An Empirical Study on Gender Bias within an Indigenous Language Revitalization Perspective and an Inclusive NMT

Ngoc Tan Le
Université du Québec à Montréal
le.ngoc_tan@uqam.ca

Oussama Hansal
Université du Québec à Montréal
Oussama.Hansal@courrier.uqam.ca

Fatiha Sadat
Université du Québec à Montréal
sadat.fatiha@uqam.ca

Abstract

Natural Language Processing applications, such as Neural Machine Translation, typically exhibit substantial biases toward sensitive factors such as gender or race. This degrades the performance of machine translation and promotes unfavorable preconceptions. The current paper examines the issues and challenges of gender bias within Inuktitut, an under-represented Indigenous language of Canada, and discusses how to enhance the performance of Inuktitut-English NMT; all with the aim of revitalizing the Indigenous language and considering an inclusive NMT. Firstly, we performed the detection of gender bias in word embeddings in Inuktitut and English. Secondly, we compared the debiasing effect with the traditional word to vectors and also based on a dictionary. Then, we adopted a strategy, within the Inuktitut-English NMT, using the two bilingual debiased word embeddings.

1 Introduction

In recent years, studies on under-represented languages within NLP and AI, such as the Indigenous and Endangered languages have drawn a number of scholars. For example, some researchers have focused their studies on new conceptualisation of language revitalisation (Grenoble and Whaley, 2006; Pine and Turin, 2017) and others proposed revitalisation strategies to strengthen communities and promote commitment (Wiltshire et al., 2022).

This can be considered as a promising factor for the development of language technologies for this category of languages. However, the complex morphology of these under-represented languages, as well as lack of resources and the presence of biases, have been regarded as serious barriers.

Gender bias, in particular, may be described as a prejudice toward one gender over the other. Bias can vary from the usage of gender defaults to the association between occupation and gender. Since language technologies become more widely used and implemented, their social influence creates issues both intrinsically and extrinsically (Hovy and Spruit, 2016; Dastin, 2018; Sun et al., 2019).

Many NLP tasks are trained using collected human language data. These applications are likely to display biases in various ways such as data, annotation, input representations, models, and research design (Hovy and Prabhumoy, 2021). To analyze the context, NLP research was investigated by Sun et al. (2019). Their research, however, is centred on monolingual applications, and the underlying hypothesis and method may not be applicable to languages other than English. Gender stereotypes have therefore been defined differently from research to study, depending on the language used and the elements considered.

As an illustration, Google Translate can translate the following statement from English to French [1], and to Inuktitut[2] :

(en) **The developer** disagreed with **the designer** because **she** did not like **her** design.

(fr) **Le développeur** n'était pas d'accord avec **le concepteur** car **elle** n'aimait pas **son** design.

(iu) **Sanaji** angiqatiqalaut**tuq titiraujaqtimit** piugilaum**magu** titiraujaqsimaninga.

We observe gender stereotypes that classified the first subject "*developer*" and the second subject "*designer*" in the masculine category, despite the fact that we used *she, her* to focus on the feminine gender. Such flaws not only challenge the development of NLP applications for under-represented languages, but they also intensify existing biases.

In Inuktitut, we notice another linguistic problem in terms of gender, since the language is gender-

[1] https://translate.google.ca/, consulted on November 18th, 2022

[2] https://www.bing.com/translator?from=pt-pt&to=iu&setlang=be, consulted on November 18th, 2022

less and does not identify a masculine or feminine category, compared to European languages.

Actually, Inuktitut people determine rather another concept of animacy, which means whether a noun is considered animate or inanimate[3]. For example, animate nouns are most obviously related to humans and animals, but other objects, such as river, tree, are considered inanimate (Hassan, 2015).

Given that word embeddings are often employed in NLP tasks, understanding how human biases are absorbed into them might help us understand bias in NLP models.

This study aims to investigate the gender bias challenges by assessing current efforts to identify and minimise gender stereotypes. Furthermore, we present an empirical case study of Inuktitut-English NMT, in which Inuktitut is considered as an under-represented language. Inuktitut is considered as an Indigenous language of Eastern Canada and also the official language of the Nunavut government. The main goal also contributes to the efforts on the revitalization and preservation of Indigenous languages.

This paper is organised as follows: Section 2 presents the relevant work. Section 3 introduces linguistic challenges in Indigenous languages, especially in Inuktitut. Sections 4 presents several detection and mitigation methods to deal with gender bias. Sections 5 presents an empirical case study of the Machine Translation downstream task on gender bias. The experiments and evaluations are presented in Sections 6 and 7 with other state-of-the-art approaches; following an error analysis and a discussion. Finally, Section 8 gives some conclusions and presents future research directions.

2 Related work

Interest in identifying, measuring, and mitigating gender bias in NLP continues to grow, with recent research demonstrating how gender differences influence language technology (Cislak et al., 2018). Recently, several approaches have been proposed in order to identify gender bias in NLP. They are classified in two main approaches, depending on what they are based on; (1) traditional word embeddings or (2) dictionary-based embeddings. In the following subsections, we give some details on the state-of-the-art of these two approaches and also that of methods for measuring gender bias.

Approach of using traditional embeddings

This technique uses some term list, i.e. occupation for male and female. The main component of a word vector consisting of gender defining word pairs defines gender directions, such as *she* and *he* (Bolukbasi et al., 2016). Besides the use of monolingual word embeddings, bilingual word embeddings are used for gender bias identification task (Liu et al., 2019), to map with similar words in related languages in the same task (Alipour et al., 2022); multilingual word embeddings (Zhao et al., 2020; Bansal et al., 2021); and contextual word embeddings with ELMo or BERT (Kurita et al., 2019; Zhao et al., 2019).

Approach of using dictionary-based embeddings

Existing debiasing algorithms usually need a pre-compiled list of seed words to indicate the bias direction, which is followed by the removal of biased information. Kaneko and Bollegala (2021) proposed a method for debiasing pre-trained word embeddings using dictionaries, without requiring access to the original training resources or any knowledge regarding the word embedding algorithms that was used. An et al. (2022) developed dictionary-guided loss functions, which promote word embeddings to be comparable to their relatively neutral Dictionary Definition (DD) representations. The authors proposed DD-GloVe, a train-time debiasing algorithm to learn word embeddings by leveraging dictionary definitions.

Besides, Ding et al. (2022) proposed a causal inference framework to leverage causal structure among bias and semantic components in order to remove gender bias.

Methods for measuring gender bias

In terms of measuring gender bias in NLP, the calculation is usually done at the profession level and at the corpus level for English, and then applied to other languages (Chaloner and Maldonado, 2019; Fabris et al., 2020; Gezici and Saygin, 2022; Jentzsch and Turan, 2022). Among other researches, Chen et al. (2021) reported measurements of gender bias in the Wikipedia corpora for nine languages, such as Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof, in the NLP pipeline.

[3] https://linguisticmaps.tumblr.com/post/169273617313/grammatical-gender-or-noun-class-categories-new

In terms of mitigating gender bias in NLP, especially in the downstream tasks, i.e. Machine Translation, the main goal consists of reducing the reliance on gender stereotypes, and also improving the translation quality (Sun et al., 2019; Stafanovičs et al., 2020; Liao, 2021; Chen et al., 2022; Kirtane and Anand, 2022).

In order to evaluate gender bias in NLP, multiple methods are proposed such as hard debias and soft debias (Bolukbasi et al., 2016). Moreover, Ravfogel et al. (2020) proposed iterative nullspace projection to evaluate on bias and fairness. Their proposed algorithm could mitigate bias in word embeddings. On the other hand, the WEAT (Word Embedding Association Test) bias detection method is usually performed for assessing word embedding bias (Caliskan et al., 2017).

3 Linguistic challenges of Inuktitut

3.1 Morphological segmentation in Inuktitut

The morphology of Indigenous languages in the Americas is very complex, with the majority being polysynthetic or agglutinative (Gasser, 2011; Littell et al., 2018; Joanis et al., 2020; Le and Sadat, 2020, 2022b).

Morphology segmentation is important in the learning of these languages. In this paper, we focus on Inuktitut, the most popular polysynthetic Indigenous language in Northern Canada, in the context of NLP research.

Polysynthetic languages, in principle, have lengthy sentence-words and a regular agglutinative and strongly suffixing morphology, which causes words to be exceedingly long and potentially unique (Lowe, 1985; Mithun, 2015). Besides, Inuktitut's morphophonemics are quite complicated.

In the following example, in Nunavut Inuktitut (Lowe, 1985), we observe no gender, only one long word corresponding to one sentence in English translation, and the verb root in the first position:

(iu) **tusaa**tsiarunnanngittualuu**JUNGA**

(en) **I** can't **hear** very well.

where *tusaa* means to *hear*, and *junga* means *I*.

We may highlight two relevant works in terms of developing a morphological segmenter for Inuktitut: (1) Uqailaut morphological analyzer (Farley, 2012), and (2) an Inuktitut Neural Network-based (NN) word segmenter (Le and Sadat, 2020).

The Uqailaut tool used a Finite-State Transducer-based method by combining several techniques such as grammar rules, linguistic knowledge and heuristics; while the NN word segmenter is trained using a set of rich features and by leveraging bicharacter-based and word-based pretrained embeddings from large-scale raw corpora. When used to perform word segmentation along with pretrained embeddings, NN-based techniques have demonstrated their usefulness. Thus, we adopted the second segmentation approach in order to prepare the experimental training data in Inuktitut.

4 Word embeddings for debiasing

Inspired by (Hansal et al., 2022) to analyse and mitigate bias in word embeddings, we adopted the three following methods to measure bias in word embeddings:

Hard debiasing (Bolukbasi et al., 2016)

The hard debiasing method aims to detect and reduce bias in word embeddings. Bolukbasi et al. (2016) presented a post-processing strategy for projecting gender-neutral words into a subspace orthogonal to the gender dimension specified by a list of gender-definitional terms (e.g., she, her, him, actor, driver), as cited in (Kaneko and Bollegala, 2021).

In particularly, it needs a set of gender-specific word pairs. Then the gender direction is computed as the difference vectors between the embeddings of the corresponding gender-definitional words.

Sent-Debias (Liang et al., 2020)

This method consists of the main following processes: (1) define bias attributes as a set of relevant words; (2) transform sentence representations using bias attributes; (3) compute bias subspace; and (4) project onto bias subspace, then remove biased sentences. As a result, the output is the general sentences debiased.

The Sent-Debias algorithm1 is described below.

Dictionary-based debiasing (Kaneko and Bollegala, 2021)

This method aims to debias pretrained word embeddings using a monolingual dictionary. It does not require the bias attributes to be annotated in the word pair lists or any prior knowledge.

Given a dictionary D containing the definition, and n-dimensional pretrained word embedding, the debiasing process is considered as the task of learning an encoder, $E(w; \theta_e)$, such that w desribes all

Algorithm 1 Sent-Debias algorithm (Liang et al., 2020).

```
To initialize pretrained sentence encoder $M_\theta$
To define bias attributes
To obtain words $D$ indicative of bias attributes
$S = Contextualize(D)$            ▷ words into sentences
for each element in $D$ do
    To get sentence representations
end for
To compute bias subspace V
for each new sentence representation $h$ do
    To project onto bias subspace $h_V$
    To subtract the projection $\widehat{h} = h - h_V$
end for
```

words in the vocabulary, that is trained to generate a debiased version of an input embedding. A loss function is computed in order to optimize the learning process. This computation is based both on the pre-trained word embeddings and on the unbiased dictionary definition of the term.

A decoder D_d allows to compute the encoded version of w, $E(w;\theta_e)$, using a parameter of θ_d and an objective function $J_d(w)$, as described in the following equation 1:

$$J_d(w) = \|s(w) - D_d(E(w;\theta_e);\theta_d)\|_2^2 \quad (1)$$

where $s(w)$ represents the dictionary-definitional vectors.

5 Methodology

5.1 Machine Translation downstream task

The purpose of this paper aims to investigate the impact of pretrained debiased word embeddings into an Inuktitut-English NMT system based on the Transformer encoder-decoder architecture (Vaswani et al., 2017).

5.2 The framework

Inspired by (Font and Costa-Jussa, 2019), we built an NMT framework by taking advantage of pretrained debiased word embeddings, and also source-target alignment information as an additional feature. Figure1 shows the architecture of the proposed architecture.

First, the pretrained debiased word embeddings are used to initialize the embedding layers of the NMT model, both in the encoder and the decoder.

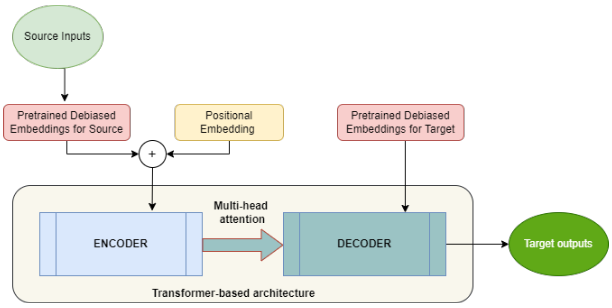
Source Inputs
Pretrained Debiased Embeddings for Source
Positional Embedding
Pretrained Debiased Embeddings for Target
+
Multi-head attention
ENCODER
DECODER
Target outputs
Transformer-based architecture

Figure 1: Architecture of our framework: Deep Learning-based debiased NMT for Indigenous language, with pretrained debiased word-based embedding for both source and target, combining with positional embedding.

We deal with the morphology complexity by applying the morpheme segmentation for Inuktitut (Le and Sadat, 2020, 2022a). Second, source-target alignment information are incorporated in the training step. We apply an unsupervised word aligner (Dyer et al., 2013) to generate symmetrical source-target alignments. Third, we inject in the decoding, the source-target morphological information, such as a bilingual lexicon. We apply lexicon extractor from Moses (Koehn et al., 2007) to prepare a bilingual lexical shortlist which is passed to the decoder.

We hypothesize that an ensemble of all the models of different types and architectures, with weights, could lead to an improved NMT performance. The following equation 2 of the objective function $f(x)$ helps in weighting all possible NMT models.

$$f(x) = \alpha Model_1 + \beta Model_2 + \theta Model_3 \quad (2)$$

where, $\alpha + \beta + \theta = 1$.

6 Evaluations

6.1 Data preparation

We performed experiments on gender bias mitigation in Inuktitut. The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (Joanis et al., 2020) is used to train and to evaluate our proposal.

As described in Table 1, the Inuktitut-English corpus contains 1,293,348 sentences pairs, 5,433 sentences pairs and 6,139 sentences pairs for the training, development and testing sets, respectively. Regarding the pretraining of word embeddings, we applied a setting of the hyper-parameters, as described in Table 2, and used *fastText* toolkit to pretrain them, as proposed by Bojanowski et al. (2017).

Dataset	Train set	Dev set	Test set
Inuktitut (iu)	1,293,348	5,433	6,139
English (en)	1,293,348	5,433	6,139

Table 1: Statistics of Nunavut Hansard for Inuktitut-English (Joanis et al., 2020).

Hyper-parameters
Epochs = *50*
Dimension size = *300*
Window size = *2*
Alpha value = *0.03*
Loss function = *softmax*

Table 2: Setting of the hyper-parameters for embedding pretraining.

6.2 Neural Machine Translation

Our experiments on NMT using the Transformer-based architecture (Vaswani et al., 2017) are described as follows:

(1) System 1: Baseline of Joanis et al. (2020) without pretrained debiased embeddings.

(2) System 2: Transformer-based model with only word alignment information as additional feature.

(3) System 3: Transformer-based model with only pretrained debiased embeddings.

(4) System 4: Transformer-based model with debiased embeddings and word alignment information as additional feature.

The configuration of the experimental environment is described in Table 3 and the relevant hyper-parameters of the NMT models are shown in Table 4.

Environment	Configuration
Operating platform	CUDA 11
Operating System	Ubuntu
Memory	32 GB
multi-GPU	6 cores
Python version	python 3.8
Tensorflow version	v2.10.0

Table 3: Configuration of the experimental environment

Hyper-parameter	Value
Maximum sentence length	128
Batch size	64
Transformer layers	12
Transformer hidden layers	768
Learning rate	0.0001
Epoch	50
Optimizer	adam

Table 4: Setting of hyper-parameters for NMT models

7 Evaluations

7.1 Results on bias mitigation

The WEAT evaluation is performed on the altered terms list which is translated into Inuktitut. We observe strong impact sizes across all standard word embeddings and several tests are significant at various levels. The WEAT results shows impact sizes on gendered tests, where a large impact size on debiased word embeddings is observed from the original models. Moreover, our evaluations show that the dictionary-based debiasing method outperforms other methods, as shown in Table 5. It effectively removes unfair biases encoded in pre-trained word embeddings, while retaining meaningful semantics.

Method	Debiased WEAT
Baseline	0.034
Hard debiasing	0.385
Sent debiasing	0.377
Dictionary-based debiasing	0.527

Table 5: The evaluation of WEAT using fastText toolkit, with significance of p-value < 0.05, against the WEAT baseline value = 0.034.

As Inuktitut is a complex-gender language, using pronouns might be challenging. Common names are utilised for males and females rather than particularly gendered terminology to identify the male and female groups (Goldfarb-Tarrant et al., 2020). We performed three tests to examine male and female name correlations to job occupations and family, art words and science.

In accordance with the projection in Figure 2, we notice that the significant association between the groups is no longer present in the tests. The experimental results show that the classes are no longer sequentially separated. This behaviour differs significantly from the sent debias and hard debias approaches, which have been found to preserve a significant amount of the closeness between female and male-biased vectors.

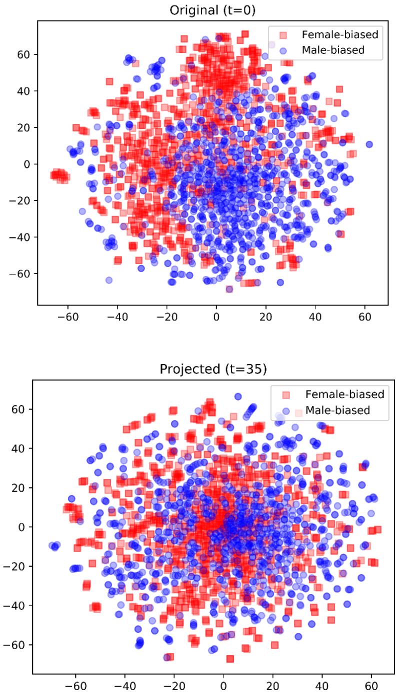

Figure 2: Projection with t-distributed stochastic neighbor embedding about different clusters between female and male biased states, at t=0 for original state and t=35 for projected state.

7.2 Results on Neural Machine Translation

We used automatic evaluation metrics such as SacreBLEU (Post, 2018) [4] for BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002), chrF++ (Popović, 2015) [5] for calculating character n-gram F-score, and translation error rate (TER).

As shown in Table 6, in the Inuktitut-English direction, systems 1, 2, and 4 outperformed the base system, with significant gains from +0.93 to +3.03 points BLEU. In the other hand, Table 7, which related to the English-Inuktitut direction, shows the best scores for System 4, with 20.5, 48, and 62.3 in terms of BLEU, chrF++, and 62.3 respectively. In both directions, we notice negative impacts on the use of debiased word embedding while injecting into NMT models. Especially, the system 3 (Table 6), with only pretrained debiased word embeddings for initialization phase, obtained 32.76 BLEU point against 35.00 BLEU point of the baseline, so a decrease of −2.24 point.

Experiment	BLEU	chrF++	TER
System 1 (base)	35.00	63.1	53.3
System 2	35.93	64.2	53.2
System 3	32.76	55.3	56.3
System 4	36.61	67.5	52.6

Table 6: Performances on Inuktitut-English NMT in terms of lowercase word BLEU score.

Experiment	BLEU	chrF++	TER
System 1	16.5	30.5	70.4
System 2	19.30	42.2	66.5
System 3	18.34	34.6	68.1
System 4	20.5	48.0	62.3

Table 7: Performances on English-Inuktitut NMT in terms of lowercase word BLEU score. We consider the system 1 as baseline.

On the other hand, with an ensemble of others models, combining both the alignment feature and debiased word embeddings, BLEU score has been improved, in all the systems, with 1.61 BLEU more, and 4 BLEU compared to the baseline, in the direction of Inuktitut-English and in the direction of English-Inuktitut, respectively.

Comparing the evaluation at the n-gram character level (Tables 6 and 7), we notice a similarity between the models, with regard to the chrF++ scores, that seems slightly more efficient than the baseline, except for System 3 that shows a drop of 0.55 point (Tables 6).

Additionally, in terms of translation error rate (TER) reductions, all systems performed better than the baseline, while applying alignment information as an additional feature, or combining both alignment information and pre-trained debiased embeddings. In the next subsection, we discuss error analysis and all possible causes.

7.3 Error analysis and discussion

Identifying the true gender direction in word embeddings is always challenging. We found a significant effect in traditional embeddings, which can be considered positive if the embeddings used ensure a more gender-neutral approach.

In addition, we have noticed, in the context of gender bias, a drawback which consists in the dependence of all 3 debiasing methods presented, like other machine learning approaches for that matter, on the data provided to it.This assumes that the training data is large enough and sampled from the

[4] https://github.com/mjpost/sacrebleu
[5] https://github.com/m-popovic/chrF

same distribution as the test data.

Another finding we made is that a large omission of masculine pronouns (less) is present in the MT outputs, compared to feminine pronouns (many). Alternatively, we notice that Inuktitut is a non-gendered language as pronouns might be dropped in translation outputs. This phenomenon makes the issue of gender or racial bias more difficult to manage in NMT systems in conjunction with underrepresented polysynthetic languages.

Statistics of errors found in accordance with the pronouns including *he, him, his, she, her* are shown in Tables 8 and 9 below.

For the NMT downstream task, we observed a decrease in the performance when initializing embedding layers with pretrained debiased embeddings. The plausible causes are related to the limited vocabulary size of pretrained embeddings. In contrast, using an ensemble of all the models outperformed all other NMT systems, with 36.61 BLEU score, and could have a better coverage of vocabulary for the model training.

Exp	he	him	his	she	her
System 1	46.79	26.32	62.04	13.89	16.13
System 2	41.28	15.79	62.04	19.44	14.52
System 3	45.87	42.11	58.33	22.22	11.29
System 4	48.62	31.58	59.26	8.33	14.52

Table 8: Precision on Inuktitut-English NMT in terms of pronouns found in accordance with {he, him, his, she, her}.

Exp	he	him	his	she	her
System 1	27.52	15.79	35.19	13.89	9.68
System 2	45.87	42.11	59.26	19.44	14.52
System 3	41.28	26.32	58.33	22.22	11.29
System 4	48.62	47.37	59.26	22.22	30.65

Table 9: Precision on English-Inuktitut NMT in terms of pronouns found in accordance with {he, him, his, she, her}.

8 Conclusion and perspectives

This research reveals that gender prejudice occurs in Inuktitut as in other languages. In this empirical study, we demonstrated how methodologies used to measure and minimise biases in English embeddings can be adapted to Inuktitut embeddings by accurately translating the data and taking into consideration the language's distinctive features. Furthermore, in Inuktitut-English NMT framework, we suggested a technique that combines bilingual debiased word embeddings with source-target alignment information.

As a future direction, we intend to examine other types of biases in Inuktitut, in close collaboration and in partnership with the indigenous community. Our primary goal is to help revitalize and preserve Indigenous languages in Canada through the use of NLP and machine learning technology and thus contribute to the effort of an inclusive AI. We hope that these preliminary results will inspire future studies on Indigenous and Endangered languages.

Acknowledgements

The authors wish to acknowledge the time and effort devoted by members of the Indigenous communities for their guidance in this research effort. We are also grateful to the anonymous reviewers for their valuable comments and feedbacks.

References

Ghafour Alipour, Jamshid Bagherzadeh Mohasefi, and Mohammad-Reza Feizi-Derakhshi. 2022. Learning bilingual word embedding mappings with similar words in related languages using gan. *Applied Artificial Intelligence*, pages 1–20.

Haozhe An, Xiaojiang Liu, and Donald Zhang. 2022. Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152, Dublin, Ireland. Association for Computational Linguistics.

Srijan Bansal, Vishal Garimella, Ayush Suhane, and Animesh Mukherjee. 2021. Debiasing multilingual word embeddings: A case study of three indian languages. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 27–34.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–

32, Florence, Italy. Association for Computational Linguistics.

Xiuying Chen, Mingzhe Li, Rui Yan, Xin Gao, and Xiangliang Zhang. 2022. Unsupervised mitigating gender bias by character components: A case study of chinese word embedding. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 121–128.

Yan Chen, Christopher Mahoney, Isabella Grasso, Esma Wali, Abigail Matthews, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender bias and under-representation in natural language processing across human languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 24–34.

Aleksandra Cislak, Magdalena Formanowicz, and Tamar Saguy. 2018. Bias against research on gender bias. *Scientometrics*, 115(1):189–200.

Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications.

Lei Ding, Dengdeng Yu, Jinhan Xie, Wenxing Guo, Shenggang Hu, Meichen Liu, Linglong Kong, Hongsheng Dai, Yanchun Bao, and Bei Jiang. 2022. Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11864–11872.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, 57(6):102377.

Benoit Farley. 2012. The uqailaut project. *URL http://www. inuktitutcomputing. ca.*

Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Michael Gasser. 2011. Computational morphology and the teaching of indigenous languages. In *Indigenous Languages of Latin America Actas del Primer Simposio sobre Enseñanza de Lenguas Indígenas de América Latina*, page 52.

Gizem Gezici and Yucel Saygin. 2022. Measuring gender bias in educational videos: A case study on youtube. *arXiv preprint arXiv:2206.09987*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Lenore A. Grenoble and Lindsay J. Whaley. 2006. Saving languages: An introduction to language revitalization. In *Cambridge: Cambridge University Press*.

Oussama Hansal, Ngoc Tan Le, and Fatiha Sadat. 2022. Indigenous language revitalization and the dilemma of gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 244–254, Seattle, Washington. Association for Computational Linguistics.

Jenna Hassan. 2015. De-colonizing gender in indigenous language revitalization efforts. In *Western Papers in Linguistics*, volume 1, Issue 2. Master's Major Research Papers and Proceedings of WISSLR 2015.

Dirk Hovy and Shrimai Prabhumoy. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in bert-measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The nunavut hansard inuktitut english parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Masahiro Kaneko and Danushka Bollegala. 2021. Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223, Online. Association for Computational Linguistics.

Neeraja Kirtane and Tanvi Anand. 2022. Mitigating gender stereotypes in hindi and marathi. *arXiv preprint arXiv:2205.05901*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume* *proceedings of the demo and poster sessions*, pages 177–180.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Ngoc Tan Le and Fatiha Sadat. 2020. Low-resource NMT: an empirical study on the effect of rich morphological word segmentation on Inuktitut. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 165–172, Virtual. Association for Machine Translation in the Americas.

Ngoc Tan Le and Fatiha Sadat. 2022a. Towards a low-resource neural machine translation for indigenous languages in Canada. *Journal TAL, special issue on Language Diversity*, 62:3:39–63.

Tan Ngoc Le and Fatiha Sadat. 2022b. Towards a low-resource neural machine translation for indigenous languages in canada. In *Journal TAL, special issue on Language Diversity*, volume 62, pages 39–63.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

Yuxin Liao. 2021. *Gender Bias in Neural Machine Translation*. Ph.D. thesis, University of Pennsylvania.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xuebo Liu, Derek F. Wong, Yang Liu, Lidia S. Chao, Tong Xiao, and Jingbo Zhu. 2019. Shared-private bilingual word embeddings for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3613–3622, Florence, Italy. Association for Computational Linguistics.

Ronald Lowe. 1985. *Basic Siglit Inuvialuit Eskimo Grammar*, volume 6. Inuvik, NWT: Committee for Original Peoples Entitlement.

Marianne Mithun. 2015. Morphological complexity and language contact in languages indigenous to north america. *Linguistic Discovery*, 13(2):37–59.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Aidan Pine and Mark Turin. 2017. Language revitalization. In *Oxford research encyclopedia of linguistics, Oxford. University Press*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Brandon Wiltshire, Steven Bird, and Rebecca Hardwick. 2022. Understanding how language revitalisation works: a realist synthesis. pages 1–17. Journal of Multilingual and Multicultural Development.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. 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 NAACL: Human Language Technology Conference, Volume 1*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.