Integration of Coreference Resolution into Machine Translation from Non-gendered to Gendered Languages

Team CLR: Catherine Mei, Linette Kunin, Ryan Conti

1 Introduction & Related Work

In linguistics, a language is said to have grammatical gender if nouns in that language are assigned gender categories regardless of whether or not they are animate or inanimate and even if the gender is unrelated to the noun's real-world properties. For example, in Spanish, a language with grammatical gender, the word for table, "mesa", is assigned feminine gender.

A large portion of NLP research is conducted in English, which lacks grammatical gender. This focus on English has resulted in models struggling to keep track of the gender of pronouns, nouns, and their associated adjectives. This issue becomes an especially potent problem in settings where English must be machine-translated into a gendered language.

While this kind of error may seem innocuous, it has real consequences that hinder machine translation's use in environments where semantic retention is crucial. Consider the potential applications of machine translation in diplomacy; machine translation can reduce communication friction, but ensuring that meaning is neither lost nor misconstrued in translation is so critical that even minor slip-ups due to gender disagreement could cause problems.

Gender-biases in machine translation occur because most translation systems function through more-or-less explicit word to word mappings across different languages. As a result, the context and relationship between subjects across sentences is not preserved, and translations from non-gendered to gendered languages often result in gender disagreements.

Coreference resolution could be one possible solution to this problem. Coreference resolution systems cluster mentions with their referent (e.g. pronouns to the corresponding proper nouns) (Pražák and Konopík, 2022). This information is useful in achieving gender agreement. If coreference resolution were integrated into the machine translation process, then our translation architectures would be less prone to mistakes in gender agreement.

The proposed research project aims to extend current machine translation systems in order to reduce gender-biases when translating from a non-gendered to a gendered language. Specifically, we wish to improve machine translation architectures built upon transformers. Transformers utilize attention mechanisms to preserve connections between parts of an input sequence (Ashish Vaswani, 2017). Current transformer-based machine translation systems can correctly produce gender-agreeing translations of contiguous phrases and subjects. In this paper, we explore how altering model inputs by including information relevant to correct gender agreement can impact a model's translation performance. Our aim is to integrate coreference resolution into the translation process.

This project experiments with machine translation between English and two gendered languages, Spanish and Irish. Spanish has a relatively well-understood translation system compared to Irish as well as a more robust dataset to work from. We chose these two languages because of the resource disparity between the two and the fact that Spanish and Irish, while both Indo-European, are still somewhat genealogically distant from one another.

Currently, much of machine translation is handled by transformers (Ashish Vaswani, 2017). Existing approaches to solving the problem of proper-gender translation include applying Seq2seq implementations of translation for less-resourced languages (Ulčar and Robnik-Šikonja, 2022), providing extra context to neural machine translation models to correct biases (Sharma et al., 2022), modifying and augmenting encode-decoder structures within the transformer (Das et al., 2022), and augmenting inputs to include information on the gender of the speaker (Vanmassenhove et al., 2018).

This project improves on these solutions to find a way to mitigate gender disagreement in machine translation from a non-gendered to a gendered language. In this project, we establish a baseline Seq2seq transformer model. We then train our model on datasets with strong gender bias. We then explore how to alter inputs in order to improve gender agreement in our model despite the initial gender bias. We also explore the effect of debiasing our training dataset on model performance.

Much work on gender in machine translation focuses on mitigating biases in professions (e.g. biases like assigning masculine to doctor and feminine to nurse) (Sharma et al., 2022). However, most of these models do not comprehensively cover gender agreement between nouns, pronouns, and their adjectives in gendered languages.

If we believe the goal of machine translation is to replace the need for manual translation in all contexts (or at least supplement it), then it is important to make translation models sound as fluent as possible. An adjective's gender provides information on what noun or pronoun the adjective is describing, especially in contexts where coreferencing is ambiguous. Improving performance on these tasks will propel NLP research closer to a reality in which machine translation from non-gendered to gendered languages is appropriate for the real-world.

2 Evaluation

Our first evaluation metric is the BiLingual Evaluation Understudy (BLEU score). The BLEU score measures the similarity of the machine-translated text to a set of high quality reference translations (Papineni et al., 2002). We use the BLEU score to evaluate the baseline performance of our replicated transformer model. After we augment the inputs, we evaluate our updated model again using the BLEU score. We look at sentences where the BLEU score changes from our baseline to our new model to qualitatively evaluate whether the improvements if any are due to better gender agreement or due to unrelated changes.

Second, we perform qualitative analysis on all of our results to ensure that sentences that were translated correctly in our baseline model are still translated correctly in our new model. We sampled and observed the accuracy of our translation model after modifications.

We train and test our model on parallel corpora, which have the same text passages translated between different languages. Using parallel corpora allows us to have target translations which we can compare to our model translation using BLEU score evaluation. We initially select a gender-biased dataset for training so our model will overemphasize gender translation errors and we can better determine if the model correctly handles cases of feminine gender agreement. A portion of the data is omitted from training and allocated for validation and testing.

For Spanish, we use selected sentence pairs from the Tatoeba Project, a collection of sentences and translations generated and translated by the public. We use data from here because it is most likely to have latent societal biases like gender biases. In addition, for our third approach, we add data from various books with predominantly female characters, obtained from OPUS, to make our dataset less biased and increase training size. For Irish, we use data from ParaCrawl, a dataset taken from selective websites compiled by Common Crawl, an open repository composed of over a decade of web-crawled data.

As a baseline heuristic, we find the percentage of feminine pronouns used. We use percentage of feminine pronouns as a proxy for gender bias because of its simplicity to calculate, but other words are also gendered. Some nouns are naturally feminine (e.g. in Spanish, table, bed, etc.) and all corresponding adjectives must agree with this gender. We are unable to find a better way to calculate this bias but we acknowledge its limitations. For the Spanish dataset, we find that only 21.6 percent of the gendered pronouns and names used are feminine. In Irish, only 12.4 percent of the gendered pronouns are feminine.

Another challenge we face is that machine translation performs poorly on small data sets. Given our limited computational resources (at most 12 hours running on Google Colab), we can only train our model on substantially smaller data sets than those of the best translation systems and train them for far fewer epochs than what is needed for convergence. We use the BLEU score to find the baseline model performance and use it to benchmark improvements.

The BLEU score tells us how closely our model translation matches with the target translation. If pronouns or adjectives have the wrong gender assignment, this corresponds to a lower BLEU score. Thus, increases in BLEU score in combination with qualitative analysis of the changes in the translation that caused this increase tell us whether our implementation improves gender assignment in translation to gendered languages.

3 Baseline Methodology

3.1 Implementation

Our baseline model is the canonical sequence to sequence transformer model (Ashish Vaswani, 2017). We trained this Seq2seq model initially with a gender-biased data set to achieve our baseline results.

We incrementally improved our model by modifying the input dataset to the transformer. In particular, we first augmented the gender-biased dataset with coreference information from the English input. Next, we added adjective-noun reference information to the input sentences. Finally, we attempted to gender-balance and unbias our dataset.

Using both the transformer model, coreference resolution, and adjective reference is justifiable here because translation often lacks relational information. This project combines established translation techniques (with transfomers) with coreference and adjective reference information to maintain gender agreement in translation.

Ultimately, this collection of data sets, evaluation metrics, and approaches should enable us to reduce gender disagreement in machine translation from English, a language without grammatical gender, to Spanish and Irish, two languages with grammatical gender.

3.2 Baseline Model

Our baseline model is a sequence-to-sequence transformer consisting of 3 components:

- `TransformerEncoder`: encodes source sentence. Encoded sentence is passed to decoder.
- `TransformerDecoder`: Predicts next word in target sequence given sequence so far.
- `PositionalEmbedding`: maintains model awareness of word order.

The following are descriptions of the three components in our baseline model.

`TransformerEncoder` Module

The encoder class consists of an attention layer, a dense NN layer, and two normalization layers. We implemented our models using `TensorFlow`. Specifically, our attention layer uses `layers.MultiheadAttention`. The dense NN layer is implemented using `layers.Dense`, and the normalization layers are implemented using `layers.LayerNormalization`.

In our `forward` method, the input data flows through the attention layer, the first normalization layer, the dense NN layer, and finally the second normalization layer, at which point it is returned as output. We chose to implement two normalization layers within our model in order to control the magnitude of our outputs and hidden layers.

`TransformerDecoder` Module

The transformer decoder consists of two attention layers, one dense NN layer, and three normalization layers. Similar to the encoder, our attention layers are implemented using `layers.MultiheadAttention`, the dense NN layer is implemented using `layers.Dense`, and the normalization layers are implemented using `layers.LayerNormalization`.

The first attention layer receives the queries, keys, and values as inputs. These inputs represent the embedded target sequence that has been augmented with positional information. The second attention layer receives the encoder output (keys and values) in addition to the normalized output of the first decoder attention layer. Then, the output of the second attention layer is passed to the dense NN layer. Between each of the layers described above, we added a normalization layer to control the magnitude of our outputs.

`PositionalEmbedding` Module

The purpose of the `PositionalEmbedding` class is to maintain word order of our input sequences. Within the `PositionalEmbedding` class, we implemented two embedding layers using `layers.Embedding`. The first embedding layer (`self.token_embeddings`) handled the embedding of the words within the input. The second embedding layer (`self.position_embeddings`) handled the embedding of the word positions.

In our `forward` method, we found both the word embeddings and position embeddings, and we returned both pieces of embedded data as the output. We chose to have these two embeddings because both the token embeddings and ordering embeddings can improve our model performance.

3.3 Baseline Model Results

Results for Spanish dataset: We trained our baseline for 5 epochs on 83,276 sentence pairs and validated on 17,844 sentence pairs. The average BLEU score of our translations was 0.338, which is considered to mean that the translations are "understandable to good," which aligns with our qualitative observations. We achieved an accuracy of 63.12%.

Our model currently displays an ability to correctly translate gender in cases of pronoun usage, e.g. "[start] Ella dio luz a su primer hijo a los veinte años de edad. [end]" or "She gave birth to her first child at 20 years old" is translated as "[start] ella le dio a su primer niño que tenía viejo años [end]" or "She gave [it] to her first boy who was old years." While the content was translated incorrectly, the pronoun "she" was correctly retained. This appears to be a consistent success in our model's translations.

That being said, gender agreement between some nouns and adjectives appears to have issues. For example, the sentence "[start] Su padre murió la semana pasada. [end]" or "His/her father died last week" was translated as "[start] su padre murió la semana pasado [end]" which has essentially the same meaning, but "pasado" is made masculine here even though it's an adjective modifying a feminine noun. Additionally, we have some examples where the translation simply reads a feminine name as a masculine object: "[start] Tom sabía dónde compró Mary sus alimentos. [end]" or "Tom knew where Mary bought her food," was translated as "[start] tom sabía dónde compró el mundo de sus [UNK] [end]" or "Tom knew where the Earth bought its [UNK]". Mary, a feminine name, is replaced by "el mundo," a masculine noun.

Overall, the model currently does a decent job at translating from English to Spanish, but it remains imperfect when it comes to the task of fully translating meaning. The model handles pronoun translation well, but it needs improvement in the case of gender agreement between nouns and adjectives.

Results for Irish dataset: We trained our baseline model for 5 epochs on 54,094 sentence pairs, which we partitioned into sets of 37,866 training sentence pairs, 8,114 validation pairs, and 8,114 testing pairs. The average BLEU score of our translations was 0.157, which is qualitatively described as "difficult to get the gist of," which makes sense based on our qualitative observations. We achieved an accuracy of 50.39%.

Our model was unable to recognize many feminine words, which may indicate the necessity for more training data. Feminine words were often marked as [UNK], such as in the example case of "Níl go ró-fhada ó shin, bean a shíl mé go raibh ag fanacht gan gaol dlúth i mo shaol," or, roughly, "Not too long ago, I thought a single woman was waiting in my life." Our model output "ní raibh aon [UNK] ann ná uair an chloig ó shin i [UNK] [UNK] [UNK] i [UNK]" for this sentence, which roughly translates as "there was no [UNK] an hour ago in [UNK] [UNK] [UNK] in [UNK]." Errors stemming from the inability to recognize feminine words frequently occurred when distinctly feminine nouns appeared in our model.

It is interesting to note that plural and masculine pronouns resulted in fewer errors. For example, "Táimid lonnaithe i Shandong le rochtain ar iompar áisiúil" or "We are located in Shandong with convenient transportation access" where "we" comes from the plural first person pronoun "Táimid" was translated as "is féidir linn a dhéanamh le [UNK] [UNK] [UNK] a fháil", which does contain a plural first person case in "is féidir linn."

The model needs more improvements with regards to its Irish translation, as it struggles with producing coherent translations in most cases. We suspect that the difference in model performance between the Spanish and Irish datasets is due to the differing sizes of the datasets.

4 Improvements on Baseline Methodology

4.1 Applying Coreference Resolution

Utilizing the same Seq2seq model as in our baseline exploration, we augmented our inputs with special tokens containing coreference information and retrained our model on these modified inputs. Specifically for every pronoun, we added the noun it references as a token right before the pronoun.

We did this by taking our input sentence, and running it through spaCy's neural coreference model. This gave us coreference clusters as a dictionary. Each noun was a key and the value was a list of all the nouns and pronouns that refer to the original noun.

For each sentence, we then used the clusters to replace each pronoun x with [the noun pronoun x references] x. As an example, after running our code, the original sentence **"The girls realized their lifelong dream of playing professional soccer, and they were content."** became **"The girls realized [The girls] their lifelong dream of playing professional soccer, and [The girls] they were content."** We add in "[The girls]" before "their" and "they", two cases where on it's own, the gender of these pronouns would be ambiguous in English. We then trained the same model with these new inputs.

We hoped that by making pronouns with ambiguous gender in English (like they or their) appear closer to nouns they refer to (like the girls), we could take advantage of transformers attention layer to learn the gender of these pronouns.

4.1.1 Results of Adding Coreference Resolution

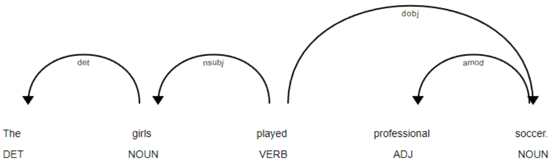
Results for Spanish Dataset: In Spanish, augmenting our input sentences with coreference information and training our model for 5 epochs resulted in a BLEU score of 0.342 and an accuracy of 63.13%. This is a minor increase from our baseline model's BLEU score and accuracy in Spanish. A sentence that was previously translated incorrectly, "She is always dressed in black" is now translated correctly to "Ella siempre esta vestida de negro." This worked correctly because "vestida" (or dressed) is feminine and agrees with the feminine subject "ella".

Results for Irish Dataset: For Irish, we attempted to run the same model, but even when training on a very small dataset, we were unable to run to completion in the Google Colab environment. We have a few hypotheses as to why the same process took so much longer for Irish than for Spanish. First, the process of producing coreference information in a sentence is roughly exponential with the sentence's length since it requires creating a syntax tree for the sentence; our Irish dataset featured longer sentences on average than our Spanish dataset, so this is likely one source of the computational complexity. Additionally, the Irish dataset does come largely from Twitter, so it is possible that the less formal setting of Twitter and abundance of different sociolects lends itself to more irregular language usage that is difficult for the model to resolve.

Comments & Interpretations: It's important to note that only some pronouns in English are actually gender ambiguous (for example: they, their, it), and we only had a few sentences where the gender of the ambiguous pronoun was referring to a feminine noun. This could explain why the increase in our BLEU score and accuracy as compared to our baseline model wasn't larger. Since many of the failed gender agreement in translating from non-gendered to gendered languages occurs in adjectives, we thought we would address this next.

4.2 Adding Adjective Reference

In order to further improve the model with coreference between nouns and their pronouns, we decided to add coreference tags for adjectives and their corresponding nouns. Our justification for this is that gender agreement for an adjective is exactly determined by the gender of the noun it modifies. In order to do this, we used the neuralcoref library to identify which nouns each adjective corresponds to; then, we modified the training sentence by appending the corresponding noun before each adjective. We retrained the model using this data augmentation. Below is an example of the syntax tree constructed after using the neuralcoref library on a sample sentence:



For example, if adding simple coreference resolution in the previous step output the sentence **"The girls realized [The girls] their lifelong dream of playing professional soccer, and [The girls] they were content."** then adding adjective reference would change things to **"The girls realized [The girls] their [dream] lifelong dream of playing [soccer] professional soccer, and [The girls] they were [The girls] content."** We add in "[dream]" before "lifelong," "[soccer]" before "professional," and "[The girls]" before "content." The first two additions may not seem especially useful since the adjective-noun pairs are adjacent in the sentence, but the power of this approach is visible in the third addition of "[the girls]" before "content," since this identifies that the adjective "content," which modifies "they" in the last independent clause of the sentence, is in fact modifying the noun phrase that "they" refers to, i.e. "the girls." Appending this information makes it more clear that the adjective is modifying a feminine noun.

4.2.1 Results of Adding Adjective Reference

Results for Spanish Dataset: After adding adjective reference to our model, we achieve a BLEU score of 0.346 and an accuracy of 63.14%. This is a minor increase over the inclusion of coreference resolution between nouns and pronouns. For example, the sentence, "Tom recommended to María that she not go there alone." was translated to "Tom le recommendó a María que no vaya ahí sola." Here, "sola" is feminine and agrees with the subject "María". We hypothesize that the lack of significant improvement in Spanish with this change is due to the high frequency of cases in which adjectives are already adjacent to the nouns they modify. In these cases, the additional adjective reference information may not add much useful information in some sentences.

Results for Irish Dataset: Our model was similarly unable to run to completion for the Irish data since it was already unable to run to completion when incorporating only coreference resolution between nouns and pronouns. Our hypotheses for the difficulty of this process are the same in this case.

We still believe the principles behind adjective-noun and noun-pronoun coreference tagging would help with gender agreement in Irish, but we would need to reevaluate the model being used and either determine if there is a more efficient way to compute this for longer sentences or simply run with more robust computational resources.

4.3 Gender Balancing Data set

We posited that the gender imbalance within our datasets may have negative impacts on the performance of the transformer model. Therefore, in this last approach, we decided to process and balance our dataset to be more gender neutral.

Since we cannot generate more feminine datasets to increase the number of feminine training examples, we decided to extend our training dataset by combining it with another dataset. For the Spanish dataset, we found a books dataset from OPUS and combined the sentences with feminine pronouns or names in this dataset with our original dataset.

In order to classify our dataset into feminine, masculine, and gender neutral, we identified common words, names, and nouns that may indicate gender. We created some sets `masculine_markers` and `feminine_markers` that contained the most common names and gendered pronouns. We used these sets to filter our dataset into different categories.

Table 1 summarizes the profile of our data. We found that the number of masculine sentences in our training data was roughly double the number of feminine examples.

	Male	Female	Neutral
Total Examples	22,683	12,609	88,751
Included Examples	11,342	12,609	88,751

Table 1: Number of sentences per category (Spanish)

Finally, we decided to include only half of the masculine training examples in order to gender balance our dataset. The exact number of examples included for each category is shown Table 1.

4.4 Results of Gender Balancing Dataset

Results for Spanish Dataset: After training the Seq2seq transformer model for 5 epochs, we achieved a BLEU score of 0.330 and accuracy of 57.98%. This is a slight decrease in both BLEU score and accuracy from our previous results. However, we believe the decrease reflects a lack sufficient number of training examples rather than poor model performance.

Results for Irish Dataset: We did not gender balance the Irish dataset because we were limited by the availability of training data. In particular, finding an Irish dataset with gendered sentences to use for our baseline model and coreference resolution approach was very difficult since most available data were sourced from political treaties and government documents. Therefore, we did not have the option of expanding our Irish dataset.

We knew that gender balancing the dataset would inevitably decrease the number of available training examples, and since our Irish dataset was already small, we felt that it didn't make sense to remove more data from the dataset. Particularly, the transformer model requires a large amount of training examples, and so even if we unbiased our dataset, it is very plausible that we would still see a decrease in model performance.

4.5 Summary of Results

Table 2 contains a summary of the results for the Spanish dataset and the Irish baseline model results.

	BLEU Score	Accuracy
Baseline Irish	0.157	50.39%
Baseline Spanish	0.338	62.34%
Coreference	0.342	63.13%
Adjective	0.346	63.14%
Unbiasing	0.330	57.98%

Table 2: Summary of Results for Spanish & Irish

5 Conclusion & Discussion

Ultimately, the changes we made to our initial model showed promise in certain instances, such as our inclusions of coreference information for nouns, pronouns, and adjectives. At the same time, we saw a dip in performance after balancing our datasets to debias them. We do believe that we would observe better performance on a debiased data set if we could increase the size of the dataset to be equal or greater than the size of our existing dataset, since gender bias in the training dataset likely lowered our model's performance on feminine gender agreement. Nevertheless, the fundamental ideas of our approach here seem to have potential as promising means for improving gender agreement in translation from non-gendered to gendered languages, or at least from English to non-gendered Indo-European languages that are sufficiently similar to Spanish and, to a lesser extent, Irish.

With regards to the differences in our model's performance on Spanish and Irish, we note that there are likely a few core reasons for the gap. First, our Irish dataset included complicated sentences that came from sources like Twitter which may have been more difficult to learn from; second, it's possible that our baseline Seq2seq model was more predisposed to perform well on Spanish; third, our Irish dataset was half the size of our Spanish dataset, and given that our Spanish dataset was small to begin with, this meant we had a rather small amount of training data for Irish for the baseline model. Finally, the length of the Irish training sentences were longer on average than the Spanish training sentences, which is likely the reason that we could not run coreference resolution for Irish within the constraints of Google Colab.

In the future, there are several steps we could take to improve to our model (in addition to the obvious of training for more epochs and on a larger dataset). We would investigate more balanced ways to include meaningful coreference information and omit redundant information (such as removing the coreference tag for a noun that is adjacent to its modifying adjective). We could also investigate ways to incorporate coreference information without directly inserting coreference tags into the training sentences themselves, with the idea that it may be better to include the original training sentence and the coreference information as two separate data structures to train on. Despite the limitations, our approach seems promising in improving gender agreement in translation from non-gendered to gender languages. Therefore, we believe further exploration in this direction could lead to lasting impacts on gender translation systems today.

6 Impact Statement

In general, the impact of improving gender agreement in machine translation is as crucial as any other improvement in machine translation: it improves meaning retention across translation and enables a future world in which speakers of different languages can communicate without confusion.

More specifically, proper gender agreement across translations can have many applications in the real world. For example, proper machine translation enables smoother diplomacy without constant need for human interpreters. Especially in such diplomatic instances, proper translation is *crucial*; it is possible to conceive of a scenario in which a translation has a gender agreement error that significantly changes the meaning across languages. For instance, if there's an entirely feminine group, and an entirely masculine group, using ”they” in English communicates ambiguous gender information — such information is often contextual in English. However, in gendered languages,“they” feminine and “they” masculine unambiguously change the group the sentence refers to. Improper gender agreement in translation may result in the target sentence referring to a different group than the source sentence was referring to. Additionally, gender disagreement across translation could produce communication friction, which could cause one group to be perceived as disrespectful or rude.

Additionally, using a person's self-identified gender in speech or text is important. Using correct pronouns and gender agreement shows social awareness, while not using them can be perceived as a purposeful insult or a sign of being intolerant. It is important that if we use translation systems in daily life, that they incorporate proper gender agreement to avoid these mistakes.

Finally, current translation systems between English and low-resource languages are error prone and not well-refined. Improving gender agreement in this space would be a big step forward in revitalization and preservation efforts for low-resourced language communities, it would give these speaker communities more exposure to the NLP community, and it would help expand inclusion and diversity efforts in the academic community.

Ultimately, improving gender agreement in translation is one of many crucial steps involved with improving machine translation in general, and all of its potential impacts are inextricably linked with the potential impact of better machine translation.

Supplemental Materials

Our code (in the form of Google Colab notebooks) and data can be accessed at: CLR_code

References

Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. 2017. Attention is all you need. *NIPS*, pages 1–15.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kumar Patra. 2022. Improving multilingual neural machine translation system for indic languages.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ondřej Pražák and Miloslav Konopík. 2022. End-to-end multilingual coreference resolution with mention head prediction.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts.

Matej Ulčar and Marko Robnik-Šikonja. 2022. Sequence to sequence pretraining for a less-resourced slovenian language.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.