Gender Bias Evaluation in Luganda-English Machine Translation

Eric Peter Wairagala kigaye.ericpeter@gmail.com
Jonathan Mukiibi jonmuk7@gmail.com
Jeremy Francis Tusubira tusubirafrancisjeremy@gmail.com
Claire Babirye clarybits68@gmail.com
Joyce Nakatumba-Nabende joyce.nabende@mak.ac.ug
Department of Computer Science, Makerere University,Uganda,Kampala

Andrew Katumba andrew.katumba@mak.ac.ug
Department of Electrical and Computer Engineering , Makerere University, Uganda Kampala

Ivan Ssenkungu ssenkungu.ivanandrew@gmail.com
Department of Languages ,literature and communication, Makerere University, Uganda Kampala

Abstract

We have seen significant growth in the area of building Natural Language Processing (NLP) tools for African languages. However, the evaluation of gender bias in the machine translation systems for African languages is not yet thoroughly investigated. This is due to the unavailability of explicit text data available for addressing the issue of gender bias in machine translation. In this paper, we use transfer learning techniques based on a pre-trained Marian MT model for building machine translation models for English-Luganda and Luganda-English. Our work attempts to evaluate and quantify the gender bias within a Luganda-English machine translation system using Word Embeddings Fairness Evaluation Framework (WEFE). Luganda is one of the languages with gender-neutral pronouns in the world, therefore we use a small set of trusted gendered examples as the test set to evaluate gender bias by biasing word embeddings. This approach allows us to focus on Luganda-Engish translations with gender-specific pronouns, and the results of the gender bias evaluation are confirmed by human evaluation. To compare and contrast the results of the word embeddings evaluation metric, we used a modified version of the existing Translation Gender Bias Index (TGBI) based on the grammatical consideration for Luganda.

1 Introduction

Uganda is a highly multilingual country with over 43 known indigenous languages and dialects (Eberhard et al., 2020). However, many languages have no existing resources for building Natural Language Processing (NLP) datasets and tools. One of the major languages spoken in Uganda is Luganda, primarily spoken in the South Eastern Buganda region, mainly along the shores of Lake Victoria and up north towards the Lake Kyoga shores (Nakayiza, 2013; Olaide and Azizi, 2019). Luganda is spoken by more than six million people, principally in central Uganda, including Kampala, the capital of Uganda. Topologically, it is a highly agglutinating, tonal language with subject-verb-object, word order, and nominative-accusative morphosyntactic alignment (Olaide and Azizi, 2019). Like many languages in sub-Saharan Africa, Luganda

has limited text and speech data resources, making it a low-resourced language.

However, due to the increase in the availability of computational resources and datasets, there has been high advancement in NLP research. Natural language processing applications approach tasks ranging from low-level processing, such as assigning parts of speech to words, to high-level tasks, such as question answering, and translating speech and text from one language to another. In this paper, *we focus on building and evaluating Machine Translation (MT) models for Luganda* as an application of natural language processing. Classically, rule-based systems were used for MT, these were replaced with statistical methods (Macketanz et al., 2017) in the 1990s. More recently, deep neural network models have achieved state-of-art results in the field of Neural Machine Translation (NMT) (Koehn, 2017).

With Neural Machine Translation and language processing tools becoming more prevalent, there has been a high interest in understanding and mitigating bias in NLP systems. This is because NLP systems are considered susceptible to social bias (Hovy and Spruit, 2016). The investigation of bias is not only a scientific and technical endeavour but also an ethical one, given the growing role of NLP applications (Bender and Friedman, 2018). Bias in MT is when an MT model systematically and unfairly discriminates against certain individuals or groups in favour of others (Savoldi et al., 2021a). Since MT systems are used daily by millions of individuals, they could impact a wide array of people in different ways (Savoldi et al., 2021a).

In the real world, the highest form of bias in machine translation systems is gender bias, which manifests itself when training data has more features and examples of a given gender stereotype compared to others. Machine translation (MT) tools trained on such data inherit the existing biases in the data.

Different languages deal with gender in different ways; for instance, some languages have gendered pronouns like *he/she/him/her* in English, (Ciora et al., 2021) whereas others, such as Luganda, Finnish, Hungarian, Turkish, etc. have neutral pronouns (Savoldi et al., 2021b). Unlike machines, a human translator can understand what the correct translation should be depending on the context. However, this can be complex with machine translation (MT) tools, except the models are contextualized to a specific domain. The gender bias problem occurs when the MT engine has to pick one pronoun over another, as dictated by the noun.

Gender bias in machine translation for gender-neutral languages is one of the most complex forms of bias. A case in point is when translating from Luganda-English; all gender-neutral pronouns are translated into gender-specific nouns. For example, consider the Luganda sentence below: *Musawo mu ddwaliro ly'e Mulago.* is translated to a gender-specific sentence: *He is a doctor at Mulago Hospital*, we show that the word *He* is neutral in Luganda hence being non-existent. The word *Musawo* is translated to *Doctor*, the word *mu* is translated into two English words *is* and *a* respectively, while the words *ddwaliro*, and *ly'e* are translated into *hospital* and *at* respectively as shown in the Figure 1. These examples show how a phrase in Luganda can be correctly translated into English with different gender variations.

Source Sentence [lg]	Target Sentence [en]
Musawo ***mu*** ddwaliro ly'e Mulago.	He ***is a*** doctor at Mulago Hospital.

Figure 1: The translation of a gender-neutral pronoun in Luganda to a gender-specific in English in a **Luganda** (lg) to **English** (en) machine translation system.

Recent approaches to bias in NLP have involved training on artificially gender-balanced versions of the original dataset (Savoldi et al., 2021b). In this work, part of the gendered datasets

have been used to understand gender bias that occurs during the translation of gender-neutral pronouns (Cho et al., 2019). This work has led to improvements in translation of gender-neutral languages. For example, Google Translate has made significant improvements to translation quality and provides both feminine and masculine translations when translating single-word queries from English to languages like French, Italian, Portuguese, and Spanish. This is also the case when translating phrases and sentences from Turkish to English. Recently, Luganda has been added as a language to the Google Translate API (Bapna et al., 2022). However, the Luganda to English translations still suffer from the same problem of returning gender-specific variants when given a gender-neutral Luganda sentence.

Due to a lack of language resources, we have seen a slow growth of machine translation for Ugandan languages. In this paper, we leverage the utility of transfer learning on a small set of trusted, gender-balanced examples to evaluate gender bias in our Luganda-English MT model. The main contributions of this paper are:

1. We build Machine Translation (MT) models for the Luganda language.

2. We create and release a gendered English-Luganda corpus of 1,000 sentences as a test set. The English-Luganda parallel corpus and gender-balanced corpus are publicly available under a CC-0 licence[1].

3. We evaluate gender bias of the Luganda-English machine translation.

The remainder of the paper is organized as follows: In Section 2, we discuss related work in machine translation and gender bias in machine translation models. In Section 3, we present the methodology used in the paper, including the dataset creation process. Section 4 discusses the model performance and evaluation. Finally, Section 7 concludes the paper.

2 Related Work

In this section, we review related work in Machine Translation (MT) of low-resourced languages, the models used, and the evaluation of gender bias in Machine Translation (MT).Neural Machine Translation (NMT) has seen a tremendous growth spurt in less than ten years. While considered the most widely used solution for Machine Translation, its performance on low-resource language pairs remains suboptimal compared to the high-resource counterparts, for example, English, German, and Spanish, among others, due to the unavailability of large parallel corpora. Therefore, the implementation of NMT techniques for low-resource language pairs has been receiving the spotlight in the recent NMT research arena, thus leading to a substantial amount of research (Ranathunga et al., 2021). Prior work in Machine Translation (MT) with a focus on low-resourced languages has been building language corpora and baseline models.

The lack of training data motivated research to compare zero-shot learning, transfer learning, and multilingual learning of three Bantu languages (Shona, isiXhosa, and isiZulu) and English (Nyoni and Bassett, 2021). In the study on Neural Machine Translation (NMT) for African Languages, the authors to (Martinus and Abbott, 2019) address the problems of the lack of datasets required for machine translation and existing research to reproduce the work on African languages.

Adelani et al. (2021) presents the MENYO20k Yoruba-English language with standardized train-test splits for model benchmarking. Researchers are leveraging several sources of text data for creating datasets focused on news headlines and text sources in their local context (Marivate et al., 2020). The authors in (Nekoto et al., 2020) propose a participatory approach to building parallel corpora and MT models to deal with the lack of language resources. Based on an

[1] https://doi.org/10.5281/zenodo.5864560

ongoing Lacuna-funded project, an effort has been made to build parallel text corpora for five Ugandan languages, i.e., Luganda, Runyokore-Rukiga, Acholi, and Lumasaaba (Babirye et al., 2022). There has been advancement in building Machine Translation (MT) models and datasets by leveraging pre-trained and multilingual models. Research that involved building datasets for African Languages and researchers adapted several multilingual pre-trained baseline models (Ifeoluwa Adelani et al., 2022). Work has been done in which a multilingual parallel corpora were created for five (5) Ugandan languages and carried out on Neural Machine Translation (NMT) models to build baseline multilingual models (Akera et al., 2022).

While translation technologies bring undeniable advantages in many contexts, it is also evident that they come with inherent risks, such as reproducing and even amplifying real-world asymmetries by codifying and entrenching various kinds of biases. One of the biases is gender bias, which affects automatic translation. This is also seen when systems are required to overly express gender in the target languages while translating from languages that do not convey such information (Vanmassenhove et al., 2019). In the paper by (Savoldi et al., 2021b), the authors present the research carried out to understand, assess and mitigate gender bias in automatic translation. The study discusses how the socio-cultural notions of gender interact with language(s) and translation and frames, which factors can contribute to the emergence of gender bias in automatic translation systems. They present the resources created to assess the biased behaviour of MT systems and the mitigation strategies developed to reduce feminine under-representation in their outputs.

The authors (Vanmassenhove et al., 2019) treat gender as a domain for machine translation, training from scratch by augmenting Europarl data with a tag indicating the speaker's gender. This does not inherently remove gender bias from the system, but allows control over the translation hypothesis of gender. Work in (Gupta et al., 2021) evaluates and quantifies the gender bias within a Hindi-English machine translation system, and they implement a modified version of the existing one. Translation Gender Bias Index (TGBI) metric is based on the grammatical considerations for Hindi. They compare the results of Word Embeddings Fairness Evaluation (WEFE) framework metrics with the pre-trained word embeddings and the ones learned by their machine translation model.

To our knowledge, there is not a lot of work done on evaluating gender bias in translation languages with gender-neutral languages. In a research study by (Cho et al., 2019), gender bias is measured in the translation of gender-neutral pronouns using Translation Bender Bias Index (TGBI). In the TGBI metric, the authors quantify the associations of "he", "she" and other related gendered words in the translated text. In this paper, we used the Word Embeddings Fairness Evaluation (WEFE) framework metrics to evaluate gender bias on word embeddings learned by our translation system. This is because word embeddings exhibit stereotypical bias towards gender, race, religion, ethnicity etc (Badilla et al., 2020). We also used a modified version of the TGBI metric on Luganda, a low-resourced language with gender-neutral pronouns.

3 Corpus Creation

In this section, we describe the process taken to create the Luganda to English parallel corpus used for training and evaluating gender bias in the Machine Translation model.

3.1 English Corpus Creation

The first step we took was to create an English Corpus, which was eventually translated to Luganda. The English corpus was compiled from various sources that included: news websites, blogs, Wikipedia, and magazines. However, the content from some of these sources was copyrighted, and the structure of the English sentences was too formal.

To deal with this, we undertook a sentence creation process whereby the extracted sen-

tences from the various sources were used as source sentences to prompt the creation of new sentences. The process involved the creation of a new conversational-like English sentence given the source sentence, as shown in Table 1.

Table 1: An example of an English sentence created and translated to Luganda during the corpus creation process

`English Source sentence`	—— Six candidates were successfully nominated.
`New English Sentence`	—— How many presidential candidates were nominated?
`Luganda Translation`	—— Abeesimbyewo bameka abaalondebwa okuvuganya ku bwa pulezidenti?

It was important to be as diverse as possible in the creation of the dataset and try to prevent topic bias. Therefore, during sentence sourcing, we collected as much data as possible from several sources relevant to the Uganda context. The data included topics around agriculture, health, politics, and laws and from the less formal sources like social media data, and blogs to the more formal sources like newspapers.

Each person was given a set of source sentences and was required to create new instances of data on the same topic of discussion. The sentences were then reviewed at two levels, (1) de-identification and (2) meaningfulness and grammatical correctness. The English sentences were created under a CC-0 licence. After the sentence creation process, the next step was the translation process where the English corpus was translated to Luganda through a crowdsourcing and iterative approach.

3.2 Creation of the Makerere English-Luganda Corpus

The English to Luganda translation process was carried out using the Pontoon system, which is a translation management system developed by Mozilla[2]. The English sentences were translated by a team of linguists from the Department of African languages at Makerere University. The translation was a three-stage process. As a first step, a linguist translated the English sentences to Luganda. In the next step, the Luganda translations were validated by a professional linguist. Finally, the translated corpus was subjected to a final check whereby the linguist randomly selected and checked the translated sentences in the parallel corpus. The linguist documented any major issues, which were sent back to the translator for any corrections. The first version of the Makerere English-Luganda corpus is available on Zenodo[3].

In addition to the Makerere English-Luganda corpus, we used other online datasets to train our MT models. These included:

1. **Bible data:** The English and Luganda versions of the Bible are publicly available. We obtained this dataset and are pre-processed, which involved verse-by-verse alignment of the English and Luganda Bible translations.

2. **Formal news articles:** We obtained English news articles from various online websites with a focus on the Ugandan context. We translated these sentences to Luganda and them as part of the parallel corpus.

3. **Gendered sentences:** We created a gendered English corpus using the same criteria as described in Section 3.1. The significant difference in this process is that we focused on only English-gendered sentences from online sources. These sentences were then translated to Luganda. We used this parallel corpus with gendered examples in Luganda as a test set

[2] `https://pontoon.mozilla.org`
[3] `https://doi.org/10.5281/zenodo.4764038`

Table 2: Statistics of the various available English-Luganda parallel corpus used to train and evaluate the Makerere Luganda to English Machine Translation model.

Dataset	Language	Sentences	Tokens	Word Types
Makerere English-Luganda Corpus	English	15,000	136,000	13,043
	Luganda	15,000	115,650	24,694
Makerere gendered corpus	English	1,000	9,920	2,588
	Luganda	1,000	8,190	3,652
Bible	English	31,000	784,708	34,029
	Luganda	31,000	609,145	93,790
News articles	English	21,000	129,005	18,261
	Luganda	21,000	118,173	26,372

to evaluate gender bias in our MT models. We openly release this corpus as the Makerere gendered corpus on Zenodo[4].

Table 2 provides a summary of the different English to Luganda parallel corpora that were used to train the MT models.

To determine the extent of gender bias in our training dataset, we developed a simple custom regex expression algorithm to extract gender pronouns from English sentences in the corpus. The algorithm focused on gender pronouns in the English monolingual corpus, since the Luganda language does not have gender pronouns. The algorithm extracts the pronouns from a sentence and returns the counts of occurrence of each pronoun in the entire text corpus. In Figure 2, we see that masculine pronouns like *He*, *His* and *Him* have the highest number of occurrences in the dataset, hence depicting representational gender bias in our training dataset. The algorithm used is not very accurate for measuring gender bias in a text corpus, but it shows how the dataset is represented across gender.

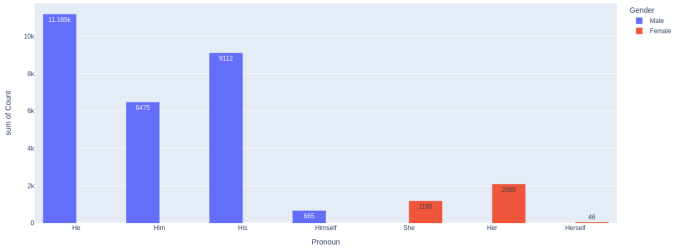

Figure 2: The distribution of masculine and feminine pronouns in the English monolingual corpus.

4 Model Training

4.1 Machine Translation

After obtaining the English to Luganda parallel corpus, the next step was to develop baseline machine translation models. We split the dataset into training (80%), testing (10%), and validation (10%) sets. We used *63,840* parallel sentences in the train, *10,640* in the test, and *10,640* in the validation sets. We created translation objects that write JavaScript object notation (JSON)

[4]https://doi.org/10.5281/zenodo.5864559

files of train, test, and validation sets. These were pushed to the Hugging Face hub [5] which provides a platform to collaborative platform for model training. We train the transformer (Vaswani et al., 2017) model the multilingual Marian MT model for our experiments (Junczys-Dowmunt et al., 2018). We leverage transfer learning on the *Helsinki-NLP/opus-mt-lg-en* and *Helsinki-NLP/opus-mt-en-lg* pre-trained multilingual models. The models were trained for 30 epochs with a batch size of 16 and 10,640 sentences from the validation set at each training step.

4.1.1 Translation Performance

Our initial results demonstrated good performance on the test set and the translation quality with a BLEU score of 26.0 for the English-Luganda model and a BLEU score of 24.6 for the Luganda-English model as shown in Table 3.

Model	Data Size	BLEU
English-Luganda	10,640	**26.0**
Luganda-English	10,640	**24.6**

Table 3: Model evaluation metrics on the test set.

4.2 Word Embeddings

We trained *Word2Vec* word embeddings on the translated sentences of the gendered-examples test set. These embeddings are used in the WEFE framework to evaluate gender bias in the Luganda-English translation system. We used *Continuous Bag of Words (CBOW) Model* and *Skip-Gram Model* word2vec architectures to create word embedding models proposed by (Mikolov et al., 2013) because we had a small gendered-examples test set.

Once the neural network is trained, it results in the vector representation of the words in the training corpus. The size of the vector is also a hyperparameter that we used to produce the best possible results (Mikolov et al., 2013).
To train the *word2vec* model with the CBOW technique, we pass *sg=0* along with other parameters like epochs, and workers. The *sg* parameter denotes the training algorithm. If *sg=1* then skip-gram is used for training and if *sg=0* then CBOW is used for training. These were then used in the WEFE framework and the results are shown in Table 4.

5 Gender Bias Evaluation

5.1 Word Embeddings Fairness Evaluation framework

For gender bias evaluation, we used the Word Embeddings Fairness Evaluation framework (WEFE) to measure gender bias in our MT system (Cho et al., 2019). In this work, we used four (4) metrics from the WEFE framework to measure and quantify gender bias in our translation system. These included (1) the Word Embedding Association Test (WEAT) Cho et al. (2019) metric, (2) WEAT Effect Size (WEAT ES) Cho et al. (2019), (3) the Relative Norm Distance (RND) Garg et al. (2018) and (4) the Relative Negative Sentiment Bias (RNSB) Sweeney and Najafian (2019).

The results of gender bias evaluation are presented in Table 4 for both Word2Vec (Skip-Gram) and Word2Vec (CBOW). WEFE takes in a query, which is a pair of two sets of target words and sets of attribute words each, which are generally assumed to be characteristics related to gender. A *Target set* also denoted by T, corresponds to a set of words intended to denote a particular social group, which is defined by a certain criterion (Badilla et al., 2020). An *Attribute set* denoted by A is a set of words representing some attitude, characteristic trait, or occupational

[5] https://huggingface.co/

field that can be associated with individuals from any social group (Badilla et al., 2020). A *query* is a pair $Q = (T, A)$ in which T is a set of target word sets, and, A is a set of attribute word sets. For example: consider target word sets

$$T_{women} = (she, woman, girl, ...), T_{men} = (he, man, boy, ...) \tag{1}$$

and the attribute word sets

$$A_{science} = (math, physics, chemistry, ...), A_{art} = (poetry, dance, literature, ...) \tag{2}$$

Then the following is the query in the WEFE framework

$$Q = ((T_{women}, T_{men}), (A_{science}, A_{art})) \tag{3}$$

The WEFE ranking process takes in an input of a set of multiple queries (Q), which serve as tests across which bias is measured, a set of pre-trained word embeddings (M), and a set of fairness metrics (F).

Model name	WEAT	WEAT ES	RND	RNSB
Word2Vec (Skip-Gram)	2 (0.268)	2(0.973)	1(0.24)	1 (0.04)
Word2Vec (CBOW)	1(0.131)	1(0.52)	2(0.594)	2(0.294)

Table 4: The results of the WEFE framework metrics that were used on the embeddings models on the Makerere gendered corpus.

In the queries we look at the male and female terms, terms in a career versus family, math versus arts, science versus arts, intelligence versus appearance, pleasant versus unpleasant, negative versus positive words, intelligence versus sensitivity, and male versus female roles. These terms are therefore used to measure gender bias in the Luganda-English translations. The individual and cumulative scores help us assess gender bias in Luganda-English translation.

5.2 Translation Gender Bias Index (TGBI)

The measure takes in a sentence `S` with each sentence containing a pronoun of which gender neutrality should be maintained in the translation, with p_w being the portion representing female in the translations,p_m male and p_n as gender-neutral Cho et al. (2019).The constraints then become,

$$p_w + p_m + p_n = 1 \tag{4}$$

$$0 \leq p_w, p_m, p_n \leq 1 \tag{5}$$

which is defined by

$$P_s = \sqrt{p_w p_m + p_n} \tag{6}$$

Using this measure, we investigated gender bias in two translation models, Google Translate and Luganda-English model. This was done on a list of seven (7) different kinds of sentences, occupation, formal, informal, polite, impolite, negative and positive.

Sentence	Size	Luganda-English model	Google Translate
Occupation	1000	0.6123(0.0487,0.3449)	0.6910(0.0064,0.4741)
Formal	1000	0.6018(0.0679,0.3206)	0.6770(0.0051,0.4556)
Informal	1000	0.6017(0.0750,0.3163)	0.6793(0.0066,0.4579)
Polite	1000	0.6057(0.0727,0.3229)	0.6864(0.0070,04674)
Impolite	1000	0.5977(0.0703,0.3139)	0.6695(0.0047,0.4457)
Negative	1000	0.5228(0.0925,0.2088)	0.6320(0.0075,0.0.950)
Positive	1000	0.6491(0.173,0.3387)	0.6720(0.0000,0.4516)
Average		**0.5987**	**0.6725**

Table 5: Evaluation results for Luganda-English model and Google Translate. For the sentence sets (occupation-positive) denote Ps (pw, pn) for each sentence set S.We calculated the average TGBI values shown in the last row, which is between 0 and 1

5.3 Human Evaluation

The interest in gender bias evaluation and mitigation in Natural Language Processing (NLP) has greatly impacted social research. This study engaged Luganda speakers and experts to validate and annotate gender in translations from Luganda. The experts annotated the output translation with the target gender and the predicted gender, but this time added a bit of context to their verdict on why they think the model was either biased or not.

Human validation of translations is time costly, therefore for this reason we sample 100 out of 1,000 sentences from the Makerere gendered corpus. One of the obvious observations was that some occupations like engineering were associated with the masculine stereotype, whereas nurse and secretary to feminine. It is believed that women are better caretakers than men, which can subsequently lead to the idea that women are better suited for domestic work rather than a professional career. This human bias affects the creation and curation of datasets used in training MT models. This makes the models susceptible to such bias. Human validations confirmed high gender bias towards the female stereotype in our machine translation system. In a gender-neutral sentence in Figure 3, we do not know the gender of the person, “beauty but not wise/intelligent” is associated with the female stereotype whereas “beauty but wise/intelligent” to the male stereotype. This is because for some reason the model has picked up a certain bias that in a given case the female or male stereotype is more likely.

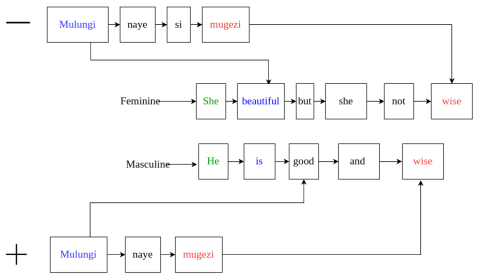

Figure 3: Translating gender-neutral sentences from Luganda-English, the Machine Translation (MT) model basically does not know who is speaking, so it picks up a gender.In this case the model seems to pick a gender variant for the positive and negative Luganda gender-neutral `SRC`, for some reason it takes the female variant for beauty but not intelligent/wise and picks up the male variant for good and wise/intelligent.

6 Discussion of Results

6.1 Machine Translation Models

The results of machine translation models were included in Table 3.BLEU scores of the two models, with the English-Luganda model performing better than the Luganda-English model by 2% on a test set of *10,640* sentences.

The performance of the model fine-tuned Marian MT model on the small 68K training set is good. However, the model had a poor performance on translation quality of single words for example days of the week, numbers, dates, and currencies. This suggests that there is a need to build more effective and better methods to fine-tune MT models, ranging from the corpora used and the models chosen.

en-lg
SRC——Farmers are encouraged to keep farm records.
TGT——Abalimi bakubirizibwa okukuuma ebiwandi- iko by'ebikolebwa ku ffaamu.
REL——Abalimi bakubirizibwa okukuuma ebi- wandiiko by'okulimirako.
lug-en
SRC——Yagamba nti nnyina tamanyi kwogera Lungereza.
TGT—— She said that her mother does not know how to speak English.
REL —— He said his mother doesn't know how to speak English.

Table 6: **Example translations** for different sentences from our test set corpus from our en-lg and lg-en models.

In Table 6, the words in blue colour in the REL sentence are the corresponding correct translations of words in purple in the SRC sentence. The table also shows where our models were not able to translate some words correctly. The words in red in REL are the wrong translations of words in orange colour in the TGT sentence.

6.2 Gender bias Evaluation

With the examples in Table 7, we see that gender bias manifests where our (MT) system attributes the nurse occupation to the female persona and the doctor occupation is attributed to the male persona. Therefore, in this study, we attempted to quantify gender bias in the *lg-en* system translations. Our main focus was on only four (4) WEFE framework metrics to measure gender bias in the *English* monolingual corpus translated by the *lg-en* model.

Luganda-English translation example
SRC —— Omusawo yatuma omubazzi mu ddwaliro kubanga yali yeegendereza nnyo.
TGT ——The nurse sent the carpenter to the hospital because he was extremely cautious.
REL ——The nurse sent the carpenter to the hospital because she was very careful.
SRC —— Ye musawo mu ddwaliro ly'e Mulago.
TGT ——She is a doctor at Mulago Hospital.
REL —— He is a doctor at Mulago Hospital.

Table 7: **Example translations** outputs for showing gender bias manifests in the *lg-en* model. Terms in red represent masculine pronouns while the terms in blue represent feminine pronouns.

Since the framework uses word embedding models to measure bias in a corpus we trained *Word2Vec* word embeddings models as shown in the Table 4. We observe that the *Word2Vec (Skip-Gram)* embeddings model is on top of the ranking of the models, hence exhibiting much more bias towards gender as shown in Figure 4. In this method, we used a small set of queries and target attributes because our test set has less representation of all gender bias occurrences in NLP. Our findings show a heavy tendency for *lg-en* MT systems to produce gendered outputs for gender-neutral pronouns.

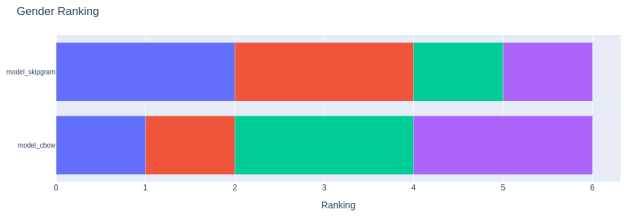

Figure 4: Cumulative rankings for the overall results of the WEFE metrics. Each colour in the plot represent a metric, for example WEAT metric,WEAT ES metric, RND metric, and RNSB metric.

From the results in table 4, we observed a slight disparity in the results of the *word2vec* Skip-Gram and CBOW embeddings. Therefore, bias is not entirely minimal considering that the models had a small set of data. This could be attributed to the fact that the results of WEAT for Family vs Career and Man roles vs Woman roles are very significant. There is a skew in most of the results, which is entirely the issue of the translation model to return gender-specific variants from gender-neutral source sentences. We point out that the model seems to associate family to women and career to men, the same is viewed where the masculine form is associated with driver whereas cook is associated with the feminine. And this shows a strong bias in the target set training data itself.

For both the WEAT ES and RND, there is a much noticeable skew. Therefore, our findings show a very high likelihood of the Luganda-English Machine Translation (MT) systems to produce gender-specific outcomes towards a specific gender stereotype given a gender-neutral source sentence.

In the TGBI measure, a score of 0 corresponds to high bias and 1 corresponds to low bias Cho et al. (2019). The bias values, in Table 5, show that both models show greater gender bias towards the female stereotype in all sentences (occupation-positive). Strong gender bias is greatly projected in the Google Translate model in positive sentiment sentences. The overall results occupation shows a high bias in the Luganda-English model, while positive projects high gender bias in the Google Translate model.

7 Conclusion and Future Work

This work provides a parallel corpus to train baseline Luganda language models. However, to our knowledge, it's very evident that there is less research invested in gender bias evaluation for Ugandan low-resourced languages. We address this problem by providing a gendered parallel corpus to support future research. We trained and tested baseline Luganda (Luganda-English) and (English-Luganda) language translation, models.

We evaluate gender bias in a Luganda-English machine translation model using the WEFE framework metrics that take in queries of data. We also compare the results of WEFE metrics with the TGBI and human evaluation. All our results show a tendency of machine translation systems to project gender bias, when translating from a gender-neutral to a language with gender-specific.

With this research, we believe it will help in future work in finding ways to mitigate gender bias in Ugandan languages with gender-neutral pronouns, given their low resourcefulness. Through this work, we look forward to creating new methods to debias such systems and metrics to measure gender bias that covers all the traits of our languages.

8 Acknowledgements

This work is funded by AI4D-African Language Program and the Lacuna NLP grant. We are grateful to the Department of African languages at Makerere University for helping with the translation of the English datasets to Luganda.

References

Adelani, D., Ruiter, D., Alabi, J. O., Adebonojo, D., Ayeni, A., Adeyemi, M., Awokoya, A., and España-Bonet, C. (2021). Menyo-20k: A multi-domain english-yorùbá corpus for machine translation and domain adaptation. *ArXiv*.

Akera, B., Mukiibi, J., Naggayi, L. S., Babirye, C., Owomugisha, I., Nsumba, S., Nakatumba-Nabende, J., Bainomugisha, E., Mwebaze, E., and Quinn, J. (2022). Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*.

Babirye, C., Nakatumba-Nabende, J., Katumba, A., Ogwang, R., Francis, J. T., Mukiibi, J., Ssentanda, M., Wanzare, L. D., and David, D. (2022). Building text and speech datasets for low resourced languages: A case of languages in east africa. In *3rd Workshop on African Natural Language Processing*.

Badilla, P., Bravo-Marquez, F., and Pérez, J. (2020). Wefe: The word embeddings fairness evaluation framework. In *IJCAI*, pages 430–436.

Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., et al. (2022). Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Cho, W. I., Kim, J. W., Kim, S. M., and Kim, N. S. (2019). On measuring gender bias in translation of gender-neutral pronouns. *arXiv preprint arXiv:1905.11684*.

Ciora, C., Iren, N., and Alikhani, M. (2021). Examining covert gender bias: A case study in turkish and english machine translation models. *arXiv preprint arXiv:2108.10379*.

Eberhard, D. M., Simons, G. F., and (eds.), C. D. F. (2020). Ethnologue: Languages of the world. twenty-third edition.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Gupta, G., Ramesh, K., and Singh, S. (2021). Evaluating gender bias in hindi-english machine translation. *arXiv preprint arXiv:2106.08680*.

Hovy, D. and Spruit, S. L. (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.

Ifeoluwa Adelani, D., Oluwadara Alabi, J., Fan, A., Kreutzer, J., Shen, X., Reid, M., Ruiter, D., Klakow, D., Nabende, P., Chang, E., et al. (2022). A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv e-prints*, pages arXiv–2205.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., et al. (2018). Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Koehn, P. (2017). Neural machine translation. *arXiv preprint arXiv:1709.07809*.

Macketanz, V., Avramidis, E., Burchardt, A., Helcl, J., and Srivastava, A. (2017). Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. *Cybernetics and Information Technologies*, 17(2):28–43.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R., and Modupe, A. (2020). Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2003.04986*.

Martinus, L. and Abbott, J. Z. (2019). A focus on neural machine translation for african languages. *arXiv preprint arXiv:1906.05685*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nakayiza, J. (2013). *The sociolinguistics of multilingualism in Uganda: A case study of the official and non-official language policy, planning and management of Luruuri-lunyara and Luganda*. PhD thesis, SOAS, University of London.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., et al. (2020). Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Nyoni, E. and Bassett, B. A. (2021). Low-resource neural machine translation for southern african languages. *arXiv preprint arXiv:2104.00366*.

Olaide, F. O. and Azizi, W. (2019). Model for translation of English language noun phrases to Luganda. *London Journal of Research in Computer Science and Technology*.

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., and Kaur, R. (2021). Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021a). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021b). Gender bias in machine translation. *arXiv preprint arXiv:2104.06001*.

Sweeney, C. and Najafian, M. (2019). A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667.

Vanmassenhove, E., Hardmeier, C., and Way, A. (2019). Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.