\sum_{i\in D} \sum_{\mathbf{y}^i:\mathbf{y}_v^i=v^*, \mathbf{y}_r^i\in M} p(\mathbf{y}^i, i)}{\sum_{i\in D} \sum_{\mathbf{y}^i:\mathbf{y}_v^i=v^*, \mathbf{y}_r^i\in M\cup W} p(\mathbf{y}^i, i)}. \end{aligned} \tag{2}$$

This bias can come from the training set D_{tr}. Here we use $b^*(v^*, male)$ to denote the "dataset bias" toward male in the training set, measured by the ratio of between male and female from the labels:

$$b^* = \frac{\sum_{i\in D_{tr}} \mathbf{1}[\hat{\mathbf{y}}_v^i = v^*, \hat{\mathbf{y}}_r^i \in M]}{\sum_{i\in D_{tr}} \mathbf{1}[\hat{\mathbf{y}}_v^i = v^*, \hat{\mathbf{y}}_r^i \in M \cup W]},$$

where $\hat{\mathbf{y}}^i$ denotes the label of instance i.

Ideally, the bias in the distribution given by CRF model should be consistent with the bias in the training set, since CRF model is trained by maximum likelihood. However, the amplification exists in practice. Here we use the difference between the bias in the posterior distribution and in training set to quantify the bias amplification, and average it over all activities to quantify the amplification in the whole dataset:

$$\begin{aligned} A(p, v^*, D) &= sgn(b^* - 0.5)[B(p, v^*, D) - b^*], \\ \bar{A}(p, D) &= \frac{1}{|V|} \sum_{v^*\in V} A(p, v^*, D). \end{aligned}$$

Note that if we use the top prediction indicator function to replace p in $A, \bar{A}$, it is the same as the definition of the bias amplification in top prediction in Zhao et al. (2017).

The corpus-level constraints aim at mitigating the bias amplification in test set D_{ts} within a predefined margin γ,

$$\forall v^*,\ |A(p, v^*, D_{ts})| \le \gamma. \tag{3}$$

5 Posterior Regularization

Posterior regularization (Ganchev et al., 2010) is an algorithm leveraging corpus-level constraints to regularize the posterior distribution for a structure model. Specifically, given corpus-level constraints and a distribution predicted by a model, we 1) define a feasible set of the distributions with respect to the constraints; 2) find the closest distribution in the feasible set from given distribution; 3) do maximum a posteriori (MAP) inference on the optimal feasible distribution.

The feasible distribution set Q is defined by the corpus-level constraints defined in Eq. (3):

$$Q = \{q \mid \forall v^*,\ |B(q, v^*, D_{ts}) - b^*| \le \gamma\}, \tag{4}$$

where $B(\cdot)$ is defined in Eq. (2).

Given the feasible set Q and the model distribution p_θ defined by Eq. (1), we want to find the closest feasible distribution q^* :

$$q^* = \arg\min_{q\in Q} KL(q \| p_\theta). \tag{5}$$

This is an optimization problem and our variable is the joint distribution q with constraints, which is intractable in general. Luckily, according to the results in Ganchev et al. (2010), if the feasible set Q is defined in terms of constraints feature functions ϕ and their expectations:

$$Q = \{q \mid \mathbb{E}_{\mathbf{y}\sim q}[\phi(\mathbf{y}) \le \mathbf{c}]\}, \tag{6}$$

Eq. (5) will have a close form solution

$$q^*(\mathbf{y}) = \frac{p_\theta(\mathbf{y}) \exp(-\lambda^* \cdot \phi(\mathbf{y}))}{Z(\lambda^*)}, \tag{7}$$

where λ^* is the solution of

$$\begin{aligned} \lambda^* &= \arg\max_{\lambda\ge 0} -\mathbf{c}\cdot\lambda - \log Z(\lambda). \\ Z(\lambda) &= \sum\nolimits_{\mathbf{y}} p_\theta(\mathbf{y}) \exp(-\lambda \cdot \phi(\mathbf{y})). \end{aligned} \tag{8}$$

Actually, we can derive the constraints into the form we want. We set $\mathbf{c} = \mathbf{0}$ and

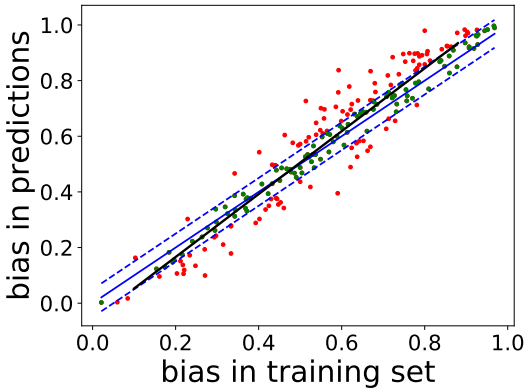
$$\phi(\mathbf{y}) = \sum\nolimits_i \phi^i(\mathbf{y}^i). \tag{9}$$

We can choose a proper $\phi^i(\mathbf{y}^i)$ to make Eq. (4) equal to Eq. (6). The detailed derivation and the definition of $\phi^i(\mathbf{y}^i)$ are shown in Appendix A.

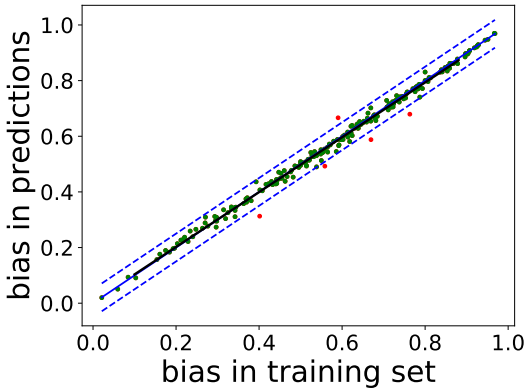
We can solve Eq. (8) by gradient-based methods to get λ^*, and further compute the close form solution in Eq. (7). Actually, considering the relation between $\mathbf{y}$ and $\mathbf{y}^i$ in Eq. (1) and (9), we can factorize the solution in Eq. (7) on instance level:

$$q^*(\mathbf{y}^i, i) = \frac{p_\theta(\mathbf{y}^i, i) \exp(-\lambda^* \cdot \phi^i(\mathbf{y}^i))}{Z^i(\lambda^*)},$$

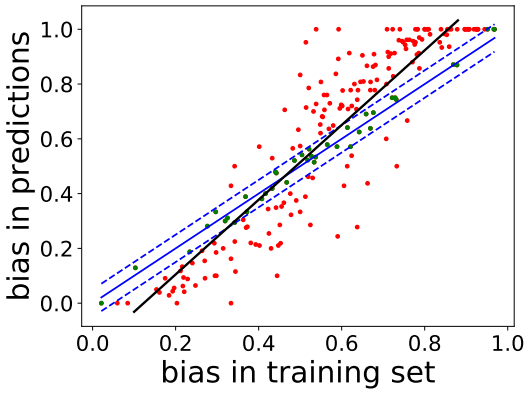
and the derivation details are in Appendix B. With this, we can reuse original inference algorithm to conduct MAP inference based on the distribution q^* for every instance seperately.



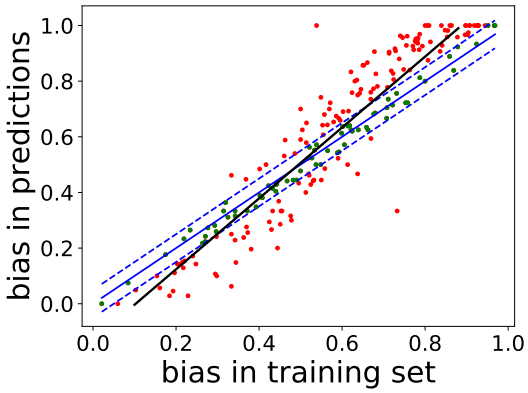
(a) bias in distribution before bias mitigation.



(b) bias in distribution after bias mitigation.



(c) bias in top predictions before bias mitigation.



(d) bias in top predictions after bias mitigation.

Figure 2: x-axis and y-axis are the bias toward male in the training corpus and the predictions, respectively. Each dot stands for an activity. The blue reference lines indicate the bias score in training is equal to that in test and the dash lines indicate the margin ($= 0.05$). The dots in red stand for being out of margin and violating the constraints. The black lines are linear regressions of the dots. Results show that we can almost remove the bias amplification in distributions (see 2a and 2b), and reduce 30.9% amplification in top predictions (see 2c and 2d) after applying posterior regularization.

6 Experiments

We conduct experiments on the vSRL task to analyze the bias amplification issue in the posterior distribution and demonstrate the effectiveness of the proposed bias mitigation technique.

Dataset Our experiment settings follow Zhao et al. (2017). We evaluate on imSitu (Yatskar et al., 2016) that activities are selected from verbs, roles are from FrameNet (Baker et al., 1998) and nouns from WordNet (Fellbaum, 1998). We filter out the non-human oriented verbs and images with labels that do not indicate the genders.

Model We analyze the model purposed together with the dataset. The score functions we describe in Sec. 3 are modeled by VGG (Simonyan and Zisserman, 2015) with a feedforward layer on the top of it. The scores are fed to CRF for inference.

6.1 Bias Amplification in Distribution

Figures 2a and 2c demonstrate the bias amplification in both posterior distribution p_θ and the top predictions $\mathbf{y}$ defined in Sec.4, respectively. For most activities with the bias toward male (i.e., higher bias score) in the training set, both the top prediction and posterior distribution are even more biased toward male, vise versa. If the bias is not amplified, the dots should be scattered around the reference line. However, most dots are on the top-right or bottom-left, showing the bias is amplified. The black regression line with $slope > 1$ also indicates the amplification. Quantitatively, 109 and 173 constraints are violated when analyzing the bias in distribution an in top predictions.

Most recent models are trained by minimizing the cross-entropy loss which aims at fitting the model's predicted distribution with observed distribution on the training data. In the inference time,

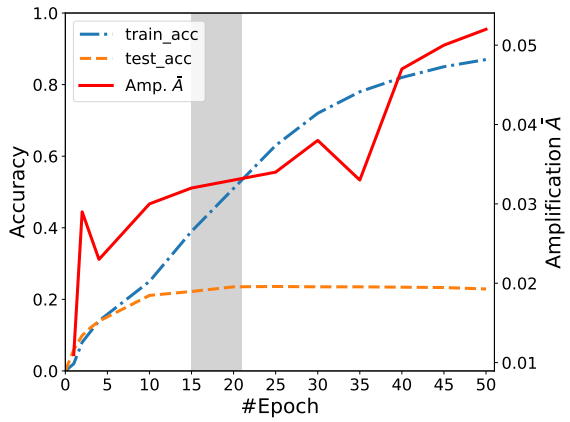


Figure 3: The curve of training and test accuracy, and bias amplification with the number of training epochs. The optimal model evaluated on the development set is found in the grey shade area.

the model outputs the top predictions based on the underlying prediction distribution. Besides, in practice, the distribution has been used as an indicator of confidence in the prediction. Therefore, understanding bias amplification in distribution provides a better view about this issue.

To analyze the cause of bias amplification, we further show the degree of amplification along with the learning curve of the model (see Fig. 3). We observed that when the model is overfitted, the distribution of the model prediction becomes more peaky[1]. We suspect this is one of the key reasons causes the bias amplification.

6.2 Bias Amplification Mitigation

We set the margin $\gamma = 0.05$ for every constraint in evaluation. However, we employ a stricter margin ($\gamma = 0.001$) in performing posterior regularization to encourage the model to achieve a better feasible solution. We use mini-batch to estimate the gradient w.r.t λ with Adam optimizer (Kingma and Ba, 2015) when solving Eq. (5). We set the batchsize to be 39 and train for 10 epochs. The learning rate is initialized as 0.1 and decays after every mini-batch with the decay factor 0.998.

Results We then apply the posterior regularization technique to mitigate the bias amplification in distribution. Results are demonstrated in Figures 2b (distribution) and 2d (top predictions). The posterior regularization effectively calibrates the bias in distribution and only 5 constraints are violated after the calibration. The average bias amplification is close to 0 ($\bar{A}$: 0.032 to -0.005). By reducing the amplification of bias in distribution, the bias amplification in top predictions also reduced by 30.9% ($\bar{A}$: 0.097 to 0.067). At the same time, the model's performance is kept (accuracy: 23.2% to 23.1%).

Note that calibrating the bias in distribution cannot remove all bias amplification in the top predictions. We posit that the requirement of making hard predictions (i.e., maximum a posteriori estimation) also amplifies the bias when evaluating the top predictions.

7 Conclusion

We analyzed the bias amplification from the posterior distribution perspective, which provides a better view to understanding the bias amplification issue in natural language models as these models are trained with the maximum likelihood objective. We further proposed a bias mitigation technique based on posterior regularization and show that it effectively reduces the bias amplification in the distribution. Due to the limitation of the data, we only analyze the bias over binary gender. However, our analysis and the mitigation framework is general and can be adopted to other applications and other types of bias.

One remaining open question is why the gender bias in the posterior distribution is amplified. We posit that the regularization and the over-fitting nature of deep learning models might contribute to the bias amplification. However, a comprehensive study is required to prove the conjecture and we leave this as future work.

Acknowledgement This work was supported in part by National Science Foundation Grant IIS-1927554. We thank anonymous reviewers and members of the UCLA-NLP lab for their feedback.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *COLING-ACL*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*.

C Fellbaum. 1998. Wordnet: An on-line lexical database.

[1]This effect, called overconfident, has been also discussed in the literature (Guo et al., 2017).

Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*.

Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2014. Learning continuous phrase representations for translation modeling. In *ACL*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *ECCV*.

Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. 2012. Question-answer topic model for question retrieval in community question answering. In *CIKM*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Tao Meng, Nanyun Peng, and Kai-Wei Chang. 2019. Target language-aware constrained inference for cross-lingual dependency parsing. In *EMNLP-IJCNLP*.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL-HLT*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *ACL*.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM*.

Michael Tonry. 2010. The social, psychological, and political causes of racial disparities in the american criminal justice system. *Crime and justice*.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*.

Bishan Yang and Claire Cardie. 2014. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *ACL*.

Mark Yatskar, Luke S. Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL-HLT*.

A Definition of the Feature Functions

The feature function for predictions $\mathbf{y}$ is defined as the summation of feature functions for each instance $\mathbf{y}^i$, which is a $2n-$dimensional vector where n is the number of constraints. Each entry is the feature function corresponding to a constraint and the inequality sign direction. Formally,

$$\phi^i_{v^*,-}(\mathbf{y}^i) = \begin{cases} 1 - b^* - \gamma & \mathbf{y}^i_v = v^*, \mathbf{y}^i_r \in M \\ -b^* - \gamma & \mathbf{y}^i_v = v^*, \mathbf{y}^i_r \in W \\ 0 & otherwise \end{cases}$$

$$\phi^i_{v^*,+}(\mathbf{y}^i) = \begin{cases} -1 + b^* - \gamma & \mathbf{y}^i_v = v^*, \mathbf{y}^i_r \in M \\ b^* - \gamma & \mathbf{y}^i_v = v^*, \mathbf{y}^i_r \in W \\ 0 & otherwise \end{cases}$$

$$\phi^i = (\phi^i_{v_1,-}, \phi^i_{v_1,+}, ..., \phi^i_{v_n,-}, \phi^i_{v_n,+})$$

$$\phi(\mathbf{y}) = \sum_i \phi^i(\mathbf{y}^i)$$

B Derivation of Feature Functions Expectation

We can derive the feature functions expection as

$$\begin{aligned} \mathbb{E}_{\mathbf{y}\sim q}[\phi(\mathbf{y})] &\leq \mathbf{0} \\ \mathbb{E}_{\mathbf{y}\sim q}\left[\sum_i \phi^i(\mathbf{y}^i)\right] &\leq \mathbf{0} \\ \sum_i \mathbb{E}_{\mathbf{y}^i\sim q(\cdot,i)}\left[\phi^i(\mathbf{y}^i)\right] &\leq \mathbf{0} \end{aligned}$$

Thus, it is equivalent as $\forall v^*$,

$$\sum_i \mathbb{E}_{\mathbf{y}^i\sim q(\cdot,i)}\left[\phi^i_{v^*,-}(\mathbf{y}^i)\right] \leq \mathbf{0},$$

$$\sum_i \mathbb{E}_{\mathbf{y}^i\sim q(\cdot,i)}\left[\phi^i_{v^*,+}(\mathbf{y}^i)\right] \leq \mathbf{0}.$$

The inequality about $\phi^i_{v^*,-}$ can be derived as

$$\begin{aligned} \sum_i \mathbb{E}_{\mathbf{y}^i\sim q(\cdot,i)}\left[\phi^i_{v^*,-}(\mathbf{y}^i)\right] &\leq \mathbf{0} \\ \sum_i \sum_{\mathbf{y}^i} q(\mathbf{y}^i, i)\phi^i_{v^*,-}(\mathbf{y}^i) &\leq \mathbf{0} \\ \sum_i \sum_{\mathbf{y}^i:\mathbf{y}^i_v=v^*,\mathbf{y}^i_r\in M} (1 - b^* - \gamma)q(\mathbf{y}^i, i) &- \\ \sum_i \sum_{\mathbf{y}^i:\mathbf{y}^i_v=v^*,\mathbf{y}^i_r\in W} (b^* + \gamma)q(\mathbf{y}^i, i) &\leq \mathbf{0} \\ \frac{\sum_i \sum_{\mathbf{y}^i:\mathbf{y}^i_v=v^*,\mathbf{y}^i_r\in M} q(\mathbf{y}^i, i)}{\sum_i \sum_{\mathbf{y}^i:\mathbf{y}^i_v=v^*,\mathbf{y}^i_r\in M\cup W} q(\mathbf{y}^i, i)} &\leq b^* + \gamma \\ B(q, v^*, \cdot) \leq b^* + \gamma \end{aligned}$$

The inequality about $\phi^i_{v^*,-}$ can be derived similarly.