Generating Gender Augmented Data for NLP

Nishtha Jain[1], Maja Popovic[2], Declan Groves[2], Eva Vanmassenhove[4]

[1]ADAPT Centre, Trinity College Dublin,
[2]ADAPT Centre, Dublin City University,
[3]Microsoft, Dublin,
[4]Department of CSAI, Tilburg University

[1,2]firstname.lastname@adaptcentre.ie, [3] degroves@microsoft.com,[4]e.o.j.vanmassenhove@tilburguniversity.edu

Abstract

Gender bias is a frequent occurrence in NLP-based applications, especially pronounced in gender-inflected languages. Bias can appear through associations of certain adjectives and animate nouns with the natural gender of referents, but also due to unbalanced grammatical gender frequencies of inflected words. This type of bias becomes more evident in generating conversational utterances where gender is not specified within the sentence, because most current NLP applications still work on a sentence-level context. As a step towards more inclusive NLP, this paper proposes an automatic and generalisable re-writing approach for short conversational sentences. The rewriting method can be applied to sentences that, without extra-sentential context, have multiple equivalent alternatives in terms of gender. The method can be applied both for creating gender balanced outputs as well as for creating gender balanced training data. The proposed approach is based on a neural machine translation (NMT) system trained to 'translate' from one gender alternative to another. Both the automatic and manual analysis of the approach show promising results for automatic generation of gender alternatives for conversational sentences in Spanish.

1 Introduction

Recent studies have exposed challenging systematic issues related to bias that extend to a range of AI applications, including Natural Language Processing (NLP) technology (Costa-jussà, 2019; Blodgett et al., 2020). Observed bias problems range from copying biases already existing in data to claims that the training process can lead to an exacerbation or amplification of observed biases (Zhou and Schiebinger, 2018; Vanmassenhove et al., 2021). The algorithms learn to maximize the overall probability of an occurrence, leading to preferences for more frequently appearing training patterns.

With this work, we propose a method for generating (more) balanced data in terms of one of the main types of bias frequently observed in language: gender bias. Gender bias can occur in language due to the fact that some languages have a way of explicitly marking (natural or grammatical) gender while others do not (Stahlberg et al., 2007). Gender bias in translation is usually manifested when animate entities (e.g. professions) are translated from gender neutral language (e.g. English) into a gendered language (e.g. Spanish) because the instances seen in training data are biased. Also, conversational utterances are prone to bias, both in machine translation as well as in other NLP applications, because systems often do not have the ability to provide multiple gender variants. Therefore, users are simply presented with the most probable option which is prone to bias. In our work, we aim to enable the generation of multiple gender variants by expanding each sentence with the missing gender variants, thus fostering inclusion in online conversations/NLP applications. Generating gender variants can and should also be used to create gender balanced conversational data that can be used to train less biased NLP models such as machine translation models, language models, chat bots, etc.

Unlike previous studies, we did not want to limit ourselves to one specific gender phenomenon, such as gender markings on professions (Zmigrod et al., 2019)) (for which the gender can easily be swapped by using hand-crafted lists) or first person personal pronouns (Habash et al., 2019)). The objective of this research aims to include as many cases as possible of gender alternatives related not only to gender of persons but also to grammatical gender of the objects referred to. In Example 1, (a) illustrates an example of two alternatives for a sentence where

Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pages 93–102

there is agreement with the grammatical gender of an object referred to in the previous sentence, while in (b) there is agreement with the gender of the speaker/writer (i.e. a person).

Example 1.
(a) [MALE] ¿Está complet**o**? – [FEMALE] ¿Está complet**a**? [1]
(b) [MALE] Estoy confundid**o**. – [FEMALE] Estoy confundid**a**.[2]

At this stage, our approach does not discriminate between human referents and objects. It is furthermore limited to the generation of binary gender alternatives. We are aware of the importance and challenge of dealing with non-binary gender(Ackerman, 2019) which we aim to tackle in future work.

The research was carried out in collaboration with an anonymous industry partner with a specific application in mind that deals with conversational sentences. Our approach aims to alleviate gender bias in the said application. We focus on one gender-rich language (Spanish), however, scalability and generalizability were kept in mind while designing the approach. Our approach can be summarized as follows:

1. Identifying (appropriate) sentences/segments that should have the opposite gender variant for some words. POS sequences were used to extract such segments from the OpenSubtitles corpus[3].

2. Creating gendered variants for the words in such segments by applying a rule-based approach.

3. Training a neural rewriter on the compiled gender-parallel Spanish data in order to be able to automatically generate gendered variants on unseen data sets. This additional step makes the approach more scalable as it removes the need for any preprocessing.

The first two steps are necessary since there is a lack of readily available open-source gender-parallel data for training. Although language knowledge and a POS tagger are necessary for these steps, the human effort and necessity for external linguistic tools are minimal (contrary to other approaches which heavily rely on linguistic tools (Zmigrod et al., 2019) or on manually created gender-parallel data (Habash et al., 2019).

2 Related Work

In the literature on gender in NLP, two main approaches for bias mitigation can be identified: (a) approaches that attempt to mitigate bias during model or word representation training, and/or (b) approaches that aim to augment the data by creating more variety in the training set (pre-processing step) or in the output (post-processing step). In the following paragraphs, we focus on the latter as it is most closely related to our approach.

There have been attempts to artificially increase the variety in already existing data sets by creating alternatives to sentences in order to decrease the overall bias (in terms of gender).[4] This approach has been referred to in the literature as 'Counterfactual Data Augmentation'(CDA) (Lu et al., 2018). Their CDA approach consists of a simple bidirectional dictionary of gendered words such as he:she, her:him/his, queen:king, etc. Zhao et al. (2018) does not use the term CDA as this was introduced later, but what they describe can be interpreted as a rudimentary approach to CDA: they augmented the existing data set by adding additional sentences in which personal pronouns 'he' and 'she' had been swapped.

Another CDA approach is described in Zmigrod et al. (2019). Similar to Lu et al. (2018), the approach relies on a bidirectional dictionary of animate nouns. Unlike Lu et al. (2018), pronouns are not handled and the languages worked on are Hebrew and Spanish, languages that have more gender markers than English. Since solely changing the nouns into their male/female counterpart often requires the enforcement of grammatical gender agreement of accompanying articles and adjectives, they introduce Markov Random Fields with optional neural parametrisation that can infer the effect of the swap on the remaining words in the segment. Their approach is limited to mitigating gender stereotypes related to animate nouns and relies on dependency trees, lemmata, POS-tags and morpho-syntactic tags in order to solve issues related to the morpho-syntactic agreement.

In the field of machine translation (MT), due

[1]English: "Is it complete?"
[2]English: "I am confused."
[3]https://opus.nlpl.eu/

[4]Different types of bias exist, however, the current approaches have focused on gender, possibly because many languages have explicit gender markers.

to specific discrepancies between the information encoded in the source and target data, there has been some work on generating the appropriate gender variant for ambiguous source sentences.[5] Vanmassenhove et al. (2019) appends gender tags to the source side of the training data indicating the gender of the speaker. As such, during testing, the desired (or multiple) gender variant(s) can be generated by adding tags. Basta et al. (2020) also experiment with incorporating a gender tag, and investigate adding the previous sentence as additional context information. Both methods result in the improvement of automatic MT scores as well as on gender accuracy for English-to-Spanish translation. Similarly, Bentivogli et al. (2020) developed NMT systems using gender tags and evaluated them specifically on gender phenomena.

The work described in Habash et al. (2019) is the most similar to ours. They proposed an approach for automatic gender reinflection ("re-gendering") for Arabic. They propose a method which consists of two components: a gender classifier and a NMT gender rewriter. In order to build the NMT rewriter, they first manually created a corpus annotated with gender information. Subsequently, each gendered sentence is re-gendered manually in order to obtain the necessary gender-parallel data for training. This way, they are able to provide gender alternatives for sentences with natural gender agreement with the first person singular.

Our research, in contrast, aims to augment existing data with gender alternatives in a broader sense: it is not limited to singular first person phenomena, ambiguity in multilingual settings, or phenomena related solely to gender agreement. It involves the gender of adjectives, past participles, and several types of pronouns for which the referent is not explicitly mentioned within the context of the sentence.

3 Generating gender-parallel data

As mentioned in the introduction, our main objective is to create an automatic gender rewriter using NMT. In order to do so, we need gender-parallel training data that consists of possible gender variants in both directions (masculine-to-feminine and feminine-to-masculine). Such data sets are, unfortunately, not publicly available, which is why we first leveraged linguistic knowledge and rules to generate a sufficient amount of gender-parallel data.

Therefore, we identified the sequences of POS classes that show gender agreement in Spanish and can thus be 're-gendered': adjectives, past participles, and several types of pronouns. A detailed description of how the different word classes are tackled to generate gender alternatives is described below. We would like to point out that our target data consisted of very short sentences, where there is at most agreement with one referent.[6] As such, our approach is limited to tackle sentences alike and cannot handle the generation of alternatives for sentences where more than two gender alternatives could be generated (due to grammatical agreement of the re-genderable word with multiple entities).

3.1 Re-genderable word classes

Past participles In principle, almost all Spanish past participles have an explicit agreement with their referent and can thus be re-gendered. However, in certain contexts they should not be: if they follow or precede a referent noun (*"Película aburrida"*, *"Acceso permitido."*) thus agreeing with the gender of the noun, or if they follow the auxiliary verb *"haber"* thus representing past tense and not a property of a person/object (*"he enviado"*, *"has descansado"*). If they appear in isolation (*"Ocupad**o**/ocupad**a**."*, *"Aburrid**o**/aburrid**a**."*), or merely surrounded by interjections or punctuation (*"Ocupad**o**/ocupad**a**, gracias."*, *"Buenos dias, recibid**o**/recibid**a**, ¡gracias!"*), adverbs (*"muy cansad**o**/cansad**a**"*), or a linking verb (*"Estoy registrad**o**/registrad**a**."*, *"Parece acabad**o**/acabad**a**."*), they can be re-gendered.

We also included pairs of past participles bound by conjunctions, referring to the same person or object, since in these sentences, both instances should be re-gendered (*"aburrid**o**/aburrid**a** y cansad**o**/cansad**a**."*, *"acabad**o**/acabad**a** y pagad**o**/pagad**a**."*).

Adjectives Many Spanish adjectives are gendered and have an explicit gender marker corresponding to the gender of its referent. However, some adjectives are gender neutral. Gendered and

[5] 'I am a teacher' or 'I am smart' in English are not marked for gender. However, in many other languages they would be morphologically marked for the male or female gender (e.g. French, Spanish...).

[6] For example, sentences such as "I am happy and they are angry." are not covered by our approach as both 'happy' and 'angry' are in agreement but with different referents, 'I' and 'they' respectively. Such sentences would require the generation of more than two alternatives since both referents are ambiguous.

neutral adjectives can (largely) be identified based on their specific suffixes (for example *"-al"*, *"-nte"*, *"-ble"*, so the adjectives *"genial"*, *"interesante"*, and *"probable"* are neutral), while other suffixes indicate gendered adjectives (for example "o/a", so the adjective *"correct**o**/correct**a**"* has variants).

In addition, similarly to past participles, the given context has to be taken into account for gendered adjectives: they should not be re-gendered if they immediately precede or follow a noun (with or without article) which determines the gender (*"Presupuestos adjuntos."*, *"¡Maravillosa idea!"*, *"La información correcta."*). Also, adjectives following neutral demonstrative pronouns *"eso"* or *"esto"* should not be re-gendered (*"Eso es bueno."*). Analogous to past participles, adjectives in isolation (*"List**o**/List**a**."*, *"perfect**o**/perfect**a**."*, *"segur**o**/segur**a**."*, *"¡fantástic**o**/fantástic**a**!"*), surrounded by punctuation (*"Correct**o**/correct**a**, saludos."*), preceding verb (*"¿Estás list**o**/list**a**?"*) or adverb (*"Es muy lind**o**/lind**a**."*) can be re-gendered.

When two adjectives are present, in a conjunction, and refer to the same referent, both should be re-gendered.

Clitic pronouns Some Spanish clitic pronouns, namely *"lo(s)"* and *"la(s)"* should be re-gendered (e.g. *"**Lo**/**la** veo."*, *"**Lo**/**la** adjunto."*) while *"le(s)"* should not be changed (*"Le veo."*, *"Le digo."*). However, in some cases *"lo"* can represent a general concept not referring to a particular object, such as in *"lo siento"* (I'm sorry), *"lo sé"* (I know). If some of these are re-gendered, the precision will decrease.

Clitic pronouns attached to verbs Clitic pronouns can be attached to a verb infinitive (*"Gracias por acabarl**o**/acabarl**a**."* (thanks for finishing it), *"Quiero verl**o**/verl**a**."* (I want to see it)). Similar to the isolated clitic pronouns, there are certain exceptions, such as *"Es bueno saberlo"* (it is good to know). If the gender neutral clitic pronoun *"le"* is attached to a verb (*"Quiero tenerle informado."* (I want to keep you/him/her informed)), it should not be re-gendered. Gendered pronouns attached to an imperative should also be re-gendered (*"Déjal**o**/Déjal**a**."* (leave it), *"Hazl**o**/Hazl**a**."* (do it)). On the other hand, clitic pronouns which refer to an indirect object, such as *"mándame"* (send me), are neutral. Finally, if there are two attached clitic pronouns, *"Mándamel**o**/Mándamel**a**."* (send it to me), only the gendered part (in this case *"lo"/"la"*) should be re-gendered.

Demonstrative pronouns Demonstrative pronouns *"esto"*, *"eso"* and *"aquello"* are neutral, while *"est**o**s/est**a**s"*, *"est**e**/est**a**"*, *"es**e**/es**a**"*, *"aquell**o**/aquell**a**"* are gendered. If the referent is missing in the sentence and the pronoun is gendered, they should be re-gendered.

3.2 Adding gender variants by rules

Whether a gender alternative translation should be generated does not solely depend on the word classes it contains but also on the structure of the sentence. If the referent is missing in a sentence, then an additional variant with the opposite gender should be generated. If the referent is present in a sentence, only one gender variant is grammatically correct, and as such, these sentences are to be left unchanged. The presence or absence of a referent can be determined by the sequence of POS tags in a sentence[7]. For example, if we want to check whether a sentence with an adjective "creo que es correct**a**" (gloss: "I believe (it) is correct-feminine") needs an additional re-gendered variant or not, its POS sequence "VERB CONJUNCTION VERB ADJECTIVE" indicates that there is no referent noun within the given context. Therefore, another variant of the adjective "correct" should be provided: "creo que es correct**o**". In contrast, the sentence "la solución es correcta" with POS sequence "ARTICLE NOUN VERB ADJECTIVE" contains a referent noun "solución", and therefore it should not be re-gendered.

For each re-genderable sentence, we apply rules for changing the ending of the corresponding word, if necessary. The POS sequences to identify re-genderable sentences and the subsequent rules used to re-gender the corresponding words in such sentences are given in detail in the Appendix. It is worth mentioning we also used POS sequences to identify neutral sentences (those which should be not re-gendered) since we wanted the parallel corpus to contain both.

4 Gender-parallel data

In order to create gender-parallel data, a set of Spanish subtitles was downloaded from the OPUS (Tiedemann, 2012) website.[8] After basic

[7] Assuming that the sentences are short- this approach would not generalize to longer sentences

[8] http://opus.nlpl.eu/

filtering (removing too long and non-alpha numeric segments), a set of short sentences with up to 10 (untokenized) words was extracted. This candidate set consisted of 22 458 968 sentences. This data set was POS tagged using Treetagger[9]. The sentences matching the POS sequences mentioned in the Appendix were extracted from this data set. This set consisted of more than 1M sentences. For each extracted re-genderable sentence, the alternative gender variant is created by applying appropriate rules described in the Appendix. After applying rules on all re-genderable structures, we joined both re-gendering directions (masculine-to-feminine and feminine-to-masculine) in order to create a balanced data set. As already mentioned, the corpus also contains a number of sentences that are not to be regendered. By including these neutral sentences in our training data, we encourage the rewriter to: (a) learn when to generate alternatives and when not to, and (b) how to generate those alternatives, if necessary. In this way, a corpus with about 2.2M gender-parallel sentences was created. This corpus was then separated into train, development (∼1k sentences) and test (∼3k sentences) sets. The rewritten parts of the development and test sets were revised manually and the errors were corrected for about 6% of sentences and 1.5% of words. The training set, being large, was not verified manually, thus it contained some noise.

In addition to OpenSubtitles, we also obtained data from the industry partner consisting of around 8 000 sentences readily available with all possible alternative versions of the sentences provided. An additional 22 000 sentences had to be revised manually in order to produce the correct gender variant for re-genderable sentences. This set was used as an additional test set for the re-writer. One part of this set can be handled by the described POS sequences and rules ("structured test 1"), while another part contains different POS sequences and cannot be handled by these rules at all ("unstructured test 1"). The latter test set will give a good estimation of the scalability of our approach. An overall split of data sets is described in Table 1. The OpenSubtitles data was split in the standard way for machine translation, namely a few thousands of segments for development and test sets and the rest for the training set.

set	segments
training (OpenSubtitles)	2 193 657
development (OpenSubtitles)	1 018
test (OpenSubtitles)	3 066
structured test1	5 648
unstructured test1	15 892

Table 1: Statistics of data used for building the NMT rewriter.

5 Neural Rewriter

Once we compiled a sufficient amount of gender-paralell data, we were able to train our automatic rewriter. The automatic rewriter is a NMT system trained on the following parallel data: original sentences as the source language, and re-gendered sentence as the target language. For neutral sentences, the source and the target parts are identical.

The NMT rewriter was built using the publicly available Sockeye[10] implementation (Hieber et al., 2018) of the Transformer architecture (Vaswani et al., 2017). The system operates on sub-word units generated by byte-pair encoding (BPE)(Sennrich et al., 2016). We set the number of BPE merging operations to 32000. We have experimented with the following setups:

- a Standard NMT system without any additional tags
- an NMT system with neutrality/re-genderability tags in the source part

The system with tags was built using the same technique as proposed in (Johnson et al., 2017) for multilingual MT systems and used for many other applications including gender-informed MT (Vanmassenhove et al., 2019). For our experiments, we added a label 'N' (neutral) or 'G' (re-genderable) to each source sentence. These tags are implicitly present in the gender-parallel data – if the source and the target parts differ, it is a re-genderable sentence, if they are identical it is neutral. Therefore, the tags are certainly available for the training and development sets, but they might not be available for the test sets. Therefore, this system was assessed in two ways:

- "NMT-T": neutrality/re-genderability tags are available for the test sets
- "NMT-AT": the tags are not available for the test sets (a realistic scenario) and therefore are

[9] https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[10] https://github.com/awslabs/sockeye

assigned automatically by the gender classifier described in the next section (which is similar to the approach described in (Habash et al., 2019).)

5.1 Gender Classifier

In order to explore potential benefits of automatic pre-classification for automatic rewriting, a classifier to distinguish between 're-genderable' (G)[11] and 'neutral' (N)[12] sentences was also designed. The tags generated by this classifier were used to assess the performance of the "NMT-AT" re-writer by appending them to the sentences.

Data

The classifier was built on the data set of about 8 000 sentences provided by the industry partner. These sentences were balanced in both directions i.e., both masculine-to-feminine as well as feminine-to-masculine counterparts of a given sentence were present and labelled as G. The rest of the sentences were labeled as N.

For the sake of designing a generalised classifier, the development set consisted of sentences from the OpenSubtitles corpus (and was the same as the development set used for the NMT system).

The final classifier was tested on two different test sets - one consisted of the 22 000 conversational sentences sourced from the industry partner and another extracted from the OpenSubtitles corpus.

Features

Following on the work of Habash et al. (2019) for the gender identification step, features using character n-grams, word n-grams and morphological information were created from the training data. To begin with, TF-IDF scores of character n-grams of length 4-7 with maximum features capped at 20 000 and of word n-grams of length 1-3 were generated. These two feature matrices were joined together along with a morphological feature that denoted the presence of a gendered word in the sentence. The resulting training data was a high dimensional data frame with around 40 000 features.

Due to the limited size of the training set, neural network based classifiers were ruled out. Instead, owing to the high dimensional nature of the data, we used a SVM based classifier for training. All the steps described in this section were implemented in Python 3.7 using sklearn[13], pandas[14] and StanzaNLP[15] libraries.

	Industry Test Set			OpenSubs		
	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.
Overall	82%	-	-	80%	-	-
G	-	96%	60%	-	97%	76%
N	-	76%	98%	-	56%	93%

Table 2: Gender Classifier Results

Precision and Recall

The SVM based classifier was tested on two sets of data as described in Section 5.1. This was done in order to assess the generalisability of the classifier. Given the small size of the training data, the performance of the classifier looks promising thus far (see Table 2).

It can be observed in Table 2 that the classifier clearly performs better on the test data set consisting of sentences sourced from the industry partner as compared to the data extracted from OpenSubtitles. While the accuracy is comparable on both sets (80%), the precision and recall of neutral sentences is higher on the industry data than the set compiled from OpenSubtitles data. The high recall of sentences labelled as G implies that the classifier is almost always successful at recognising sentences that need to be re-gendered (i.e. sentences that need an alternative variant). However, it incorrectly predicts the labels of a substantial number of N-labelled sentences, which in turn results in a low precision of re-genderable sentences. As we want to avoid generating (incorrect) gender alternatives for neutral sentences, our aim was to first attain a high precision for neutral sentences and then aim towards a high recall for the same. The tags generated by this classifier for the industry sourced data and OpenSubtitles data were used to test the "NMT-AT" rewriter.

6 Results for generating gender variants

Our first experiment consisted of using the implementation of CDA by (Zmigrod et al., 2019) to generate gendered variants. However, this work only tackled animate nouns, which rarely occur in the conversational sentences we investigated in this

[11] Grammatical gender markings are not related to a referent within the sentence, therefore these markings have to be expanded.

[12] No gender markers that need to be expanded.

[13] https://scikit-learn.org/stable/

[14] https://pandas.pydata.org/

[15] https://stanfordnlp.github.io/stanza/

work. Our re-implementation of their approach generated the correct gender variant for only 1% of the sentences. Because of the very low recall, this implementation was not directly applicable for our research. In addition to this, since our work aims to tackle multiple gender related word classes, we explored extending the implementation by augmenting the list with character adjectives. On doing so, we found that this implementation generated the correct gendered variant in only 9% of the cases. An important point to note is that 3% of the neutral sentences (for which variants should not have been generated) were also converted as opposed to the 1% with only animate nouns, attributed to the presence of more words in the hand-crafted lists. In order to cover more words and improve the performance of this implementation on our data set, we considered augmenting the hand-crafted list with past participles and/or clitic pronouns. However, that increased the size of the list exponentially and made the approach prone to errors, inefficient and not scalable to other languages.

6.1 Automatic evaluation of neural rewriter

The results in the form of error rates are shown in 3. Since we are not performing typical machine translation, namely converting one language into another one, but only converting a few words in the sentence into a sentence in the same language, these error rates are not related to any of the typical automatic evaluation metrics (such as TER, etc.) but to the amount of incorrectly converted words. For each system, numbers in the left column represent the count of incorrectly converted words normalised by the total number of sentences, while numbers in the right column represent the count of incorrectly converted words normalised by the total number of words in the corpus. The numbers in the first row and first two columns can be interpreted as follows: left: 6.4% of all sentences have incorrectly converted words in ; right: 1.50% of all words are incorrectly converted.

First, it can be noted that the error rates are lower for the template-based "in-domain" test sets than for the unstructured "out-of-domain" test sets, which is in line with our expectations. The change in error rate is mainly due to discrepancies in the re-genderable segments. The error rates in the neutral segments are comparable in the out-of-domain and in-domain test sets.

Adding manual tags indicating whether a sentence should get a gender alternative or not (e.g. 'neutral' vs 'regenderable') reduces the error rates on all test sets for both types of segments. A similar performance can not be achieved by adding automatic tags. Automatic tags deteriorate the performance on neutral segments, but reduce the error rates for re-genderable segments, especially for the unstructured "out-of-domain" test set. The manually tagged results indicate the potential of a classifier. These results tie up with the results of the gender classifier (Section 5.1) which is good at classifying the re-genderable sentences as denoted by a high recall of sentences labelled 'G', however it doesn't do very well at labelling neutral sentences as 'N'. It tends to mislabel many of those sentences as 'G', resulting in a low recall and, consequently, incorrect re-gendering.

For the sake of completeness, error rates are reported for the rule-based rewriter, too. The error rates for re-genderable sentences are lower than the NMT rewriter without tags and for neutral sentences the error rate is 0%; it should be noted that the rules are applicable only to data sets which strictly conform to the described template structures.

6.2 Qualitative manual inspection of errors

In order to better understand the nature of errors and remaining challenges, a qualitative manual inspection was carried out on all test sets. First of all, it is observed that in general, the NMT re-writer does not intervene on large portions of a sentence but addresses only specific words, which is exactly what it is expected to do. This is a positive result, as generating gender variants implies changing specific gendered words and does not involve changing entire segments. It also facilitates the evaluation since manual inspection is needed only to identify the nature of incorrect words.

The analysis revealed that the most frequent error for neutral sentences are re-gendered pronouns and adjectives which should not be changed. Also, the most frequent error in re-genderable sentences is leaving them unchanged. These types of errors are predominant in structured sentences, and two examples, one for neutral and one for regenderable sentence, can be seen in Table 4(a). It can also be seen that adding tags can help in some cases.

For unstructured sentences, there are more error types especially for neutral sentences, and examples can be seen in Table 4(b). In the first three

set	type	NMT		NMT-T		NMT-AT		rules	
test	all	6.4	1.50	4.5	1.03	17.9	4.21	6.1	1.43
(structured)	neutral	5.3	1.13	2.5	0.48	33.3	7.07	0.0	0.0
	re-genderable	7.1	1.81	6.0	1.51	6.0	1.72	6.1	1.43
test1	all	2.4	0.54	1.3	0.27	4.5	0.99	3.2	0.7
(structured)	neutral	4.8	0.95	2.2	0.43	8.7	1.73	0.0	0.0
	re-genderable	0.8	0.19	0.6	0.14	1.6	0.38	3.2	0.7
test2	all	11.9	2.13	5.2	0.93	10.4	1.87	not	
(unstructured)	neutral	3.3	0.58	0.3	0.04	6.0	1.07	applicable	
	re-genderable	57.3	10.7	31.1	5.84	33.4	6.26		

Table 3: Results for NMT rewriter: error rates (%): count of incorrectly converted words normalised by the total number of sentences (left columns) and normalised by the total number of words (right columns).

(a) structured sentences

type	original	correct	NMT	NMT-T
N	esto es perfecto	esto es perfecto	esto es **perfecta**	esto es perfecto
G	está adjunt**o**	está adjunt**a**	está adjunt**o**	está adjunt**o**

(b) unstructured sentences

	type	original	correct	NMT	NMT-T
1)	N	no son lo mismo	no son lo mismo	no son **la misma**	no son lo mismo
2)	N	aquello fue encantador	aquello fue encantador	aquello fue **encantadora**	aquello fue encantador
3)	N	¿a quién aprovecha?	¿a quién aprovecha?	¿a quién **aprovecho**?	¿a quién aprovecha?
4)	N	indíqueme la disponibilidad	indíqueme la disponibilidad	indíqueme la **emperbilidad**	indíqueme la **evelbilidad**
5)	N	indíqueme su disponibilidad	indíqueme su disponibilidad	indíqueme su disponibilidad	indíqueme su **escorpibilidad**
6)	N	unos momentos extraordinarios	unos momentos extraordinarios	unos momentos **extraordinarias arios**	unos momentos extraordinarios
7)	N	indíquenos cuánto	indíquenos cuánto	**indíquenas** cuánto	indíquenos cuánto
8)	G	esta es la adecuada	este es el adecuado	**esta** es ***la*** **adecuada**	**esta** es ***lo*** **adecuada**
9)	G	esta la hemos recibido	este lo hemos recibido	**esta la** hemos recibido	**esta** lo hemos recibido

Table 4: Examples of incorrectly generated sentence variants for (a) structured sentences and (b) unstructured sentences.

sentences, the same error type as for structured sentences can be seen, namely some words are changed which should not be changed. Adding tags helped in both cases. However, some other error types can be seen, such as converting some (not gender-related) words into non-existing words in sentences 4) and 5). For sentence 5), generating a non-existing word was triggered by adding tags. Sentence 6) shows an unnecessary re-gendering as well as adding non-existing words. This was also resolved by adding tags. In sentence 7), a word which is not at all related to gender was converted, and this was prevented by adding tags.

As for regenderable sentences, the vast majority of errors are again the unchanged words which had to be changed. If there is more than one word to be regendered, sometimes they all remain unchanged (sentence 8) and sometimes only some of them are regendered (sentence 9). Tags can help to some extent, but only for some words, not all.

Adding tags generated by the classifier also increases the number of correctly re-gendered structures at the cost of a small number of additions of non-existing words.

7 Conclusions and Future Work

In this paper, we describe an initial approach towards enriching short conversational sentences with their gender variants. Unlike other related work, our approach is not limited to tackling the first person singular phenomena, swapping third person pronouns or merely dealing with occupa-

tional or generally animate nouns. In addition, with our approach, the reliance on linguistic knowledge and tools is kept to a minimum in order to facilitate real-world deployment.

The main hurdle for this type of research is the absence of large training sets. Although provided with some manually annotated data from the industry partner, the data provided was far from sufficient to train a state-of-the-art automatic gender re-writer.

Therefore, training data was extracted from OpenSubtitles using linguistic knowledge about the targeted language, namely Spanish. Re-genderable types of words (POS classes) were identified and then frequently occurring 're-genderable' as well as 'neutral' POS patterns were extracted. By applying the corresponding rules to the re-genderable sentences, a large gender-parallel Spanish data set was compiled.

Next, an NMT rewriter was trained in order to 'translate' each re-genderable sentence into its gender alternative which showed promising performance both in terms of automatic as well as of manual evaluation.

In addition, it is shown that providing additional information regarding the need for rewriting in the form of tags could be helpful for the NMT system, as similar tags have shown to be useful for other applications such as multilingual translation, controlling politeness and gender in MT, etc. While gold standard labels show better performance than the labels generated by the gender classifier, the classifier shows promising results given the very small training set. Further experiments should investigate a classifier trained on larger amount of data.

In future work, we would like to explore how a similar approach can be applied on more sentence structures in Spanish, as well as for different languages which exhibit distinct gendering rules. Furthermore, different NMT architectures, e.g. character-level NMT or an NMT system with linguistically motivated subword units could be an interesting extension to the conducted experiments, given that gender is usually marked by specific morphemes (usually not more than one or two specific characters). In addition to that, the performance of the gender classifier can be improved to produce more accurate tags by using larger annotated training sets, adding more morphological information in features and using word embeddings instead of TF-IDF scores.

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1).

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2020. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, pages 1–14.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia Antonino Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the must-she corpus. *arXiv preprint arXiv:2006.05754*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Marta R Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, pages 1–2.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. In *Transactions of the Association of Computational Linguistics, Volume 5:1*, pages 339–351, Vancouver, Canada.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender Bias in Natural Language Processing. In *arXiv:1807.11714*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1715–1725, Berlin, Germany.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social communication*, pages 163–187.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3003–3008.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of The Thirty-first Annual Conference on Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008, Long Beach, CA, USA.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4847–4853, Brussels, Belgium.

J Zhou and L Schiebinger. 2018. AI Can be Sexist and Racist – It's Time to Make it Fair. In *Nature 559*, pages 324–326. https://www.nature.com/articles/d41586-018-05707-8.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 1651–1661, Florence, Italy.