Science and Technology Journal Vol. 13 Issue: 1 Jan 2025 ISSN: 2321-3388

Handling Gender Bias in Neural Machine Translation: A Focus on English-Khasi Language Pair

Aiusha Vellintihun Hujon[*,1,#], Thoudam Doren Singh[2], and Khwairakpam Amitab[3]
[1,3]Department of Information Technology, North Eastern Hill University, Shillong, India
[2]Department of Computer Science and Engineering, National Institute of Technology Meghalaya Meghalaya, India
[#]Department of Computer Science, St. Anthony's College, Shillong, India
E-Mail : avhujon@gmail.com[1], thoudam.doren@gmail.com[2], khamitab@gmail.com[3]
*Corresponding Author

Abstract—Machine Translation systems have progressed tremendously over the years and have enhanced cross-lingual communication. Although machine translation has advanced from a rule-based to a neural machine translation method, a new issue occurring in most machine translation systems, such as gender bias, cannot be ignored. Gender bias not only affects translation accuracy but also has an impact on societal prejudices and often promotes inequalities and stereotypes. This study attempts to address the challenges of gender bias in neural machine translation for English-Khasi, a low-resource language pair, where data scarcity increases the risk of biased translations. To the best of our knowledge, this is the first attempt to report a study on gender bias in the neural machine translation task of English-Khasi language pair. We use two different methods; a data augmentation technique and a transfer learning method tailored to the linguistic and socio-cultural characteristics of the target language. To implement the two methods, we manually build a sizeable gender-balanced English-Khasi parallel corpora to handle gender bias in English-Khasi neural machine translation systems. Through empirical evaluation of low-resource language pairs, English-Khasi, we demonstrate the effectiveness of the transfer learning approach in reducing gender bias while maintaining translation quality.

Keywords: Gender bias information, Khasi, Data augmentation, Transfer learning, Neural machine translation.

INTRODUCTION

In an age of rapidly advancing technologies, machine translation has become essential in bridging linguistic gaps and enabling communication across cultural domains. Gender bias, a subtle yet significant issue persists beneath this technological progress, which is a serious concern about fairness and the inclusiveness of societal prejudices. Gender bias in machine translation refers to the tendency of translation algorithms to produce gender-related imbalances, which often occur due to the morphological inflection of nouns and pronouns while translating a source language to a target language. This issue is a growing concern that can affect the fairness of machine translation systems and therefore presents a significant challenge that needs to be addressed.

Neural machine translation (NMT) systems rely on word embedding (Espana-Bonet *et al.*, 2017; Sennrich and Haddow, 2016) in which a sentence is represented in the form of vectors of real numbers. These vectors can capture the semantic relationship between words in sentences, thus

https://doi.org/10.22232/stj.2025.13.01.17

allowing an NMT system to translate each word efficiently. Word embedding is often trained on large corpora of text data. Training data can contain inherent biases in a society where certain occupations or roles can be predominantly associated with a specific gender. The system learns these biases from the training data, which can lead to a biased representation in the word embedding. The biased word embedding used by the NMT system can propagate the gender biases learned throughout translation. The system may learn to associate certain words with the specific genders based on biased word embedding, thus leading to biased translations.

NMT for gendered languages in which different words or forms are used based on the gender of the subject, can amplify gender bias in translations and reinforce stereotypes. This phenomenon usually occurs if one of the language pairs is a gendered language, such as English-Spanish (Escud´e Font and Costa-juss`a, 2019) or English-Khasi. Spanish and Khasi are both gendered languages, whereas English is not a gendered language. Thus, cases of gendered pronoun errors often occur in NMT tasks of such language pairs. An occurrence of gender disagreement in English-Khasi NMT (Hujon *et al.*, 2024) was recently reported in our previous work. One of the major causes of this error is the inadequate information required to aid in the correct translation of gendered pronouns in the parallel corpora. The results often exhibit gender stereotypes, which occur for various occupations, such as 'Doctor', is translated as a masculine pronoun, whereas 'Nurse' is often translated as a feminine pronoun. An example, 'Jane is a good doctor and she works in the village hospital' is often translated as 'Ka Jane ka dei *u* doktor ba bha bad *u* trei ha ka hospital shnong' with masculine *u* instead of the feminine *ka*.

Existing datasets like Winobias (Zhao *et al.*, 2018) and Winogender(Rudinger *et al.*, 2018) are available for a few high-resource languages for performing gender bias studies. Few works on NMT for English-Khasi language pair(Hujon *et al.*, 2024) (Singh and Hujon, 2020; Hujon *et al.*, 2023a) are reported, however, no report is found on gender bias. A dataset for the Khasi language that can aid in the English-Khasi NMT for addressing the issue of gender bias does not exist so far. This paper aims to delve into this issue and attempt to reduce gender bias in the machine translation of English to Khasi. An English-Khasi gender-balanced dataset is specifically created manually and experimented with using two different methods to handle gender bias. The NMT systems are evaluated using quantitative and qualitative metrics. The empirical results demonstrate that these methods can significantly reduce gender bias in the English-Khasi NMT.

The significant contributions of this paper are:

1. A sizeable parallel corpus is built for English-Khasi which is manually digitized and aligned.
2. A sizeable gender-balanced parallel corpus is built for English-Khasi which is manually translated and aligned.
3. The transfer learning method and data augmentation method are adapted and implemented for handling gender bias in the NMT systems of English-Khasi.
4. Quantitative and qualitative evaluation and analysis of the two methods are performed.

The paper's structure is as follows: Section 2 delves into the related work, while Section 3 outlines the research methods. Furthermore, Section 4 explains the experimental setup, Section 5 analyzes and discusses the experimental results of the various models, and finally, Section 6 presents the conclusion.

RELATED WORKS

Over the years, numerous works have contributed to the advancement of NMT. With the introduction of encoder-decoder architecture (Sutskever *et al.*, 2014), a crucial foundation was made for NMT. Based on this architecture, various works are reported for many language pairs across the globe. Subsequent research to improve translation accuracy and multiple aspects of NMT is performed and reported, such as the attention mechanism (Bahdanau *et al.*, 2016). Recently, transfer learning has emerged as a new approach for NMT (Zoph *et al.*, 2016), especially in dealing with issues of translation accuracy for low-resource languages. Another report using transfer learning has shown that transfer learning can also be applied to languages that are not linguistically related (Kocmi and Bojar, 2018).

Some existing works have attempted to mitigate gender bias in machine translations. One of the early works reported is a gender-enhanced NMT (Vanmassenhove *et al.*, 2018). The technique applied is by tagging the dataset in which gender information is injected into the dataset to reduce gender bias. The experiments mainly focused on language pairs that express grammatical gender. The results report significant improvements over state-of-the-art baseline systems. Word embedding is found to be another cause of gender bias in machine translations. The report shows that

word embedding can be severely biased, and a technique by debiasing the word embedding has been shown to reduce gender bias (Bolukbasi *et al.*, 2016a) significantly. Another work was reported using debiasing of word embeddings as a technique to handle gender bias in neural machine translations (Bolukbasi *et al.*, 2016b). Algorithms for debiasing and metrics to quantify direct and indirect gender biases in embeddings are implemented with significant improvement. The issue of gender bias also occurs in text emotion detection tasks. A report in line with this used a data augmentation method (Odbal *et al.*, 2022) where the gendered words are swapped using a bidirectional dictionary of gender pairs and a method of adversarial training, which showed a significant result in mitigating gender bias. Another report (Costa-juss`a *et al.*, 2022) explores gender bias in neural machine translation systems and investigates how different multilingual architectures can affect bias. For the study, two methods are proposed; one is based on word embedding, and the second is based on the attention mechanism. Gender information is encoded by using an SVM to classify occupations into three groups: male, female, and neutral. The systems are trained with the Europarl datasets for four languages: English, Spanish, French, and German. The results are analyzed based on BLEU and gender accuracy. The result shows that gender bias can occur in the source embedding, and algorithms of the implemented systems can also amplify the problem. A study on machine bias in terms of gender by Google Translate was performed and reported (Prates *et al.*, 2020). It reported a comprehensive study of twelve different languages. A set of sentences was prepared using the format 'He\She is an < *Job position* >', where < *job position* > is replaced by different job positions for both males and females. These sentences were created for twelve gendered neutral languages and translated to English using Google Translate. The report shows that Google Translate produces outputs that are predominantly male-dominated for most job positions. The study on the distribution of translated gender pronouns for each occupation category shows that 71.624% are male pronouns for Science, Technology, Engineering, and Mathematics occupations which are grouped under the STEM occupation category whereas only 4.219% are female pronouns and 11.181% are neutral pronouns. Another report on the study of a gender translation error in NMT systems (Wisniewski *et al.*, 2022) presented the results of the analyses of the flow of gender information between the Encoders and Decoders. A control test set consisting of 3394 gender-balanced parallel sentences for the French-English language pair is used. The sentences are created using the following template:

Type 1: (DET) (N) a termin´e son travail.

Type 2: The (N) has finished (PRO) work.

In the above template, (N) is the noun based on occupations, (DET) is the determiner for the French language, while (PRO) is the English possessive pronoun. The experiments and study are performed on the French-English parallel corpus from the WMT'15 'News' task (Bojar *et al.*, 2015). The method to mitigate gender bias through margin-augmented training loss has shown a significant result in improving the information flow between source and target in NMT systems.

Winobias (Zhao *et al.*, 2018) is a dataset designed to explore gender bias in terms of coreference resolution. The dataset consists of 3160 sentences covering around 40 stereotype occupations created in Winograd-style (Rahman and Ng, 2012) in two forms as given in the templates given below:

Type 1 : (entity1) (interacts with) (entity2) (conjunction)(pronoun) (circumstances)

Type 2: (entity1) (interacts with) (entity2) and then (interacts with) (pronoun) for (circumstances)

Winogender (Rudinger *et al.*, 2018) is another dataset which is also designed to resolve coreferences in gender bias. It is similar to Winobias but additionally includes gender-neutral pronouns.

One of the existing works on the English-Khasi language pairs (Hujon *et al.*, 2024) reported on neural machine translation systems using Long-Short-Term-Memory (LSTM), Gated Recurrent Unit (GRU), and transformer. The evaluation is performed on two different test data, within the domain and out of the domain. The study is performed on the English-Khasi dataset, which consists of 41529 parallel sentences. Considering the close linguistic relationship between languages of the same classification, Vietnamese, which belongs to the same group of Austroasiatic languages as the Khasi language, is chosen for the study to improve the translation accuracy. The NMT system achieved a superior performance in comparison to all the other models in the quantitative and qualitative evaluation results. The report also discusses its findings on gender disagreement observed in the outputs of the various NMT systems implemented. Transfer learning methods have shown promising results for low-resource languages (Zoph *et al.*, 2016; Nguyen and Chiang, 2017; Kocmi and Bojar, 2018), and reports show that when the parent models are trained on a high resource language pair and the child model are trained on the low

resource language pair, appears to have a positive impact on the performance of the child model. Another NMT system is also reported using an English-French language pair on the parent model and an English-Khasi language pair on the child model (Hujon *et al.*, 2023b) with significant results.

METHOD

Recurrent neural networks (RNN) have been successful in many neural machine translation tasks for various language pairs. However, recently with the introduction of the transformer (Vaswani *et al.*, 2017), NMT systems have achieved more accurate translations. The central core of the state-of-the-art transformer architecture is the self-attention mechanism and the positional encoding.

$$PE_{(pos2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d\,\text{model}}}}\right) \quad (1)$$

$$PE_{(pos2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d\,\text{model}}}}\right) \quad (2)$$

This mechanism allows the model to capture long-range dependencies and focus on essential words in a sentence to generate the translated output. The positional encoding is performed by using a *cos* function as in Equation 1 and *sin* function as in Equation 2 on the odd index and even index of the input vector, respectively.

$$\textit{Multi– Head (output)} = \textit{Concant}\ (z_0, z_2, z_3, z_4, z_5, z_6, z_7) = x, d_{\text{model}} \quad (3)$$

$$\textit{Attention}\ (Q, K, V) = \textit{softmax}\left(\frac{QK^T}{\sqrt{^d K}}\right)V \quad (4)$$

A multi-head attention is performed before going through self-attention and this is accomplished by Equations 3 and 4. In Equation 3, z_i is the output of the multi-head attention, d_{model} is the input dimension and x is a sequence of words, and in Equation 4, the terms Q, K and V are the query, key which is the vector representation of the words in a sequence, and value respectively. The transformer has achieved state-of-the-art performance on various NMT tasks with improved translation quality. For this reason, we have implemented our models using the transformer architecture (Vaswani *et al.*, 2017).

[1]https://storage.googleapis.com/gresearch/translate-gender-challenge-sets/data/Translated/Wikipedia/Biographies/-/EN_ES.csv

One of the significant causes of gender bias is the imbalances of data in the parallel corpora for all genders. Since most of the data collected are from existing sources originally documented by humans, the presence of biases in the text is typical. Usually, most documents are influenced by societal prejudices. Various occupations are associated only with a specific gender. A few examples of occurrence of such association for occupations and genders are:

Doctor	Male
Nurse	Female
Babysitter	Female
Captain	Male
Police Officer	Male
Professor	Male
Receptionist	Female
Pilot	Male
Surveyor	Male

This problem manifests itself in the NMT systems of English to Khasi, which leads to gender bias in the output of the systems.

TECHNIQUE FOR REDUCING GENDER BIAS

To handle gender bias in the neural machine translation of English to Khasi, we prepare a gender-balanced dataset Dataset GB. This dataset is manually created to handle the issue of imbalances of gender information in the English-Khasi parallel corpora. The Translated Wikipedia Biographies are professionally translated and are available in two pairs of languages, English-Spanish and English-German. The sentences in the dataset represent an entity identified in the biographies as feminine or masculine, a band, or a team. Each dataset instance consists of 8 to 15 sentences referring to the central entity. Our gender-balanced dataset consists of two sets of data; 1363 English sentences are extracted from the existing Translated Wikipedia Biographies and are then manually translated from English to Khasi. The second set consists of 586 pairs of gender-balanced sentences, which are manually created for both English and Khasi. The sentences are short and include various occupations and gender entities to complement the main dataset in terms of gender information. These sentences are generally of two forms:

Form1: (entity1) (verbal phrase) (pronoun) (circumstances)
Form2: (entity1) (verbal phrase) with (pronoun)(entity2) (circumstances)

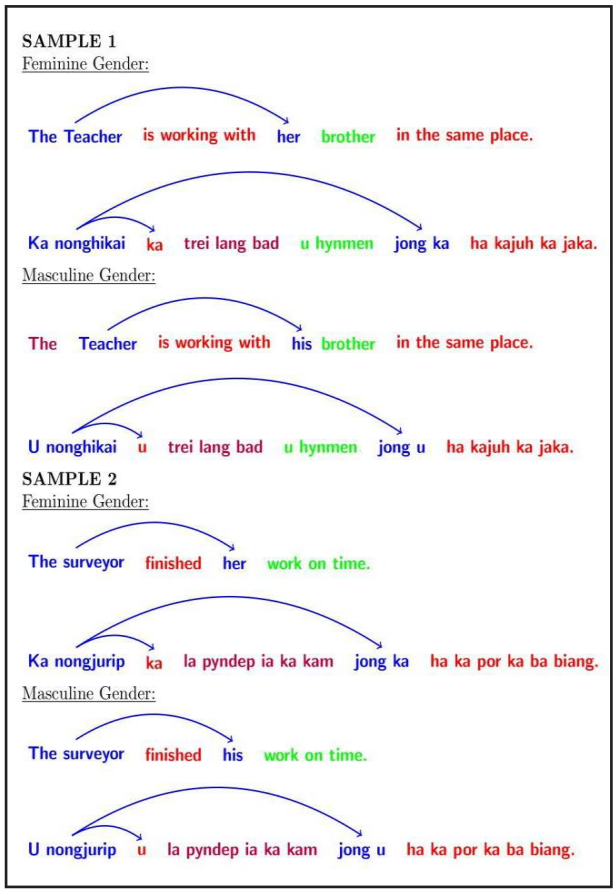


Fig. 1: Gender balanced sentences for English and Khasi pair

For each occupation, we prepare two sets of sentences in English language and Khasi language, associated with masculine and feminine gender as in Figure 1. Each sentence in Figure 1 consists of an entity and a gendered pronoun associated with it, indicated by the direction of the arrow.

We built three datasets as in Table 1; $Dataset_{EK}$ is the main dataset of 50691 parallel sentences. $Dataset_{EK}$ is built using online data from the Bible (Life.Church/YouVersion, 2021a; Life.Church/YouVersion, 2021b) and existing books (Hujon and Singh, 2018) in Khasi language that are translated from an English source. The English-Khasi parallel corpora are manually digitized and aligned. $Dataset_{GB}$ is the gender-balanced dataset, and $Dataset_{CD}$ is the augmented dataset consisting of text from $Dataset_{EK}$ and $Dataset_{GB}$. The $Dataset_{EK}$ dataset, the $Dataset_{GB}$ dataset, and the $Dataset_{CD}$ consists of *9,28,495* words, *34,348*, and *9,62,843* words in total in the English corpus, whereas the $Dataset_{EK}$ dataset, the $Dataset_{GB}$ dataset, and the $Dataset_{CD}$ consists of *12,31,938* words, *49,584*, and *12,81,522* words in total in the Khasi corpus and are distributed for train, validation and test data as shown in Table 1.

We initially implemented the baseline model $NMT_{baseline}$ using a transformer that is trained on the English-Khasi parallel corpora. The baseline model is trained on the main dataset $Dataset_{EK}$. We attempt to reduce gender bias by using two different methods.

In the first method as shown in the schematic diagram in Figure 2, we take the English-Khasi parallel corpora and the gender-balanced parallel corpora and perform data augmentation. In the process, we make sure that data of the gender-balanced dataset is distributed proportionately as train, validation, and test data. Next, the augmented English-Khasi dataset $Dataset_{CD}$, goes through a word segmentation process. We apply two different word segmentation methods, that is, word segmentation method by tokenization using Mosesdecoder toolkit (Koehn *et al.*, 2007) and word segmentation method by subword byte pair encoding (Sennrich *et al.*, 2016) technique. The tokenized augmented English-Khasi dataset and the subword BPE augmented English-Khasi Dataset are trained on different NMT systems.

In the second method as shown in the schematic diagram in Figure 3, the English-Khasi parallel corpora and the gender-balanced parallel corpora initially go through a word segmentation process. Similar to the previous method, we apply two different word segmentation methods, that is, word segmentation method by tokenization using Mosesdecoder toolkit (Koehn *et al.*, 2007) and word segmentation method by subword byte pair encoding(Sennrich *et al.*, 2016) technique, and these give rise to two different pair of datasets; the tokenized English-Khasi dataset and tokenized

Table 1: Datasets Statistics

Dataset	Number of sentences			English(in number of words)			Khasi(in number of words)		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
$Dataset_{EK}$	46263	1772	2656	864465	27356	36674	1145840	36684	49414
$Dataset_{CD}$	47838	2026	2776	892553	31251	39039	1186444	42343	52735
$Dataset_{GB}$	1575	254	120	28088	3895	2365	40604	5659	3321

English-Khasi gender balanced dataset, the subword BPE English-Khasi dataset, and subword BPE English-Khasi gender balanced dataset. Using the two tokenized datasets, we create a joint tokenized vocabulary, and similarly, using the two subword BPE datasets, we create a joint subword BPE vocabulary. Next, we implement two different NMT systems using the transfer learning approach.

Both the parent model and the child model share a joint vocabulary. For the first NMT system using the transfer learning approach, we train the parent model using the transformer architecture on the tokenized $Dataset_{EK}$ dataset till convergence, and the weights of the parent model are used to initialize the child model. The child model is trained on the tokenized English-Khasi gender-balanced dataset, $Dataset_{GB}$, and shares the joint tokenized vocabulary with the parent model. For the second NMT system using the transfer learning approach, we train the parent model and also use the transformer architecture on the subword BPE $Dataset_{EK}$ dataset till it converged, and the weights of the parent model are used to initialize the child model. Next, the child model is trained on the subword BPE English-Khasi gender-balanced $Dataset_{GB}$ dataset. The child model shares the joint subword BPE vocabulary with the parent model.

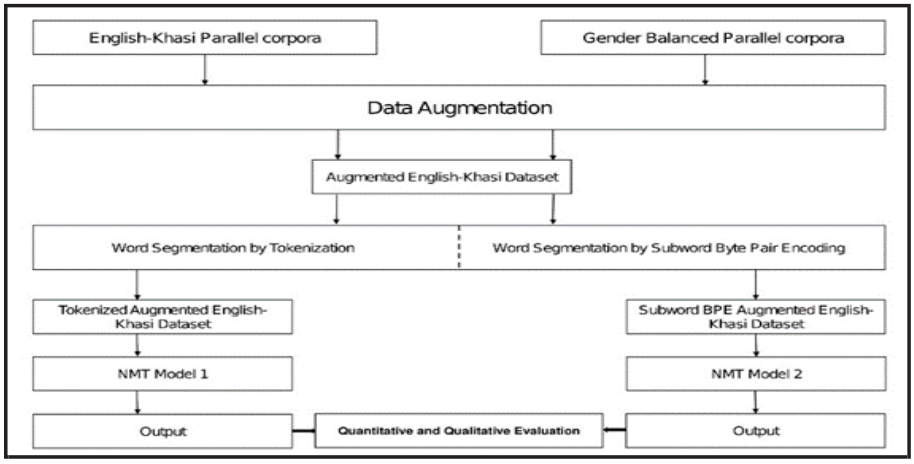


Fig. 2: Schematic Diagram for the Data Augmentation Method

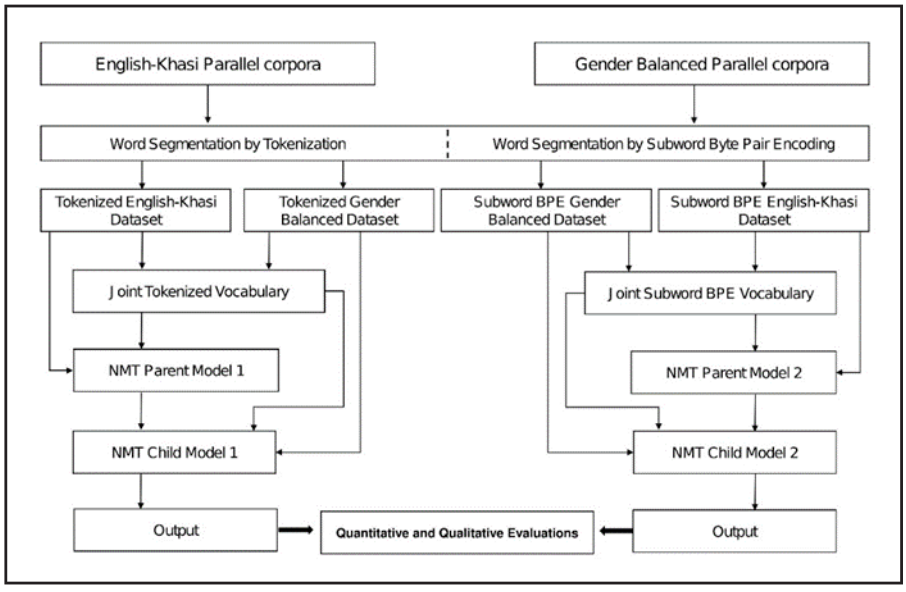


Fig. 3: Schematic Diagram for the Transfer Learning Method

Evaluation Method

Some commonly used evaluation methods are reported for NMT models of low resource languages (Chatzikoumi, 2020; Ranathunga *et al.*, 2023). Evaluation of the output of each model is performed using the sacreBleu (Post, 2018; Papineni *et al.*, 2002) script for automatic evaluation. We also performed a statistical evaluation (Koehn, 2009) based on *Precision*, *Recall*, and *F*1–*Measure* on a scale of 1 to 100. With regards to qualitative evaluation, two annotators evaluate the output on a scale of 1-5 based on *adequacy*- number of correct translated words, *fluency* - grammatical correctness based on word order agreement of the target language, and *gender accuracy* - number of correct translated pronoun in co-reference to a noun in the sentence. We also use the Spearman correlation coefficient as in Equation 5 for determining the inter-annotator agreement. An analysis of the correlation between the quantitative evaluation and qualitative evaluation is performed using the Pearson correlation coefficient as in Equation 6.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{5}$$

$$r = \frac{n\sum xy - \sum x \sum y}{\left(\sqrt{(n\sum x^2 - (\sum x)^2}\left(n\sum y^2 - (\sum y)^2\right)\right)} \tag{6}$$

EXPERIMENTAL SETUP

The two datasets $Dataset_{CD,}$ *and* $Dataset_{GB}$ which are processed using two different word segmentation methods are used to train separate NMT systems to implement the two approaches. We implement two NMT systems which are trained using the data augmented datasets. The system NMT_{CD1} is trained on the $Dataset_{CD}$ tokenized dataset, and NMT_{CD2} is trained on the $Dataset_{CD}$ subword BPE processed dataset. Another two NMT systems using the transfer learning method are trained on the tokenized datasets and the subword BPE datasets. The parent model of the NMT_{TL1} system is trained on the $Dataset_{EK}$ tokenized dataset, and the child model is trained on the $Dataset_{GB}$ tokenized dataset. Similarly, the parent model of the NMT_{TL2} system is trained on the $Dataset_{EK}$ subword BPE dataset and the child model is trained on the $Dataset_{GB}$ subword BPE dataset.

The NMT systems are implemented using the transformer architecture (Vaswani *et al.*, 2017) with six layers and eight heads. We train the models using similar parameters: Adam optimizer, a learning rate of 2.0, a dropout rate of 0.2, and an attention dropout of 0.1.

The translation accuracy of the systems is evaluated on two test sets- $Testset_{CD}$ which is the test set of the augmented dataset and $Testset_{GB}$ which is the test set of the gender-balanced dataset. The quantitative evaluation uses BLEU, ChrF2, and TER. The qualitative evaluation is performed by human judgment based on *accuracy* and *fluency*. For qualitative evaluation, a Python script is used to randomly select 100 samples of sentences for each system. To evaluate the effectiveness of the systems in reducing gender bias, we evaluate the systems on another test set of 100 samples taken from the test set $Testset_{GB}$. The qualitative evaluations are performed by two annotators using two metrics- *adequacy* and *fluency* and *gender accuracy* by two annotators. To assess the inter-annotator agreement, we use Spearman correlation for both *adequacy*, *fluency*, and *gender accuracy*. Using Pearson correlation, we also correlate the BLEU scores with *adequacy* and *fluency*.

RESULTS AND ANALYSIS

The automatic scores are shown in Table 2 for the two test sets, $Testset_{CD}$ and $Testset_{GB}$. NMT_{CD2} achieved the highest score of 54.7 BLEU, 69.6 ChrF2 and Translation Error Rate of around 31 with $Testset_{CD}$, NMT_{CD2} also performed equally well with a slight difference of -0.1 BLEU compared to NMT_{CD2} while NMT_{TL2} achieved the highest score of 12.0 BLEU with $Testset_{GB}$. Compared to the baseline model, NMT_{CD2} shows an improvement of +0.7 BLEU and +0.4 ChrF2, whereas NMT_{TL2} achieved a slightly lower score for $Testset_{CD}$ but shows a significant improvement of +6.8 compared to the baseline model for $Testset_{GB}$. NMT_{CD1} and NMT_{CD2} also performed equally well with score of 10.8 BLEU and 11.3 BLEU on the gender-balanced test set. We observed that the models that are trained on the subword BPE datasets outperformed the other models; NMT_{CD2} with $Testset_{CD}$ and NMT_{TL2} with $Testset_{GB}$. The word embedding of the subword BPE data appears to have a positive impact on the quality of the output translated by the NMT systems.

Table 2: Automatic Scores

Model	$Testset_{CD}$			$Testset_{GB}$		
	BLEU	**ChrF2**	**TER**	**BLEU**	**ChrF2**	**TER**
$NMT_{Baseline}$	54.0	69.2	31.9	5.2	24.8	85.5
NMT_{CD1}	54.6	69.6	31.4	10.8	31.8	80.7
NMT_{CD2}	**54.7**	69.6	31.7	11.3	31.2	82.7
NMT_{TL1}	22.4	41.6	58.6	9.6	25.3	71.2
NMT_{TL2}	21.8	42.0	58.8	**12.0**	30.5	90.3

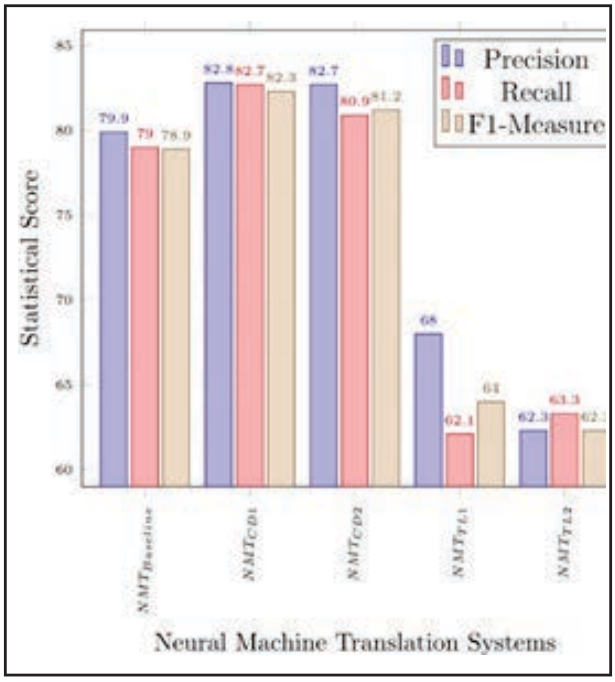
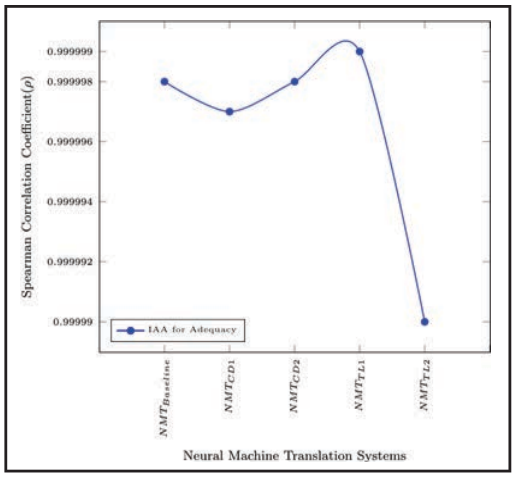


Fig. 4: Statistical Evaluations

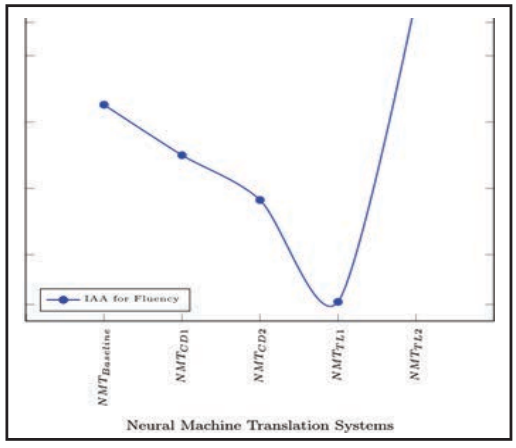
The statistical evaluation is performed, and the results are shown in Figure 4. We observed that the data augmentation models outperformed the other models in the experiment. NMT_{CD1} scored the highest with 82.3 F1-measure score, while NMT_{CD2} scored 81.2 F1-measure score. Both the data augmentation models, NMT_{CD1} and NMT_{CD2} outperformed the baseline model $NMT_{Baseline}$ by an improvement of +3.4 and 2.3 F1-measure score, respectively. The transfer learning models show a lower performance compared to the other models, but both models, NMT_{TL1} and NMT_{TL2} achieved good scores of > 60 F1-measure. On analyzing the automatic score in Table 2 and the statistical score in Figure 4, we find that the automatic scores of $Testset_{CD}$ correlate with the statistical scores, in which the data augmentation models achieved the highest scores in the quantitative evaluations. Considering Figure 4 and Table 3, we also observed a similar pattern in the results of the evaluations, where the proposed data augmentation models showed a higher performance than the other models and outperformed the baseline model.

Table 3: Human Evaluated Scores

NMT System	Adequacy Score	Fluency Score
$NMT_{Baseline}$	3.95	4.48
NMT_{CD1}	4.10	4.55
NMT_{CD2}	4.04	4.52
NMT_{TL1}	3.10	3.47
NMT_{TL2}	3.18	3.97



(a) IAA for Adequacy



(b) IAA for Fluency

Fig. 5: Inter Annotator Agreement (IAA)

The results of human evaluations as performed by the annotators are shown in Table 3. The Spearman correlation is implemented as in Equation 5. The Spearman correlation coefficients for the various NMT systems show a favourable inter-annotator agreement with overall scores of +0.9 for both *adequacy* and *fluency* as shown in Figure 5a and Figure 5b. Comparing the Spearman correlation coefficient between *adequacy* and *fluency*, we observed that human judgment by the annotators on *adequacy* is more correlated to one another than on *fluency*, as the task of evaluating the number of correct translated words can be performed using the same method, while the task of evaluating *fluency* is more difficult and human judgment can slightly differ while ranking each sample. The two data augmentation models NMT_{CD1} and NMT_{CD2} have outperformed the baseline model $NMT_{Baseline}$ in both *adequacy* and *fluency*. Even though NMT_{TL1} and NMT_{TL2} achieved slightly lower scores in human

evaluations among the models, the two models achieved significantly high scores of > 3 in *adequacy* and *fluency*.

We compared the automatic scores and human-evaluated scores using Pearson correlation as in Equation 6. The Pearson correlation coefficient for the various NMT systems is shown in Table 4. All the models show positive scores close to +1, which depicts a favourable correlation between automatic and human evaluation.

Table 4: Pearson Correlations between automatic and human evaluated scores

Model	BLEU Score and Adequacy(r_1)	BLEU Score and Fluency(r_2)
$NMT_{Baseline}$	0.408386899	0.456298011
NMT_{CD1}	0.687616319	0.344257435
NMT_{CD2}	0.653555117	0.407419022
NMT_{TL1}	0.408699657	0.225584008
NMT_{TL2}	0.585296315	0.408386899

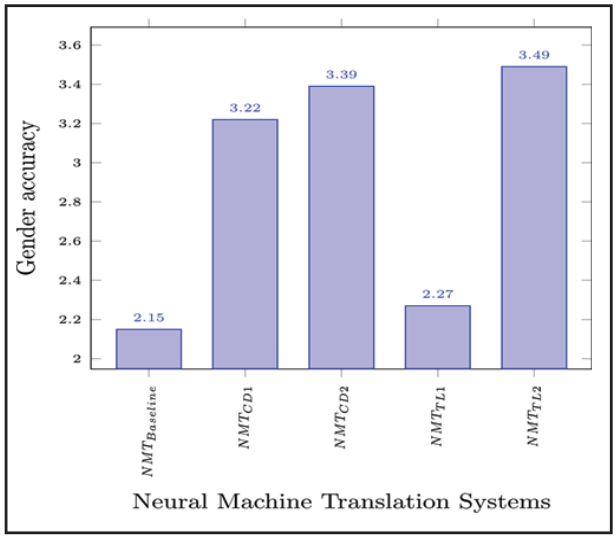


Fig. 6 : Gender Accuracy Scores of the NMT Systems

Gender accuracy is evaluated based on how many pronouns are translated correctly in coreference to each gendered entity in a sentence. Two annotators scored each sample on a scale of 1-5 with 5 as the highest. The scores are as shown in Figure 6. Just as in *adequacy* and *fluency*, the Spearman correlation of gender accuracy as shown in Figure 7 also illustrates a strong inter-annotator agreement with overall scores of +0.9 for *gender accuracy*. NMT_{TL2} achieved the highest score with 3.46875 and outperformed the baseline model by +1.32, closely followed by NMT_{CD1} and NMT_{CD2} which also outperformed the baseline models with an improvement of +1.069 and +1.23 respectively. Referring to Table 2 and Figure 6, we observed that the scores in *gender accuracy* correlate with the BLEU scores achieved by the models for $Testset_{GB}$ in which NMT_{TL2} model outperformed the other models. We also observed that the transfer learning systems NMT_{TL1} and NMT_{TL2} shows a slightly lower performance in translation accuracy as seen in automatic scores in Table 2, statistical scores in Figure 4, and human judgment score on *adequacy* and *fluency* in Table 3 but it shows a good improvement in terms of gender accuracy as seen in Figure 6. NMT_{TL1} outperformed the baseline system, while NMT_{TL2} achieved the highest performance compared to all the NMT systems.

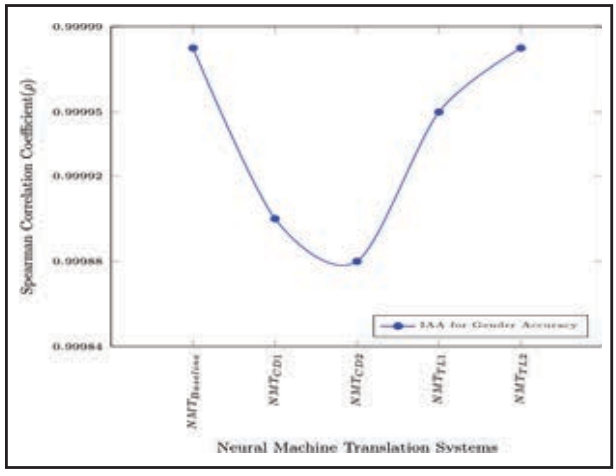


Fig. 7 : Inter Annotator Agreement (IAA) for Gender Accuracy

Output Analysis

We present a study on randomly selected sentences of the NMT systems experimented based on translation accuracy and gender accuracy concerning inputs and the corresponding reference texts.

Table 5: Sample Output-I: Short Sentence

Input	he will defeat kings all over the earth.
Reference	un jop ïa ki syiem baroh jong ka pyrthei.
$NMT_{Baseline}$	un jop ïa ki syiem ha kylleng ka pyrthei
NMT_{CD1}	un jop ïa ki syiem ha kylleng ka pyrthei.
NMT_{CD2}	un jop ïa ki syiem ha kylleng ka pyrthei.
NMT_{TL1}	un jop ia ki syiem jong ka pyrthei baroh kawei.
NMT_{TL2}	un jop ia ki syiem baroh ha ka pyrthei.

Table 6: Sample Output-II: Long Sentence

Input	What then? The people of Israel did not find what they were looking for. It was only the small group that God chose who found it; the rest grew deaf to God's call .
Reference	Kaei pat te ngin ïa-ong? Ki paid Israel kim shym la shem ïa kaei kaba ki ïawad. Ka la long tang ka kynhun kaba rit kaba U Blei u la jied, kaba la shem ïa kata; kiba sah na kita ki la set kyllut ïa ka jingkhot jong U Blei.
$NMT_{Baseline}$	Te uei ? Ki paid Israel kim shym shem aïu ki ïawad . Ka dei tang ka kynhun kaba rit kaba U Blei u la jied ïa ka ; kiba sah ki la nang san sha ka jingkhot U Blei.
NMT_{CD1}	Te? Ki paid Israel kim shym shem ïa kaei kaba ki wad. Ka long tang ka kynhun kaba rit kaba U Blei u la jied ïa ka; kaba sah ka la nang san sha ka jingkhot U Blei.
NMT_{CD2}	Kaei te ? Ki paid Israel kim shym shemïa kaei kaba ki wad . Ka dei tang ka kynhun kaba rit kaba U Blei u la jied ïa ka ; kaba sah ki la nang kyllut ban khot ïa U Blei.
NMT_{TL1}	Nangta pat ? Ki paid Israel kim shym shem ia kata ba ki lah khmih lynti ba ki ioh . Ka long tang ka kynhun kaba rit ba lah jied ban shem ia ka
NMT_{TL2}	Nangta ki paid Israel kim shym shem kat kaba ki dang wad . Ka dei tang ka kynhun kaba rit ba U kpa jong ki uba la lap ia ka ; kiwei pat ki la san ha ka jingkhot jong U pyrthei.

Considering translation accuracy, two samples of the outputs of the NMT systems are shown in Table 5 and Table 6 for short sentences and long sentences respectively. In Table 5, the input text 'he will defeat kings all over the earth' with the corresponding reference text 'un jop ïa ki syiem baroh jong ka pyrthei' is translated more or less accurately by all the NMT systems.

Table 7: Sample Output-III: Gender Accuracy

Input	*She* is *the new jailer* appointed recently
Reference	*ka* dei *ka nongap byndi ba thymmai* ba dang shu sl thung shen
$NMT_{Baseline}$	*Ka* dei *u nongpeit phatok uba thymmai*
NMT_{CD1}	*Ka* dei *ka nongpeit phatok kaba thymmai* ba la thu
NMT_{CD2}	*Ka* dei *ka nongap ing byndi*
NMT_{TL1}	*Ka* dei *ka nongap ing byndi ba thymmai*
NMT_{TL2}	*Ka* dei *ka nongap ing byndi thymmai* ba la thung

Each of the NMT systems maintains the Subject-Verb-Object word order agreement of the target language in their respective outputs and therefore achieves the fluency of the target language. Although the input phrase 'all over' is translated to 'ha kylleng' by $NMT_{Baseline}$, NMT_{CD1}, NMT_{CD2} instead of 'baroh kylleng', the output sentences of these systems convey a similar meaning with fluency. In Table 6, the phrase 'The people of Israel did not find what they were looking for' is translated more correctly to 'Ki paid Israel kim shym shem ia kaei kaba ki wad' by the data augmentation NMT systems and the baseline system. The outputs of the transfer learning NMT systems also efficiently translate the input sentence as in the reference text, with a few disagreements with the word order of the target language. Both the data augmentation NMT systems and the transfer learning NMT systems can translate most of the words as in the reference text. We also observed that the predicted outputs of the NMT systems correlate with the performance of the NMT systems in the quantitative evaluations and qualitative evaluations as in Table 2, Figure 4, and Table 3.

Considering gender accuracy, we also present a study on two randomly selected sentences as we have done for the analysis of translation accuracy. In Table 7, we observe that the occupation 'jailer' refers to a feminine gender in the input sentence, and the correct translation in the reference is 'ka nongpeit byndi', the word 'byndi' or 'patok' in the Khasi language can be used interchangeably. So 'ka nongpeit patok' is also a correct translation. We observed that the baseline model $NMT_{Baseline}$ translates into a masculine form 'u nongpeit patok'. The proposed models show a correct translation and also translate the correct gender pronoun of 'She' to 'ka'. Similarly, in Table 8, we observed that the occupation 'doctor' refer to a feminine gender in the input sentence and the translation to Khasi is 'ka doktor' as shown in the reference text. The baseline model $NMT_{Baseline}$, the NMT_{CD1} model and the NMT_{CD2} model translate to a masculine form 'u doktor', whereas the NMT_{TL1} model and the NMT_{TL2} model, both translate to the correct gender form. The translation in Khasi of the word 'sister' can be either 'ka para', 'ka para kynthei', 'ka hynmen' or 'ka hynmen kynthei', and we notice that all the models generate the correct translation in Khasi. Furthermore, the NMT_{TL1} model generates an almost exact translation as the reference text.

Table 8: Sample Output-III: Gender Accuracy

Input	My sister is working as a doctor in our local hospital and she is very good with kids
Reference	*Ka hynmen kynthei* jong nga ka trei kum *ka doktor* ha ka hospital shnong bad *ka* ba iadei bha bad ki khynnah rit
$NMT_{Baseline}$	*Ka para* jong nga ka trei kum *u doktor* uba stad tamïa kaei kaba *ka* la jia ha shnong rit ka bym salia eiei ruh shaphang ka jingïadei hapdeng jong ngi bad ka Elisabet.
NMT_{CD1}	*Ka para* jong nga ka trei kum *u doktor* ha ka hospital jong ngi bad ka, *ka* long kaba bha shibun bad ka bymju riej ïano ïano ruh.
NMT_{CD2}	*Ka para* jong nga ka trei kum *u doktor* ha ka hospital jong ngi , bad *ka* long kaba bha shibun ïa ka dur jong ki .
NMT_{TL1}	*Ka hynmen* jong nga ka trei kum *ka doktor* ha ka bad ka tbit bha ha ka
NMT_{TL2}	*Ka para* jong nga ka trei kum *ka doktor* ha ka hospital jong ngi bad *ka* long kaba bha bad ki.

CONCLUSION

Addressing gender bias in NMT systems is critical to facilitate inclusive and unbiased communication across linguistic barriers. Our study emphasizes the importance of handling gender bias, and the method using data augmentation and transfer learning with the transformer, a cutting-edge architecture as a backbone, has shown promising results. A sizeable gender-balanced dataset is built and presented for the empirical study. Quantitative and qualitative evaluation and analysis are performed on the output of the systems. Qualitative evaluations are performed by two annotators, and the inter-annotator agreement scores show a good correlation of 0.9 for *adequacy*, *fluency* in terms of translation accuracy, and *gender accuracy* for reducing gender bias. The result of the evaluation shows that the models using the data augmentation method performed better in terms of automatic scores for test data $Testset_{CD}$, in *adequacy* as well as in *fluency*. The transfer learning models also show a significant performance in automatic scores with test data $Testset_{GB}$. The NMT_{CD2} model achieved the highest automatic score for test data $Testset_{CD}$, while the models using the transfer learning method, NMT_{TL2} achieved the highest automatic score with the test data $Testset_{GB}$. The statistical evaluation also correlates with automatic scores in the quantitative evaluations. In the case of gender accuracy, the results of the evaluations show that the transfer learning models performed better than the data augmentation models. The output analysis also correlates with the evaluation of gender accuracy. The NMT_{TL2} achieved the highest gender accuracy score despite giving lesser *adequacy* and *fluency* scores compared to other models. In transfer learning, using the English-Khasi gender-balanced dataset for the child model has shown its effectiveness in reducing gender bias. Our method has shown significant results; however, overcoming the challenge of gender bias for English-Khasi language pairs remains a formidable task, indicating that considerable progress is still to be made. Our efforts represent an initial stride towards this objective. In the future, we plan to experiment using different word embedding techniques in the NMT tasks of English-Khasi language pair.

DECLARATIONS

Competing Interest: The authors have no relevant financial or non-financial interests to disclose.

Ethical standards: This work described in this manuscript is original and has not been under consideration for publication elsewhere. All authors read and approved the final manuscript. The research in this manuscript does not involve human participants and animals.

Data availability and access: Data will be made available on reasonable request.

Authors contribution: **Aiusha Vellintihun Hujon**: Conceptualization, Data curation, Methodology, Investigation, Software, Validation, Formal analysis, Writing-Original draft preparation. **Thoudam Doren Singh**: Supervision, Writing - Review & Editing. **Khwairakpam Amitab**: Supervision, Writing - Review & Editing.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

REFERENCES

Cristina Espana-Bonet, Adam Csaba Varga, Alberto Barron-Cedeno, and Josef van Genabith. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11 (8):1340–1350, December 2017. ISSN 1941-0484. doi: 10.1109/jstsp.2017.2764273. URL http://dx.doi.org/10.1109/JSTSP.2017.2764273.

Rico Sennrich and Barry Haddow. Linguistic input features improve neural machine translation, 2016.

Joel Escud´e Font and Marta R. Costa-juss`a. Equalizing gender bias in neural machine translation with word embeddings techniques. In Marta R. Costa-juss`a, Christian Hardmeier, Will Radford, and Kellie Webster, editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3821. URL https: //aclanthology.org/W19-3821.

Aiusha Vellintihun Hujon, Thoudam Doren Singh, and Khwairakpam Amitab. Neural machine translation systems for english to khasi: A case study of an austroasiatic language. *Expert Systems with Applications*, 238:121813, 2024. ISSN 0957-4174. doi: https://doi.org/10. 1016/j.eswa.2023.121813. URL https://www.sciencedirect.com/science/article/pii/S0957417423023151.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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*, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Thoudam Doren Singh and Aiusha Vellintihun Hujon. Low resource and domain specific english to khasi smt and nmt systems. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 733–737. IEEE, 2020.

Aiusha Vellintihun Hujon, Khwairakpam Amitab, and Thoudam Doren Singh. Convolutional sequence to sequence learning for english-khasi neural machine translation. In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–4, 2023a. doi: 10.1109/I3CS58314.2023.10127426.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016. URL https://arxiv.org/abs/1409.0473.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, Nov 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1163. URL https://aclanthology.org/D16-1163.

Tom Kocmi and Ondrej Bojar. Trivial transfer learning for low-resource neural machine translation. *CoRR*, abs/1809.00357, 2018. URL http://arxiv.org/abs/1809.00357.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1334. URL https://aclanthology.org/D18-1334.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016a. URL http://arxiv.org/abs/1607.06520.

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

Odbal, Guanhong Zhang, and Sophia Ananiadou. Examining and mitigating gender bias in text emotion detection task. *Neurocomputing*, 493:422–434, 2022. ISSN 0925-2312. doi: https: //doi.org/10.1016/j.neucom.2022.04.057. URL https://www.sciencedirect.com/science/ article/pii/S0925231222004374.

Marta R. Costa-juss`a, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. Interpreting gender bias in neural machine translation: Multilingual architecture matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 11855–11863, Jun. 2022. doi: 10.1609/aaai.v36i11.21442. URL https://ojs.aaai.org/ index.php/AAAI/article/view/21442.

Marcelo O. R. Prates, Pedro H. Avelar, and Lu´ıs C. Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Comput. Appl.*, 32(10):6363–6381, may 2020. ISSN 0941-0643. doi: 10.1007/s00521-019-04144-6. URL https://doi.org/10. 1007/s00521-019-04144-6.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and Fran¸cois Yvon. Analyzing gender translation errors to identify information flows between the encoder and decoder of a NMT system. In Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe, editors, *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 153–163, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/ 2022.blackboxnlp-1.13. URL https://aclanthology.org/2022.blackboxnlp-1.13.

Ond˘rej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 workshop on statistical machine translation. In Ond˘rej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3001.

URL https://aclanthology.org/W15-3001.

Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: The Winograd schema challenge. In Jun'ichi Tsujii, James Henderson, and Marius Pa¸sca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1071.

Toan Q. Nguyen and David Chiang. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan, nov 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/ I17-2050.

Aiusha V Hujon, Thoudam Doren Singh, and Khwairakpam Amitab. Transfer learning based neural machine translation of english-khasi on low-resource settings. *Procedia Computer Science*, 218:1–8, 2023b. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2022.12.396. URL https://www.sciencedirect.com/science/article/pii/S1877050922024899. International Conference on Machine Learning and Data Engineering.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Life.Church/YouVersion. Gnb bible youversion, 2021a. URL https://www.bible.com/en-GB/ bible/296/GEN.1.GNB. Accessed: March 2021.

Life.Church/YouVersion. Khasiclbsi bible youversion, 2021b. URL https://www.bible.com/ en-GB/bible/1865/EXO.1.KHASICLBSI. Accessed: March 2021.

Aiusha Vellintihun Hujon and Thoudam Doren Singh. Existing english to khasi translated documents for parallel corpora development: A survey. *International Journal on Natural Language Computing (IJNLC)*, 7(5):81–91, 2018.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ond˘rej Bojar, Alexandra Constantin, and Evan Herbst. 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, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-2045.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers) 1715–1725 association for computational linguistics https://www. aclweb. org/ anthology. *P16-1162*, 2016.

Eirini Chatzikoumi. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161, 2020. doi: 10.1017/S1351324919000469.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11), February 2023. ISSN 0360-0300. doi: 10.1145/3567592. URL https://doi.org/10.1145/3567592. Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/ anthology/W18-6319.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Philipp Koehn. *Evaluation*, page 217–246. Cambridge University Press, 2009. doi: 10.1017/ CBO9780511815829.009.