Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology

Ran Zmigrod[1]   **Sebastian J. Mielke**[2]   **Hanna Wallach**[3]   **Ryan Cotterell**[1]

[1] University of Cambridge   [2] Johns Hopkins University   [3] Microsoft Research

`rz279@cam.ac.uk` `sjmielke@jhu.edu`
`wallach@microsoft.com` `rdc42@cam.ac.uk`

Abstract

Gender stereotypes are manifest in most of the world's languages and are consequently propagated or amplified by NLP systems. Although research has focused on mitigating gender stereotypes in English, the approaches that are commonly employed produce ungrammatical sentences in morphologically rich languages. We present a novel approach for converting between masculine-inflected and feminine-inflected sentences in such languages. For Spanish and Hebrew, our approach achieves F_1 scores of 82% and 73% at the level of tags and accuracies of 90% and 87% at the level of forms. By evaluating our approach using four different languages, we show that, on average, it reduces gender stereotyping by a factor of 2.5 without any sacrifice to grammaticality.

1 Introduction

One of the biggest challenges faced by modern natural language processing (NLP) systems is the inadvertent replication or amplification of societal biases. This is because NLP systems depend on language corpora, which are inherently "not objective; they are creations of human design" (Crawford, 2013). One type of societal bias that has received considerable attention from the NLP community is gender stereotyping (Garg et al., 2017; Rudinger et al., 2017; Sutton et al., 2018). Gender stereotypes can manifest in language in overt ways. For example, the sentence *he is an engineer* is more likely to appear in a corpus than *she is an engineer* due to the current gender disparity in engineering. Consequently, any NLP system that is trained such a corpus will likely learn to associate *engineer* with men, but not with women (De-Arteaga et al., 2019).

To date, the NLP community has focused primarily on approaches for detecting and mitigating gender stereotypes in English (Bolukbasi et al., 2016; Dixon et al., 2018; Zhao et al., 2017). Yet, gender stereotypes also exist in other languages
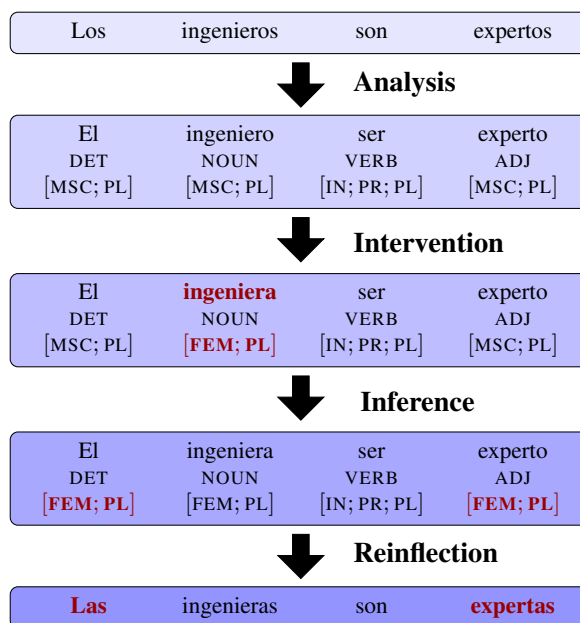

Figure 1: Transformation of *Los ingenieros son expertos* (i.e., *The male engineers are skilled*) to *Las ingenieras son expertas* (i.e., *The female engineers are skilled*). We extract the properties of each word in the sentence. We then fix a noun and its tags and infer the manner in which the remaining tags must be updated. Finally, we reinflect the lemmata to their new forms.

because they are a function of society, not of grammar. Moreover, because English does not mark grammatical gender, approaches developed for English are not transferable to morphologically rich languages that exhibit gender agreement (Corbett, 1991). In these languages, the words in a sentence are marked with morphological endings that reflect the grammatical gender of the surrounding nouns. This means that if the gender of one word changes, the others have to be updated to match. As a result, simple heuristics, such as augmenting a corpus with additional sentences in which *he* and *she* have been swapped (Zhao et al., 2018), will yield ungrammatical sentences. Consider the Spanish phrase *el ingeniero experto* (*the skilled engineer*). Replacing *ingeniero* with *ingeniera* is insufficient—*el* must

also be replaced with *la* and *experto* with *experta*.

In this paper, we present a new approach to counterfactual data augmentation (CDA; Lu et al., 2018) for mitigating gender stereotypes associated with animate[1] nouns (i.e., nouns that represent people) for morphologically rich languages. We introduce a Markov random field with an optional neural parameterization that infers the manner in which a sentence must change when altering the grammatical gender of particular nouns. We use this model as part of a four-step process, depicted in Fig. 1, to reinflect entire sentences following an intervention on the grammatical gender of one word. We intrinsically evaluate our approach using Spanish and Hebrew, achieving tag-level F_1 scores of 83% and 72% and form-level accuracies of 90% and 87%, respectively. We also conduct an extrinsic evaluation using four languages. Following Lu et al. (2018), we show that, on average, our approach reduces gender stereotyping in neural language models by a factor of 2.5 without sacrificing grammaticality.

2 Gender Stereotypes in Text

Men and women are mentioned at different rates in text (Coates, 1987). This problem is exacerbated in certain contexts. For example, the sentence *he is an engineer* is more likely to appear in a corpus than *she is an engineer* due to the current gender disparity in engineering. This imbalance in representation can have a dramatic downstream effect on NLP systems trained on such a corpus, such as giving preference to male engineers over female engineers in an automated resumé filtering system. Gender stereotypes of this sort have been observed in word embeddings (Bolukbasi et al., 2016; Sutton et al., 2018), contextual word embeddings (Zhao et al., 2019), and co-reference resolution systems (Rudinger et al., 2018; Zhao et al., 2018) *inter alia*.

A quick fix: swapping gendered words. One approach to mitigating such gender stereotypes is counterfactual data augmentation (CDA; Lu et al., 2018). In English, this involves augmenting a corpus with additional sentences in which gendered words, such as *he* and *she*, have been swapped to yield a balanced representation. Indeed, Zhao et al. (2018) showed that this simple heuristic significantly reduces gender stereotyping in neural co-reference resolution systems, without harming system performance. Unfortunately, this approach

is only applicable to English and other languages with little morphological inflection. When applied to morphologically rich languages that exhibit gender agreement, it yields ungrammatical sentences.

The problem: inflected languages. Many languages, including Spanish and Hebrew, have gender inflections for nouns, verbs, and adjectives— i.e., the words in a sentence are marked with morphological endings that reflect the grammatical gender of the surrounding nouns.[2] This means that if the gender of one word changes, the others have to be updated to preserve morpho-syntactic agreement (Corbett, 2012). Consider the following example from Spanish, where we wish to transform Sentence (1) to Sentence (2). (Parts of words that mark gender are depicted in bold.) This task is not as simple as replacing *el* with *la*—*ingeniero* and *experto* must also be reinflected. Moreover, the changes required for one language are not the same as those required for another (e.g., verbs are marked with gender in Hebrew, but not in Spanish).

(1) ***El*** *ingenie**ro*** *alemán*
The.MSC.SG engineer.MSC.SG German.MSC.SG
es *muy expert**o***.
is.IN.PR.SG very skilled.MSC.SG
(The German engineer is very skilled.)

(2) ***La*** *ingenie**ra*** *aleman**a***
The.FEM.SG engineer.FEM.SG German.FEM.SG
es *muy expert**a***.
is.IN.PR.SG very skilled.FEM.SG
(The German engineer is very skilled.)

Our approach. Our goal is to transform sentences like Sentence (1) to Sentence (2) and vice versa. To the best of our knowledge, this task has not been studied previously. Indeed, there is no existing annotated corpus of paired sentences that could be used to train a supervised model. As a result, we take an unsupervised[3] approach using dependency trees, lemmata, part-of-speech (POS) tags, and morpho-syntactic tags from Universal Dependencies corpora (UD; Nivre et al., 2018). Specifically, we propose the following four-step process:

1. Analyze the sentence (including parsing, morphological analysis, and lemmatization).

---

[1] Specifically, we consider a noun to be animate if WordNet considers *person* to be a hypernym of that noun.

[2] The number of grammatical genders varies for different languages, with two being the most common non-zero number (Dryer and Haspelmath, 2013). The languages that we use in our evaluation have two grammatical genders (male, female).

[3] Because we do not have any direct supervision for the task of interest, we refer to our approach as being unsupervised even though it does rely on annotated linguistic resources.
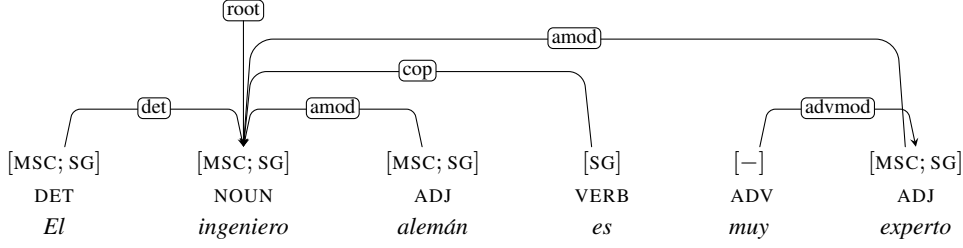
Figure 2: Dependency tree for the sentence *El ingeniero alemán es muy experto*.

2. Intervene on a gendered word.

3. Infer the new morpho-syntactic tags.

4. Reinflect the lemmata to their new forms.

This process is depicted in Fig. 1. The primary technical contribution is a novel Markov random field for performing step 3, described in the next section.

3 A Markov Random Field for Morpho-Syntactic Agreement

In this section, we present a Markov random field (MRF; Koller and Friedman, 2009) for morpho-syntactic agreement. This model defines a joint distribution over sequences of morpho-syntactic tags, conditioned on a labeled dependency tree with associated part-of-speech tags. Given an intervention on a gendered word, we can use this model to infer the manner in which the remaining tags must be updated to preserve morpho-syntactic agreement.

A dependency tree for a sentence (see Fig. 2 for an example) is a set of ordered triples (i, j, ℓ), where i and j are positions in the sentence (or a distinguished root symbol) and $\ell \in L$ is the label of the edge $i \to j$ in the tree; each position occurs exactly once as the first element in a triple. Each dependency tree T is associated with a sequence of morpho-syntactic tags $\mathbf{m} = m_1, \ldots, m_{|T|}$ and a sequence of part-of-speech (POS) tags $\mathbf{p} = p_1, \ldots, p_{|T|}$. For example, the tags $m \in M$ and $p \in P$ for *ingeniero* are $[\text{MSC}; \text{SG}]$ and NOUN, respectively, because *ingeniero* is a masculine, singular noun. For notational simplicity, we define $\mathcal{M} = M^{|T|}$ to be the set of all length-$|T|$ sequences of morpho-syntactic tags.

We define the probability of $\mathbf{m}$ given T and $\mathbf{p}$ as

$$\Pr(\mathbf{m} \mid T, \mathbf{p}) \propto \prod_{(i,j,\ell) \in T} \phi_i(m_i) \cdot \psi(m_i, m_j \mid p_i, p_j, \ell), \quad (1)$$

where the binary factor $\psi(\cdot, \cdot \mid \cdot, \cdot, \cdot) \geq 0$ scores how well the morpho-syntactic tags m_i and m_j

agree given the POS tags p_i and p_j and the label ℓ. For example, consider the amod (adjectival modifier) edge from *experto* to *ingeniero* in Fig. 2. The factor $\psi(m_i, m_j \mid \text{A}, \text{N}, \text{amod})$ returns a high score if the corresponding morpho-syntactic tags agree in gender and number (e.g., $m_i = [\text{MSC}; \text{SG}]$ and $m_j = [\text{MSC}; \text{SG}]$) and a low score if they do not (e.g., $m_i = [\text{MSC}; \text{SG}]$ and $m_j = [\text{FEM}; \text{PL}]$). The unary factor $\phi_i(\cdot) \geq 0$ scores a morpho-syntactic tag m_i outside the context of the dependency tree. As we explain in §3.1, we use these unary factors to force or disallow particular tags when performing an intervention; we do not learn them. Eq. (1) is normalized by the following partition function:

$$Z(T, \mathbf{p}) = \sum_{\mathbf{m}' \in \mathcal{M}} \prod_{(i,j,\ell) \in T} \phi_i(m'_i) \cdot \psi(m'_i, m'_j \mid p_i, p_j, \ell).$$

$Z(T, \mathbf{p})$ can be calculated using belief propagation; we provide the update equations that we use in App. A. Our model is depicted in Fig. 3. It is noteworthy that this model is delexicalized—i.e., it considers only the labeled dependency tree and the POS tags, not the actual words themselves.

3.1 Parameterization

We consider a linear parameterization and a neural parameterization of the binary factor $\psi(\cdot, \cdot \mid \cdot, \cdot, \cdot)$.

Linear parameterization. We define a matrix $W(p_i, p_j, \ell) \in \mathbb{R}^{c \times c}$ for each triple (p_i, p_j, ℓ), where c is the number of morpho-syntactic subtags. For example, $[\text{MSC}; \text{SG}]$ has two subtags MSC and SG. We then define $\psi(\cdot, \cdot \mid \cdot, \cdot, \cdot)$ as follows:

$$\psi(m_i, m_j \mid p_i, p_j, \ell) = \exp\left(\overline{m}_i^\top W(p_i, p_j, \ell) \overline{m}_j\right),$$

where $\overline{m}_i \in \{0, 1\}^c$ is a multi-hot encoding of m_i.

Neural parameterization. As an alternative, we also define a neural parameterization of $W(p_i, p_j, \ell)$ to allow parameter sharing among
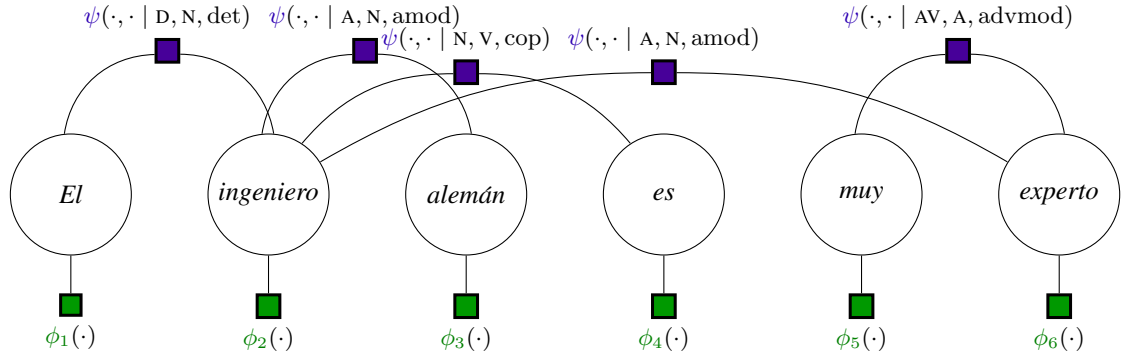
Figure 3: Factor graph for the sentence *El ingeniero alemán es muy experto*.

edges with different parts of speech and labels:

$$W(p_i, p_j, \ell) = \\ \exp\left(U \tanh(V [\mathbf{e}(p_i); \mathbf{e}(p_j); \mathbf{e}(\ell)])\right)$$

where $U \in \mathbb{R}^{c \times c \times n_1}$, $V \in \mathbb{R}^{n_1 \times 3n_2}$, and n_1 and n_2 define the structure of the neural parameterization and each $\mathbf{e}(\cdot) \in \mathbb{R}^{n_2}$ is an embedding function.

Parameterization of ϕ_i. We use the unary factors only to force or disallow particular tags when performing an intervention. Specifically, we define

$$\phi_i(m) = \begin{cases} \alpha & \text{if } m = m_i \\ 1 & \text{otherwise,} \end{cases} \qquad (2)$$

where $\alpha > 1$ is a strength parameter that determines the extent to which m_i should remain unchanged following an intervention. In the limit as $\alpha \to \infty$, all tags will remain unchanged except for the tag directly involved in the intervention.[4]

3.2 Inference

Because our MRF is acyclic and tree-shaped, we can use belief propagation (Pearl, 1988) to perform exact inference. The algorithm is a generalization of the forward-backward algorithm for hidden Markov models (Rabiner and Juang, 1986). Specifically, we pass messages from the leaves to the root and vice versa. The marginal distribution of a node is the point-wise product of all its incoming messages; the partition function $Z(T, \mathbf{p})$ is the sum of any node's marginal distribution. Computing $Z(T, \mathbf{p})$ takes polynomial time (Pearl, 1988)—specifically, $\mathcal{O}(n \cdot |M|^2)$ where M is the number of morpho-syntactic tags. Finally, inferring the highest-probability morpho-syntactic tag sequence $\mathbf{m}^\star$ given T and $\mathbf{p}$ can be performed using the max-product modification to belief propagation.

---

[4]In practice, α is set using development data.

Language	Accuracy	Language	Accuracy
French	93.17	Italian	98.29
Hebrew	95.16	Spanish	97.78

Table 1: Morphological reinflection accuracies.

3.3 Parameter Estimation

We use gradient-based optimization. We treat the negative log-likelihood $-\log(\Pr(\mathbf{m} \mid T, \mathbf{p}))$ as the loss function for tree T and compute its gradient using automatic differentiation (Rall, 1981). We learn the parameters of §3.1 by optimizing the negative log-likelihood using gradient descent.

4 Intervention

As explained in §2, our goal is to transform sentences like Sentence (1) to Sentence (2) by intervening on a gendered word and then using our model to infer the manner in which the remaining tags must be updated to preserve morpho-syntactic agreement. For example, if we change the morpho-syntactic tag for *ingeniero* from [MSC;SG] to [FEM;SG], then we must also update the tags for *el* and *experto*, but do not need to update the tag for *es*, which should remain unchanged as [IN; PR; SG]. If we intervene on the ith word in a sentence, changing its tag from m_i to m_i', then using our model to infer the manner in which the remaining tags must be updated means using $\Pr(\mathbf{m}_{-i} \mid m_i', T, \mathbf{p})$ to identify high-probability tags for the remaining words.

Crucially, we wish to change as little as possible when intervening on a gendered word. The unary factors ϕ_i enable us to do exactly this. As described in the previous section, the strength parameter α determines the extent to which m_i should remain unchanged following an intervention—the larger the value, the less likely it is that m_i will be changed.

Once the new tags have been inferred, the final step is to reinflect the lemmata to their new forms.

Language	Training Size	Annotated Test Size
Hebrew	5,241	111
Spanish	14,187	136
French	14,554	–
Italian	12,837	–

Table 2: Language data.

This task has received considerable attention from the NLP community (Cotterell et al., 2016, 2017). We use the inflection model of Wu et al. (2018). This model conditions on the lemma **x** and morpho-syntactic tag m to form a distribution over possible inflections. For example, given *experto* and [A; FEM; PL], the trained inflection model will assign a high probability to *expertas*. We provide accuracies for the trained inflection model in Tab. 1.

5 Experiments

We used the Adam optimizer (Kingma and Ba, 2014) to train both parameterizations of our model until the change in dev-loss was less than 10^{-5} bits. We set $\beta = (0.9, 0.999)$ without tuning, and chose a learning rate of 0.005 and weight decay factor of 0.0001 after tuning. We tuned $\log \alpha$ in the set $\{0.5, 0.75, 1, 2, 5, 10\}$ and chose $\log \alpha = 1$. For the neural parameterization, we set $n_1 = 9$ and $n_2 = 3$ without any tuning. Finally, we trained the inflection model using only gendered words.

We evaluate our approach both intrinsically and extrinsically. For the intrinsic evaluation, we focus on whether our approach yields the correct morpho-syntactic tags and the correct reinflections. For the extrinsic evaluation, we assess the extent to which using the resulting transformed sentences reduces gender stereotyping in neural language models.

5.1 Intrinsic Evaluation

To the best of our knowledge, this task has not been studied previously. As a result, there is no existing annotated corpus of paired sentences that can be used as "ground truth." We therefore annotated Spanish and Hebrew sentences ourselves, with annotations made by native speakers of each language. Specifically, for each language, we extracted sentences containing animate nouns from that language's UD treebank. The average length of these extracted sentences was 37 words. We then manually inspected each sentence, intervening on the gender of the animate noun and reinflecting the sentence accordingly. We chose Spanish and Hebrew because gender agreement operates differ-

		Tag			Form
	P	**R**	**F1**	**Acc**	**Acc**
Hebrew–BASE	**89.04**	40.12	55.32	86.88	83.63
Hebrew–LIN	87.07	62.35	72.66	90.5	**86.75**
Hebrew–NN	87.18	**62.96**	**73.12**	**90.62**	86.25
Spanish–BASE	**96.97**	51.45	67.23	90.21	86.32
Spanish–LIN	92.74	**73.95**	**82.29**	93.79	89.52
Spanish–NN	95.34	72.35	82.27	**93.91**	**89.65**

Table 3: Tag-level precision, recall, F_1 score, and accuracy and form-level accuracy for the baselines ("–BASE") and for our approach ("–LIN" is the linear parameterization, "–NN" is the neural parameterization).

ently in each language. We provide corpus statistics for both languages in the top two rows of Tab. 2.

We created a hard-coded $\psi(\cdot, \cdot \mid \cdot, \cdot, \cdot)$ to serve as a baseline for each language. For Spanish, we only activated, i.e. set to a number greater than zero, values that relate adjectives and determiners to nouns; for Hebrew, we only activated values that relate adjectives and verbs to nouns. We created two separate baselines because gender agreement operates differently in each language.

To evaluate our approach, we held all morpho-syntactic subtags fixed except for gender. For each annotated sentence, we intervened on the gender of the animate noun. We then used our model to infer which of the remaining tags should be updated (updating a tag means swapping the gender subtag because all morpho-syntactic subtags were held fixed except for gender) and reinflected the lemmata. Finally, we used the annotations to compute the tag-level F_1 score and the form-level accuracy, excluding the animate nouns on which we intervened.

Results. We present the results in Tab. 3. Recall is consistently significantly lower than precision. As expected, the baselines have the highest precision (though not by much). This is because they reflect well-known rules for each language. That said, they have lower recall than our approach because they fail to capture more subtle relationships.

For both languages, our approach struggles with conjunctions. For example, consider the phrase *él es un ingeniero y escritor* (*he is an engineer and a writer*). Replacing *ingeniero* with *ingeniera* does not necessarily result in *escritor* being changed to *escritora*. This is because two nouns do not normally need to have the same gender when they are conjoined. Moreover, our MRF does not include co-reference information, so it cannot tell that, in this case, both nouns refer to the same person. Note
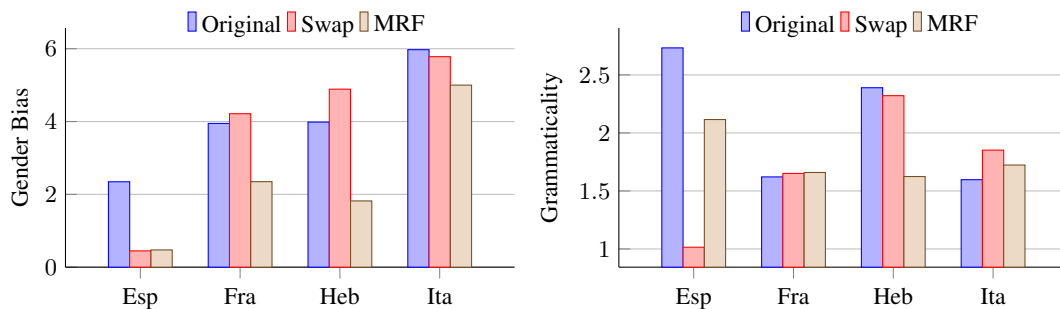
1655

Figure 4: Gender stereotyping (left) and grammaticality (right) using the original corpus, the corpus following CDA using naïve swapping of gendered words ("Swap"), and the corpus following CDA using our approach ("MRF").

that including co-reference information in our MRF would create cycles and inference would no longer be exact. Additionally, the lack of co-reference information means that, for Spanish, our approach fails to convert nouns that are noun-modifiers or indirect objects of verbs.

Somewhat surprisingly, the neural parameterization does not outperform the linear parameterization. We proposed the neural parameterization to allow parameter sharing among edges with different parts of speech and labels; however, this parameter sharing does not seem to make a difference in practice, so the linear parameterization is sufficient.

5.2 Extrinsic Evaluation

We extrinsically evaluate our approach by assessing the extent to which it reduces gender stereotyping. Following Lu et al. (2018), focus on neural language models. We choose language models over word embeddings because standard measures of gender stereotyping for word embeddings cannot be applied to morphologically rich languages.

As our measure of gender stereotyping, we compare the log ratio of the prefix probabilities under a language model P_{lm} for gendered, animate nouns, such as *ingeniero*, combined with four adjectives: *good*, *bad*, *smart*, and *beautiful*. The translations we use for these adjectives are given in App. B. We chose the first two adjectives because they should be used equally to describe men and women, and the latter two because we expect that they will reveal gender stereotypes. For example, consider

$$\log \frac{\sum_{\mathbf{x} \in \Sigma^*} P_{lm}(\text{BOS } \textit{El ingeniero bueno } \mathbf{x})}{\sum_{\mathbf{x} \in \Sigma^*} P_{lm}(\text{BOS } \textit{La ingeniera buena } \mathbf{x})}.$$

If this log ratio is close to 0, then the language model is as likely to generate sentences that start with *el ingeniero bueno* (*the good male engineer*) as it is to generate sentences that start with *la*

Language	No. Animate Noun Pairs	% of Animate Sentences
Hebrew	95	20%
Spanish	259	20%
Italian	150	10%
French	216	7%

Table 4: Animate noun statistics.

ingeniera bueno (*the good female engineer*). If the log ratio is negative, then the language model is more likely to generate the feminine form than the masculine form, while the opposite is true if the log ratio is positive. In practice, given the current gender disparity in engineering, we would expect the log ratio to be positive. If, however, the language model were trained on a corpus to which our CDA approach had been applied, we would then expect the log ratio to be much closer to zero.

Because our approach is specifically intended to yield sentences that are grammatical, we additionally consider the following log ratio (i.e., the grammatical phrase over the ungrammatical phrase):

$$\log \frac{\sum_{\mathbf{x} \in \Sigma^*} P_{lm}(\text{BOS } \textit{El ingeniero bueno } \mathbf{x})}{\sum_{\mathbf{x} \in \Sigma^*} P_{lm}(\text{BOS } \textit{El ingeniera bueno } \mathbf{x})}.$$

We trained the linear parameterization using UD treebanks for Spanish, Hebrew, French, and Italian (see Tab. 2). For each of the four languages, we parsed one million sentences from Wikipedia (May 2018 dump) using Dozat and Manning (2016)'s parser and extracted taggings and lemmata using the method of Müller et al. (2015). We automatically extracted an animacy gazetteer from WordNet (Bond and Paik, 2012) and then manually filtered the output for correctness. We provide the size of the languages' animacy gazetteers and the percentage of automatically parsed sentences that contain an animate noun in Tab. 4. For each sentence containing a noun in our animacy gazetteer, we created a copy of the sentence, intervened on
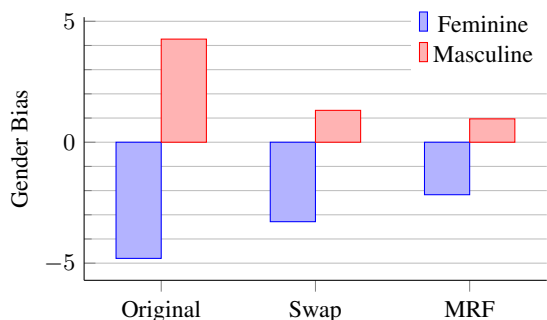
Figure 5: Gender stereotyping for words that are stereotyped toward men or women in Spanish using the original corpus, the corpus following CDA using naïve swapping of gendered words ("Swap"), and the corpus following CDA using our approach ("MRF").

Phrase	Original	Swap	MRF
1. *El ingeniero bueno*	-27.6	-27.8	-28.5
2. *La ingeniera buena*	-31.3	-31.6	-30.5
3. **El ingeniera bueno*	-32.2	-27.1	-33.5
4. **La ingeniero buena*	-33.2	-32.8	-33.6
Gender stereotyping	3.7	6.2	2
Grammaticality	3.25	0.25	4.05

Table 5: Prefix log-likelihoods of Spanish phrases using the original corpus, the corpus following CDA using naïve swapping of gendered words ("Swap"), and the corpus following CDA using our approach ("MRF"). Phrases 1 and 2 are grammatical, while phrases 3 and 4 are not (dentoted by "*"). Gender stereotyping is measured using phrases 1 and 2. Grammaticality is measured using phrases 1 and 3 and using phrases 2 and 4; these scores are then averaged.

the noun, and then used our approach to transform the sentence. For sentences containing more than one animate noun, we generated a separate sentence for each possible combination of genders. Choosing which sentences to duplicate is a difficult task. For example, *alemán* in Spanish can refer to either a German man or the German language; however, we have no way of distinguishing between these two meanings without additional annotations. Multilingual animacy detection (Jahan et al., 2018) might help with this challenge; co-reference information might additionally help.

For each language, we trained the BPE-RNNLM baseline open-vocabulary language model of Mielke and Eisner (2018) using the original corpus, the corpus following CDA using naïve swapping of gendered words, and the corpus following CDA using our approach. We then computed gender stereotyping and grammaticality as described above. We provide example phrases in Tab. 5; we provide a more extensive list of phrases in App. C.

Results Fig. 4 demonstrates depicts gender stereotyping and grammaticality for each language using the original corpus, the corpus following CDA using naïve swapping of gendered words, and the corpus following CDA using our approach. It is immediately apparent that our approch reduces gender stereotyping. On average, our approach reduces gender stereotyping by a factor of 2.5 (the lowest and highest factors are 1.2 (Ita) and 5.0 (Esp), respectively). We expected that naïve swapping of gendered words would also reduce gender stereotyping. Indeed, we see that this simple heuristic reduces gender stereotyping for some but not all of the languages. For Spanish, we also examine specific words that are stereotyped

toward men or women. We define a word to be stereotyped toward one gender if 75% of its occurrences are of that gender. Fig. 5 suggests a clear reduction in gender stereotyping for specific words that are stereotyped toward men or women.

The grammaticality of the corpora following CDA differs between languages. That said, with the exception of Hebrew, our approach either sacrifices less grammaticality than naïve swapping of gendered words and sometimes increases grammaticality over the original corpus. Given that we know the model did not perform as accurately for Hebrew (see Tab. 3), this finding is not surprising.

6 Related Work

In contrast to previous work, we focus on mitigating gender stereotypes in languages with rich morphology—specifically languages that exhibit gender agreement. To date, the NLP community has focused on approaches for detecting and mitigating gender stereotypes in English. For example, Bolukbasi et al. (2016) proposed a way of mitigating gender stereotypes in word embeddings while preserving meanings; Lu et al. (2018) studied gender stereotypes in language models; and Rudinger et al. (2018) introduced a novel Winograd schema for evaluating gender stereotypes in co-reference resolution. The most closely related work is that of Zhao et al. (2018), who used CDA to reduce gender stereotypes in co-reference resolution; however, their approach yields ungrammatical sentences in morphologically rich languages. Our approach is specifically intended to yield grammatical sentences when applied to such languages. Habash et al. (2019) also focused on morphologically rich

languages, specifically Arabic, but in the context of gender identification in machine translation.

7 Conclusion

We presented a new approach for converting between masculine-inflected and feminine-inflected noun phrases in morphologically rich languages. To do this, we introduced a Markov random field with an optional neural parameterization that infers the manner in which a sentence must change to preserve morpho-syntactic agreement when altering the grammatical gender of particular nouns. To the best of our knowledge, this task has not been studied previously. As a result, there is no existing annotated corpus of paired sentences that can be used as "ground truth." Despite this limitation, we evaluated our approach both intrinsically and extrinsically, achieving promising results. For example, we demonstrated that our approach reduces gender stereotyping in neural language models. Finally, we also identified avenues for future work, such as the inclusion of co-reference information.

Acknowledgments

References

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 4349–4357.

Francis Bond and Kyonghee Paik. 2012. A survey of WordNets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.

Jennifer Coates. 1987. *Women, Men and Language: A Sociolinguistic Account of Sex Differences in Language*. Longman.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press.

Greville G. Corbett. 2012. *Features*. Cambridge University Press.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Kate Crawford. 2013. The hidden biases in big data.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word embeddings quantify 100 years of gender and ethnic stereotypes. *CoRR*, abs/1711.08412.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In *Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing*, Florence, Italy.

Labiba Jahan, Geeticka Chauhan, and Mark Finlayson. 2018. A new approach to animacy detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1–12. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: Principles and techniques*. MIT Press.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *CoRR*, abs/1807.11714.

Sebastian J. Mielke and Jason Eisner. 2018. Spell once, summon anywhere: A two-level open-vocabulary language model. *CoRR*, abs/1804.08205.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Kamil Kopacewicz, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguy~ên Thị, Huy`ên Nguy~ên Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Judea Pearl. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers.

Lawrence R. Rabiner and Biing-Hwang Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16.

Louis B. Rall. 1981. *Automatic Differentiation: Techniques and Applications*, volume 120 of *Lecture Notes in Computer Science*. Springer.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14. Association for Computational Linguistics.

Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini. 2018. Biased embeddings from wild data: Measuring, understanding and removing. *CoRR*, abs/1806.06301.

Shijie Wu, Pamela Shapiro, and Ryan Cotterell. 2018. Hard non-monotonic attention for character-level transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4425–4438. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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. pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. pages 15–20.

A Belief Propagation Update Equations

Our belief propagation update equations are

$$\mu_{i\to f}(m) = \prod_{f'\in N(i)\setminus\{f\}} \mu_{f'\to i}(m) \qquad (3)$$

$$\mu_{f_i\to i}(m) = \phi_i(m)\,\mu_{i\to f_i}(m) \qquad (4)$$

$$\mu_{f_{ij}\to i}(m) = \sum_{m'\in M} \psi(m',m \mid p_i, p_j, \ell)\,\mu_{j\to f_{ij}}(m') \qquad (5)$$

$$\mu_{f_{ij}\to j}(m) = \sum_{m'\in M} \psi(m,m' \mid p_i, p_j, \ell)\,\mu_{i\to f_{ij}}(m') \qquad (6)$$

where $N(i)$ returns the set of neighbouring nodes of node i. The belief at any node is given by

$$\beta(v) = \prod_{f\in N(v)} \mu_{f\to v}(m). \qquad (7)$$

B Adjective Translations

Tab. 6 and Tab. 7 contain the feminine and masculine translations of the four adjectives that we used.

Adjective	French	Hebrew	Italian	Spanish
good	bonne	טובה	buona	buena
bad	mauvaise	רעה	cattiva	mala
smart	intelligente	חכמה	intelligenti	inteligente
beautiful	belle	יפה	bella	hermosa

Table 6: Feminine translations of *good*, *bad*, *smart*, *beautiful* in French, Hebrew, Italian, and Spanish

Adjective	French	Hebrew	Italian	Spanish
good	bon	טוב	buono	bueno
bad	mauvais	רע	cattivo	malo
smart	intelligent	חכם	intelligente	inteligente
beautiful	bel	יפה	bello	hermoso

Table 7: Masculine translations of *good*, *bad*, *smart*, *beautiful* in French, Hebrew, Italian, and Spanish

C Extrinsic Evaluation Example Phrases

For each noun in our animacy gazetteer, we generated sixteen phrases. Consider the noun *engineer* as an example. We created four phrases—one for each translation of *The good engineer*, *The bad engineer*, *The smart engineer*, and *The beautiful engineer*. These phrases, as well as their prefix log-likelihoods are provided below in Tab. 8.

Phrase	Original	Swap	MRF
El ingeniero bueno	-27.63	-27.80	-28.50
La ingeniera buena	-31.34	-31.65	-30.46
*El ingeniera bueno	-32.22	-27.06	-33.49
*La ingeniero buena	-33.22	-32.80	-33.56
El ingeniero mal	-30.45	-30.90	-30.86
La ingeniera mala	-31.03	-29.63	-30.59
*El ingeniera mal	-34.19	-30.17	-35.15
*La ingeniero mala	-33.09	-30.80	-33.81
El ingeniero inteligente	-26.19	-25.49	-26.64
La ingeniera inteligente	-29.14	-26.31	-27.57
*El ingeniera inteligente	-29.80	-24.99	-30.77
*La ingeniero inteligente	-31.00	-27.12	-30.16
El ingeniero hermoso	-28.74	-28.65	-29.13
La ingeniera hermosa	-31.21	-29.25	-30.04
*El ingeniera hermoso	-32.54	-27.97	-33.83
*La ingeniero hermosa	-33.55	-30.35	-32.96

Table 8: Prefix log-likelihoods of Spanish phrases using the original corpus, the corpus following CDA using naïve swapping of gendered words ("Swap"), and the corpus following CDA using our approach ("MRF"). Ungrammatical phrases are denoted by "*".